# Diagnosis of Pleural Mesothelioma using Machine Learning

by

Olaoluwa Julianah Abejide

A thesis submitted in partial fulfillment.

of the requirements for the degree of

MSc Computational Sciences

The Office of Graduate Studies

Laurentian University

Sudbury, Ontario, Canada

**THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE**
**Laurentian Université/Université Laurentienne**
Office of Graduate Studies/Bureau des études supérieures

| | | |
|---|---|---|
| Title of Thesis<br>Titre de la thèse | Diagnosis of Pleural Mesothelioma using Machine Learning | |
| Name of Candidate<br>Nom du candidat | Abejide, Olaoluwa Julianah | |
| Degree<br>Diplôme | Master of Science | |
| Department/Program<br>Département/Programme | Computational Sciences | Date of Defence<br>Date de la soutenance October 26, 2023 |

**APPROVED/APPROUVÉ**

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewak
(Committee member/Membre du comité)

Dr. Oumar Gueye
(Committee member/Membre du comité)

Dr. Jioti Singh Kirar
(External Examiner/Examinateur externe)

Approved for the Office of Graduate Studies
Approuvé pour le Bureau des études supérieures
Tammy Eger, PhD
Vice-President Research (Office of Graduate Studies)
Vice-rectrice à la recherche (Bureau des études supérieures)
Laurentian University / Université Laurentienne

**ACCESSIBILITY CLAUSE AND PERMISSION TO USE**

# Abstract

Mesothelioma is cancer that develops in the pleura. The most common cause of this disease is contact with asbestos. Patients with mesothelioma have a better chance of surviving if they are diagnosed quickly. This study utilizes a variety of machine learning to enhance pleural mesothelioma diagnosis. The possibility of misclassification was decreased by extracting features from a preexisting dataset. SVM, Decision Trees, and Random Forests are only a few machine learning classifiers trained using essential and foundational features. Accuracy, precision, recall, and F1-score were just a few measures used to evaluate these classifiers' performance in cross-validation. SVM demonstrated excellent accuracy, precision, recall, and F1-score when classifying individuals as either healthy or having mesothelioma. The results show the potential of machine learning techniques for early diagnosis of pleural mesothelioma. Machine learning algorithms improve diagnosis accuracy and turnaround time, improving patient outcomes. Using the results of this research, a fully automated technique for diagnosing mesothelioma might be developed, allowing clinicians more time to provide better care for their patients.

**Keywords:** Mesothelioma, Machine Learning, SVM, Anfis, Decision Trees, and Random Forests

# Acknowledgements

This thesis is an outcome of the hard work and dedication of the people who helped directly and indirectly during my master's academic journey. To this end, I would like to appreciate their contributions.

My deepest appreciation goes to my thesis supervisor, Dr. Kalpdrum Passi for his unwavering guidance, insightful feedback, and continuous support throughout the process.

I acknowledge my beloved family- children, husband, siblings, and loving mother for their encouragement and prayers throughout my study, research, and thesis writing.

Finally, I extend my appreciation to the academic community, mentors, friends, and all those who have indirectly played a role in this endeavor. This thesis is a culmination of collective efforts, and I am honored to have had the opportunity to contribute to the realm of knowledge.

# Table of Contents

Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Pleural mesothelioma (Lung Cancer)

Rare and fatal, pleural mesothelioma develops in the thin membrane that lines the chest cavity and lungs. Asbestos exposure causes pleural cells to have mutations in their DNA. In recent decades, there has been a global rise in the prevalence of pleural mesothelioma, which now claims the lives of an estimated 43,000 people annually. Pleural mesothelioma can be hard to diagnose since its symptoms are vague and can be mistaken for those of other diseases like pneumonia or lung cancer. Pleural mesothelioma patients have a better chance of survival if their condition is diagnosed and treated early. Currently, a pleural mesothelioma diagnosis is made by a combination of imaging techniques (such as a chest X-ray, CT scan, or MRI) and a biopsy (the removal and examination of a tissue sample). However, the accuracy and non-invasiveness of these techniques are not optimal. Although CT and MRI scans can produce high-resolution images of the lungs and their structures, they may not be able to distinguish between benign and malignant lesions and have a high false-positive rate. In contrast, a biopsy can provide an absolute diagnosis, but it is a surgical treatment that isn't always possible or appropriate for patients.

Recent developments in machine learning algorithms have demonstrated significant potential for increasing the accuracy of medical diagnostics, particularly the identification of cancer.

Machine learning has not only been investigated for its ability to improve diagnosis, but it has also been investigated for its capacity to predict treatment results and survival rates in individuals who have pleural mesothelioma.

## 1.2 Objectives of the Study

The overall objective of this study is to diagnose pleural mesothelioma using deep learning while the specific objectives of the study include:

- Obtain the best and most significant features from the available dataset because some irrelevant features may lead the algorithm to false classification.
- Train different machine learning classifiers like SVM, Decision Trees, and Random Forest, using significant and basic features obtained from the previous step.
- Present classification report using different evaluation parameters like accuracy, Precision, Recall, and area under the curve.

## 1.3 Organization of the Thesis

In this thesis, we utilize a variety of machine learning and deep learning techniques to enhance pleural mesothelioma diagnosis. This thesis is organized as follows:

- Chapter 1 introduces the topic of pleural mesothelioma and objective of our study.
- Chapter 2 is a literature review of the previous studies in pleural mesothelioma diagnosis.
- Chapter 3 provides the definition of data and methods utilized in the proposed model.
- Chapter 4 is about the evaluation process and the obtained results.
- Chapter 5 provides a conclusion of the research and discusses the potential future direction.

# Chapter 2

# Literature Review

Mesothelioma is a rare and severe disease produced by exposure to asbestos fibers. It can damage the lining of the lungs, abdomen, or heart, and it has a lousy prognosis due to its aggressive nature and resistance to typical cancer treatments. Mesothelioma can be caused by exposure to asbestos fibers (Institute, 2021). It is challenging to identify mesothelioma since the diagnostic procedures that are now in use, including imaging tests and biopsies, have some limits regarding their accuracy and reliability. The potential of machine learning to improve the accuracy of a mesothelioma diagnosis has been studied in previous research. For instance, using an approach that enables feature selection and machine learning, this article proposes a classification and detection method for mesothelioma cancer. The CFS correlation-based feature selection approach is first used in the feature selection process. It acts as a filter, selecting just the traits that are relevant to the categorization. The accuracy of the categorization model is improved as a direct consequence of this. After that, classification is carried out with the help of Naive Bayes, fuzzy SVM, and the ID3 algorithm Shobana et al in 2022). CFS was evaluated by experiments on artificial and natural datasets. Three machine learning algorithms were used: C4.5 (a decision tree learner), IB 1 (an instance-based learner), and Naive Bayes. Experiments on artificial datasets showed that CFS quickly identifies and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features. On natural domains, CFS typically eliminated well over half the features. In most cases, classification accuracy using the reduced feature set equaled or bettered accuracy using the complete feature set. Feature

selection degraded machine learning performance in cases where some features were eliminated which were highly predictive of very small areas of the instance space.

Another study that looked into using a machine learning model for the automated detection of pleural mesothelioma. (Komal Saxena et al., 2022). According to the findings of the research, they investigated AI methodologies and consider the exact DM (Diabetes Mellitus) conclusion outcomes in this study, which focuses on DM determination. K-nearest neighborhood, linear-discriminant analysis, Naive Bayes, decision-tree, random forest, support vector machine, and logistic regression analyses have been used in clinical decision support systems in the detection of mesothelioma. To test the accuracy of the evaluated categorizers, the researchers used a dataset of 350 instances with 35 highlights and six execution measures. LDA, NB, KNN, SVM, DT, LogR, and RF have precisions of 65%, 70%, 92%, 100%, 100%, 100%, and 100%, correspondingly. In count, the calculated complication of individual approaches has been evaluated. Every process is chosen on the basis of its characterization, exactness, and calculated complications. SVM, DT, LogR, and RF outclass the others.

Karapınar et al in 2020, roposed a machine learning based early detection system for this fatal disease. An open database is used for the experiments and different methods have been applied to the problem of diagnosing mesothelioma disease. Accuracy and sensitivity performance metrics were used for the evaluation of the methods. The results show the diagnostic performance of different machine learning methods and present a successful early diagnosis system (Karapınar et al, 2020). Patients have a CT (Computed Tomography)-scan and lung x-ray traditionally, but the exact method is biopsy. There are also different biopsy methods for its diagnosis. Its prevalence is one or two in a million around the world, but for Turkey it is disastrous. Five hundred people are diagnosed as mesothelioma every year in Turkey. This serious rate makes early diagnosis systems

crucial for mesothelioma. In this paper, a machine learning based early detection system has been proposed for this fatal disease. An open database is used for the experiments and different methods have been applied to the problem of diagnosing mesothelioma disease. Accuracy and sensitivity performance metrics were used for the evaluation of the methods. The results show the diagnostic performance of different machine learning methods and present a successful early diagnosis system with 80% accuracy.

Khan et al in 2018 prioritized to identify the disease in its premature stage, to control the invasive nature of cancer. The most prominent causes of cancer are environmental issues, life style and genetic heritage. Malignant Mesothelioma (MM) is one of the fastest growing neoplasm tumour in human body, that originates due to mesothelium cells in various parts of the human body, and directly affects the pleura. The main causes of MM are asbestos exposure, exposure to the high doses of radiation to the chest or abdomen, genetics disposition and the infection of simian virus 40. In this paper MM tumour classification is performed using Support Vector Machine (SVM). Tumour is classified as either malignant or benign. SVM is trained on features extracted in the form of symptoms of MM cancer. SVM is trained on features extracted in the form of symptoms of MM cancer. The combination of SVM and feature extraction method is compared with Probabilistic Neural Network (PNN) and Multi-layered neural networks (MLNN) and shows better results than both PNN and MLNN (Khan et al., 2018). Among the 324 data examples of patients, 40% were chosen for testing and other 60% were for the training phase. The accuracy for SVM is 98 percent and 93%, 94 % for PNN and MLNN respectively.

Pleural Mesothelioma (PM) is an unusual, belligerent tumor that rapidly develops into cancer in the pleura of the lungs. Pleural Mesothelioma is a common type of Mesothelioma that accounts for about 75% of all Mesothelioma diagnosed yearly in the U.S. Diagnosis of Mesothelioma takes

several months and is expensive. Given the risk and constraints associated with PM diagnosis, early identification of this ailment is essential for patient health. In this study, they use artificial intelligence algorithms recommending the best fit model for early diagnosis and prognosis of Malignant Pleural Mesothelioma (MPM). He retrospectively retrieved patients' clinical data collected by Dicle University, Turkey and applied multilayered 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). Choudhury, 2021 evaluated the models, compared and tested them 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). SGD, AdaBoost.M1, KLR, MLP, VFDT generate optimal results with the highest possible performance measures with more than 75% accuracy. Accuracy of classifications were: SGD 69.23 % , AdaBoost.M1 75.29%, KLR 69.51%, MLP 64.11 %, VP 70.38 %, VFDT 70.38%, CC 70.38%, s-Pegasos 67.03%.

C-reactive protein, platelet count, duration of symptoms, gender, and pleural protein were found to be the most relevant predictors that can prognosticate Mesothelioma. This study confirms that data obtained from biopsy and imaging tests are strong predictors of Mesothelioma but are associated with a high cost; however, they can identify Mesothelioma with optimal accuracy (Choudhury, 2021). Summary of Literature Review is given in Table 2.1.

**Table 2.1. Literature Review summary**

| Reference | Study purpose | Dataset | Methods used | performance |
|---|---|---|---|---|
| Shobana et.al 2022 | Classification and Detection of Mesothelioma Cancer Using Feature Selection-Enabled Machine Learning Technique | Mesothelioma's disease data set, UCI | Combine CFS feature selection and fuzzy SVM | Accuracy:100% |
| Komal et al 2022 | using a machine learning model for the automated detection of pleural mesothelioma | A dataset of 350 instances with 35 highlights and six execution measures | LDA, NB, KNN, SVM, DT, LogR, and RF have precisions | LDA precisions 65%, NB precisions 70%, KNN precisions SVM precisions LogR precisions 92%, And RF precisions 100% |
| Karapınar et al 2020 | a machine learning based early detection system for detection of pleural mesothelioma | Mesothelioma's disease data set, UCI | Gradient Boosted Trees, SVM, KNN, Random Forest without feature selection | Gradient Boosted Trees accuracy: 80% KNN accuracy:81.79% SVM accuracy: 100% Random Forest accuracy:81.54% |
| Khan et al. 2018 | Classification of Malignant Mesothelioma Cancer Using Support Vector Machine | Mesothelioma's disease data set, UCI | SVM is compared with Probabilistic Neural Network (PNN) classification method and Multi-layered neural networks (MLNN). | Probabilistic Neural Networks accuracy: 93.1% MLNN Accuracy 94.41% Support Vector Machine 98% |
| Choudhury, 2021 | early diagnosis and prognosis of Malignant Pleural Mesothelioma | clinical data collected by Dicle University | applied multilayered perceptron (MLP), voted perceptron (VP), Clojure classifier (CC), kernel logistic regression (KLR), | Accuracy of classifications: SGD 69.23 % ,AdaBoost.M1 75.29, KLR 69.51 %, MLP 64.11 %, |

| | | | | |
|---|---|---|---|---|
| | | | stochastic gradient decent (SGD), adaptive boosting (AdaBoost), Hoeffding tree (VFDT), and primal estimated sub-gradient solver for support vector machine (s-Pegasos) | VP 70.38 %, VFDT 70.38 %, CC 70.38 %, s-Pegasos 67.03%. |

# Chapter 3

# Data and Methods

The methodology employed in this study involved the following steps: (1) data collection and preprocessing, (2) feature selection and engineering, (3) model selection and training, (4) hyper-parameter tuning and optimization, and (5) evaluation and analysis. The first step involved obtaining the necessary data and preparing it for use by removing any noise, outliers, or missing values. The second step involved selecting the most relevant features and engineering new ones to improve the model's performance. The third step involved choosing the most appropriate model based on the problem and training it on the preprocessed data. The fourth step involved tuning the model's hyper-parameters using grid or random search techniques to achieve the best possible results. Finally, the last step involved evaluating the model's performance using accuracy, precision, recall, and F1 score metrics and analyzing the results to draw conclusions and insights. Figure 3.1 illustrates the proposed methodology.

## 3.1 Dataset details

The term "malignant mesothelioma" (MM) refers to a highly aggressive pleural tumor type. The presence of these malignancies has been linked to asbestos exposure. On the other hand, it could have something to do with past infection with simian virus 40 (SV40) and a possible genetic susceptibility. In addition, molecular pathways can potentially play a role in the progression of mesothelioma. In machine learning, datasets are an integral part of the methodology. Data collection is one of the most challenging tasks, especially when it is related to healthcare data. A health record of 324 patients with mesothelioma from the Diyarbakir region of southeast Turkey has been analysed. It was found that the area has a prevalence of natural asbestos fibers. Therefore,

the dataset provides evidence that the population more exposed to asbestos has a high mesothelioma ratio. The dataset includes socio-economic, geographical, histopathological, and clinical features. The target class is known as the class of diagnosis' which had two defined types, 1 Patient, 2 Healthy. The dataset has been imbalanced because patients were only 96, and healthy individuals were 228. Mesothelioma is linked to living in rural areas, which increases the risk of getting the disease.

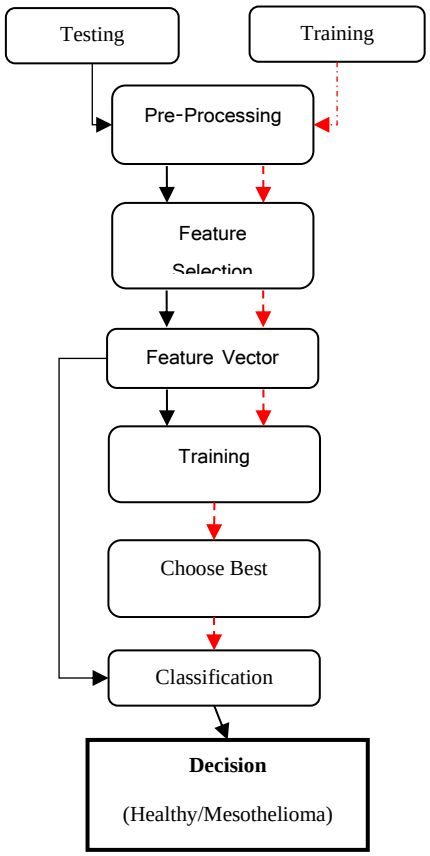


**Figure 3.1. Proposed Research Methodology**

The Faculty of Medicine at Dicle University in Turkey compiled the mesothelioma illness data sets[1]. Data from 324 patients diagnosed with mesothelioma. Every sample in the dataset has 34

[1] Mesothelioma's disease data set - UCI Machine Learning Repository

features, and the output is divided into two categories: healthy and mesothelioma. Table 1.3 shows feature details along with the output.

**Table 3.1. Feature Details of Dataset available on UCI[2]**

| **Input Features (34)** | **Output Class** |
|---|---|
| Age, gender, city, asbestos exposure, type of MM, duration of asbestos exposure, pleural lactic dehydrogenise, pleural protein, pleural albumin, pleural glucose, dead or not, pleural effusion, pleural thickness on tomography, pleural level of acidity (pH), C-reactive protein (CRP), Glucose, albumin, total protein, alkaline phosphatise (ALP), blood lactic dehydrogenise (LDH), ache on chest, weakness, habit of cigarette, performance status, white blood, WBC, HGB, PLT, sedimentation, diagnosis method, keep side, cytology, duration of symptoms, dyspnoea | Healthy and Mesothelioma |

## 3.2 Preprocessing of data

It is necessary to preprocess the medical data to diagnose pleural mesothelioma using machine learning techniques. Preprocessing involves cleaning and transforming the raw data into a format that can be easily used by machine learning algorithms. The preprocessing steps ensured that the machine learning algorithms could accurately diagnose pleural mesothelioma based on the relevant medical data. We can use one of the following feature selection algorithms. This research used four methods for feature selection as there is no need for feature extraction because they were already designed in dataset for analysis.

### 3.2.1 Principal Component Analysis (PCA)

PCA (Principal Component Analysis) is primarily a dimensionality reduction technique rather than a feature selection method. However, it can be used indirectly for feature selection in certain scenarios. PCA and used for feature selection is described below (Guo et.al, 2002).

[2] University of California, Irvine

1. Dimensionality Reduction: PCA is a mathematical technique used to reduce the dimensionality of a dataset while preserving as much variance as possible. It does this by transforming the original features into a new set of orthogonal (uncorrelated) features called principal components. These principal components are linear combinations of the original features and are ordered by the amount of variance they explain in the data. The first principal component explains the most variance, the second explains the second most, and so on.
2. Feature Selection: While PCA itself does not directly select or eliminate individual features; it can be used as a feature selection method in the following way:
    a. Calculate Principal Components: Run PCA on the dataset to compute the principal components.
    b. Variance Explained: Analyze the variance explained by each principal component. A cumulative explained variance plot shows how much of the total variance in the data is explained by including each principal component.
    c. Choose the Number of Components: Decide how many principal components to keep based on the desired level of variance retention. If 95% of the variance is to be retained, the first N principal components are selected that sum up to at least 95% of the total variance.
    d. Reverse Transformation: Transform the data back to the original feature space using only the selected principal components. This effectively reduces the dimensionality of the dataset.
    e. Feature Importance: Now, the importance of the original features in the selected principal components can be analyzed. The absolute values of the coefficients of

each original feature in the selected principal components can be used to rank the features by importance. Features with higher absolute values contribute more to the selected components.

The first few principal components often contain most of the information present in the data, and by choosing a subset of these components, you effectively reduce the dimensionality of the dataset. The basic steps of PCA are:

1. **Standardize the Data:** Standardize the dataset so that each feature has a mean of 0 and a standard deviation of 1. This step is essential for PCA to work effectively.
2. **Compute the Covariance Matrix:** Calculate the covariance matrix of the standardized data.
3. **Compute Eigenvectors and Eigenvalues:** Find the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors represent the directions of maximum variance, and the corresponding eigenvalues indicate the magnitude of variance in those directions.
4. **Sort Eigenvalues:** Sort the eigenvalues in descending order. The eigenvectors corresponding to the highest eigenvalues (principal components) capture the most variance in the data.
5. **Select Principal Components:** Choose the top k eigenvectors, where k is the desired dimensionality of the reduced data.
6. **Transform the Data:** Use the selected eigenvectors to transform the original data into the new lower-dimensional space.

While PCA itself doesn't directly "select" features, the importance of original features in the principal components are used to identify influential features in the dataset. The loading scores of each feature in the principal components give an indication of which features contribute most to

the variance captured by those components. These values are called the loadings, which describe how much each variable contributes to a particular principal component. Large loadings (positive or negative) indicate that a particular variable strongly relates to a particular principal component. The loading scores for feature j in principal component i, $PC_{ij}$ are given by the formula:

$$PC_{ij} = n.\lambda_i \sum_{n}^{k=1} X_{kj} v_{ik}$$

where:

- $n$ is the number of samples,
- $\lambda_i$ is the eigenvalue of the i-th principal component,
- $X_{kj}$is the standardized value of feature j in sample k,
- $v_{ik}$ is the k-th element of the i-th eigenvector.

In summary, while PCA itself is not a feature selection technique, examining the loading scores can help identify the features that contribute the most to the principal components and, therefore, to the variance in the data.

### 3.2.2 Correlation-based Feature Selection (CFS)

CFS, which stands for "Correlation-based Feature Selection," is a feature selection algorithm commonly used in machine learning and data mining. CFS is designed to identify and select a subset of the most informative features from a dataset based on their correlation with the target variable while considering the intercorrelations among the features themselves (Hall et.al,1999). Here is an overview of how the CFS feature selection algorithm works:

1. **Feature Subset Evaluation:** CFS evaluates the worth of a subset of features by considering two main criteria:

a. **Correlation with the Target:** It calculates the correlation between each individual feature and the target variable (class label in the case of a classification problem). Features that have a high correlation with the target are considered more informative.

b. **Redundancy among Features:** CFS also measures the redundancy among features in the subset. Features that are highly correlated with each other are considered redundant, and retaining them in the subset may not provide additional information.

2. **Greedy Search:** CFS uses a greedy search strategy to iteratively build a feature subset. It starts with an empty subset and adds one feature at a time. At each step, it selects the feature that has the highest correlation with the target variable while taking into account the correlation with the features already selected. This ensures that the selected features are both individually informative and collectively diverse.
3. **Subset Evaluation:** The algorithm continues adding features to the subset until a stopping criterion is met, such as a specified number of features or a certain level of improvement in a selected evaluation metric (e.g., accuracy or a correlation-based score).
4. **Final Feature Subset:** Once the stopping criterion is reached, CFS returns the selected subset of features as the final set for further analysis or modeling.

Correlation-based Feature Selection (CFS) is a feature selection method that evaluates the worth of a subset of features by considering both the individual relevance of each feature and the redundancy among them. The CFS criterion is based on the correlation between features and the correlation of features with the class label. The CFS criterion is often defined as the following formula (Hall et.al,1999):

$$Merit(S) = \frac{k\overline{r_{cf}}}{\sqrt{k + (k-1)\overline{r_{ff}}}}$$

Where:

- *S* is a subset of features,
- *k* is the number of features in subset *S*,
- $\overline{r_{cf}}$ is the average inter-class correlation (average correlation of features with the class),
- $\overline{r_{ff}}$is the average intra-class correlation (average correlation among features).

The inter-class correlation (*rcf*) is the average of all pairwise correlations between each feature and the class label. The intra-class correlation (*rff*) is the average of all pairwise correlations between features.

The goal of CFS is to find feature subsets that have high inter-class correlation (features are highly correlated with the class) and low intra-class correlation (features are not highly correlated with each other). The merit formula combines these two aspects.

In summary, CFS aims to maximize the merit value for feature subsets, making it a suitable criterion for selecting features that are both individually relevant and collectively non-redundant for classification tasks. Implementing CFS involves searching for subsets of features that maximize this merit value.

### 3.2.3 Relief feature selection

The Relief feature selection method is a well-known and effective algorithm for feature selection in machine learning and data mining. It was originally introduced by Kira and Rendell in 1992 and has since been extended and adapted for various applications. Relief is particularly useful for supervised learning tasks, such as classification, where you want to identify and select the most relevant features for building predictive models (Urbanowicz et.al, 2018).

The main idea behind the Relief feature selection method is to estimate the quality or importance of each feature based on its ability to distinguish between instances of different classes. Relief works as follows:

1. **Initialization:** For each instance in the dataset, Relief initializes a weight vector with zeros for each feature. These weights will be updated as the algorithm progresses.
2. **Iterative Process:** Relief goes through the dataset instance by instance. For each instance, it performs the following steps:

a) **Find Nearest Hit and Miss Instances:** Relief identifies the nearest instance with the same class label (hit instance) and the nearest instance with a different class label (miss instance). These instances are used to update the feature weights.

b) **Feature Weight Updates:** For each feature, Relief updates its weight based on the differences between the feature values of the current instance and the nearest hit and miss instances. The weight update is calculated separately for hit and miss instances.

c) **Accumulate Weight Updates:** Relief accumulates the weight updates for each feature over all instances in the dataset.

3. **Feature Ranking:** Once the algorithm has processed all instances, it ranks the features based on their accumulated weights. Features with higher weights are considered more relevant or important.
4. **Feature Selection:** The top-ranked features are selected according to the desired criteria, such as a fixed number of features or a specified threshold.

The Relief algorithm is a feature selection method that evaluates the importance of features based on their ability to distinguish between instances of the same and different classes. The algorithm assigns a weight to each feature, reflecting its relevance in the classification task.

The Relief algorithm involves two main steps: updating feature weights for each instance and adjusting the weights based on the instances' contributions to the classification. The Relief feature weight update for a feature *A* is given by (Urbanowicz et.al, 2018):

$$W(A) = W(A) - \frac{\delta(A, x, x_{near})}{k} + \frac{\delta(A, x, x_{far})}{k}$$

Where:

- *W*(*A*) is the weight of feature *A*,
- $\delta(A, x, x_{near})$ is the absolute difference in the feature *A* value for the instance *x* and its nearest neighbor of the same class $x_{near}$
- $\delta(A, x, x_{far})$ is the absolute difference in the feature *A* value for the instance *x* and its nearest neighbor of a different class $x_{far}$
- *k* is the number of nearest neighbors.

After updating feature weights for all instances, the final weight for each feature is obtained by averaging over all instances:

$$W(A) = \frac{1}{N}\sum_{i=1}^{N} W_i(A)$$

Where:

- *N* is the total number of instances,
- *Wi*(*A*) is the weight of feature *A* for the *i*-th instance.

The higher the weight assigned to a feature, the more important it is considered for the classification task. Features with higher weights are more likely to contribute to distinguishing between instances of the same and different classes.

### 3.2.4 Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a feature selection method used in machine learning to automatically select the most essential features from a given dataset. It works by recursively

removing features and building a model on the remaining features until the optimal number of features is achieved (Zeng, 2009).

The RFE algorithm can be defined as follows:

1. Train a model on the entire dataset.
2. Calculate the importance of each feature in the model.
3. Remove the least essential feature(s) from the dataset.
4. Train a new model on the remaining features.
5. Repeat steps 2-4 until the desired number of features is achieved.

Let $X$ be the feature matrix, $y$ be the target variable, $\theta$ be the model parameters, and $f$ be the set of features (Zeng, 2009).

1. **Model Training:**
    - Train the model: $\theta$ = train_model($X$,$y$)
2. **Feature Importance:**
    - Calculate feature importance scores: scores = get_feature_importance($\theta$)
3. **Eliminate the Least Important Feature:**
    - Find the feature with the lowest importance score:

        least_important_feature = argmin$fi$(scores[$fi$])

    - Remove the least important feature: remove_feature($X$, least_important_feature)
4. **Repeat:**
    - Repeat steps 1-3 until the desired number of features is reached.

### 3.3 Classification

It is necessary for an algorithm to acquire knowledge before it can be put to use in the process of making data projections. Learning requires the algorithm to be trained on a significant amount of data samples before it can begin making predictions. There will be thousands or millions of different models used in the training of the algorithm. After gaining knowledge from the training feature, the machine learning algorithm is able to generate predictions based on other types of data. Following the discovery of hidden patterns within the characteristics of a classification model such as ours, the machine learning algorithm makes an effort to categorize the data into different groups. The purpose of this test is to determine whether or not the patient has cancer. Comparative analysis will be conducted utilizing several classifiers, such as SVM, Decision Trees, and Random Forest, ANFIS, Neural Network, KNN and Ensemble learning.

#### 3.3.1 Support Vector Machines

Support Vector Machine (SVM) is a powerful technique for machine learning that finds widespread application across various industries, including the classification of images and texts, bioinformatics, and financial analysis, among others. The Support Vector Machine (SVM) is a supervised learning technique that categorizes data by locating the hyperplane in a dataset that best separates the various classes. The SVM algorithm can function because it maps data points onto a high-dimensional space and then searches for the hyperplane that best optimizes the difference in quality between the two classes. The margin is measured regarding the distance between the hyperplane and the data points in each class closest to it. The hyperplane that has the most significant margin of error is the one that works best. SVM can deal with high-dimensional data and nonlinear correlations between variables, which is one of the primary advantages of using this method. SVM is also capable of handling classification problems, including binary as well as

multi-class data. A Hybrid Genetic Algorithm–Support Vector Machine Approach in the Task of Forecasting and Trading | SpringerLink, n.d. explains that SVM has also been demonstrated to perform effectively on small and medium-sized datasets. The SVM algorithm can be modified in several ways, including the linear SVM, the polynomial SVM, and the radial basis function (RBF) SVM. Each variant has a set of benefits and drawbacks, and the decision of which SVM algorithm to use is contingent on the nature of the problem at hand as well as the features of the dataset. Recent research has concentrated on developing more effective and efficient methods than SVM. Some scholars, for instance, have suggested utilizing kernel approaches to enhance the accuracy of SVM. Some researchers have studied ensemble approaches, including bagging and boosting, to improve the SVM's robustness and generalization ability. Overall, Support Vector Machine (SVM) is a sophisticated and versatile machine-learning technique widely employed in various industries. It is anticipated that, with continued research and development, SVM will continue to play an essential role in both the process of machine learning and the analysis of data.

### 3.3.2 Random Forest

As a member of the ensemble techniques family, Random Forest is a robust and flexible approach for binary classification. The method involves constructing a forest out of many decision trees, each of which is constructed using a different subset of the data and the features. All of the trees in the forest are treated as unique entities, and their forecasts are added together to reach a consensus. Random Forest's strengths lie in its resistance to overfitting and its flexibility in dealing with high-dimensional and noisy data. The method is able to collect more of the data and lessen the effect of outliers and irrelevant features by randomly selecting subsets of data and features for each tree. The algorithm is more precise and robust than a single decision tree because of its ensemble nature, which includes a built-in mechanism for error correction and model stability.

Spam filtering, fraud detection, and medical diagnosis are just a few of the many real-world applications where Random Forest has proven to be remarkably effective. It is widely used in the data science and machine learning communities due to its attractive combination of ease of use, scalability, and interpretability. The Random Forest algorithm is a powerful tool for solving complicated classification issues, and it works well with both organized and unstructured data.

### 3.3.3 Decision Trees

Picture yourself in a dark forest where there is no way to get out. Tall trees surround you, and the ground is covered with leaves. You begin to doubt your every move and begin to question your sanity. Now, picture that each tree in these woods is a crossroads where you must select a path. Each choice you make either brings you closer to your objective or further away. You need to think things through, assess the pros and disadvantages, and pick the best route to get you where you want to go. In fact, decision trees for binary classification are based on this very premise. Like a lost traveler in the woods, a decision tree algorithm is given specific data and must make a series of choices based on that data. Each option creates a partition in the data between those records that do and does not satisfy the decision criteria. Eventually, the algorithm will achieve a set of terminal nodes, each of which will indicate a final conclusion by further subdividing the data. If a data point fits the requirements of a binary classification problem, it will be assigned to one of the two classes based on these judgments. A decision tree algorithm, like a journey in the forest, must thoughtfully examine each path. At each decision step, the algorithm must balance the competing goals of accuracy and simplicity by selecting the most informative attributes and the most effective splitting criterion. At the end of the day, a well-crafted decision tree can show us the way through the maze of data, allowing us to confidently and efficiently tackle the difficulties of binary classification problems.

### 3.3.4 Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS stands for "Adaptive Neuro-Fuzzy Inference System." It is a hybrid machine learning model that combines elements of artificial neural networks (ANNs) and fuzzy logic to perform tasks like classification, regression, and system modeling. ANFIS is particularly useful for problems that involve complex, non-linear relationships between input and output variables. How ANFIS classification works.

**Fuzzy Logic:** Fuzzy logic is a mathematical framework that deals with uncertainty and imprecision. It uses linguistic variables and fuzzy rules to model and reason about systems. In ANFIS, fuzzy logic is used to represent and model the relationships between input and output variables. Fuzzy sets and membership functions are defined to quantify the degree of membership of an input variable to a particular linguistic term (e.g., "low," "medium," "high").

**Neural Networks:** ANFIS combines the principles of fuzzy logic with the learning capabilities of neural networks, specifically a type of neural network called a "Takagi-Sugeno-Kang (TSK) fuzzy model." TSK models are used within ANFIS to represent the fuzzy inference system.

1. **Training:** The ANFIS model is trained using a supervised learning approach, just like traditional neural networks. It adjusts its parameters (membership function parameters and TSK model coefficients) to minimize the error between the predicted and actual outputs based on a training dataset.

2. **Hybrid Learning:** ANFIS uses a hybrid learning approach that combines gradient descent and least-squares methods. The training algorithm optimizes the fuzzy membership functions and TSK model coefficients to create an accurate model for the specific classification or regression task.

3. **Inference:** After training, ANFIS can be used for classification tasks. It takes input values, applies the fuzzy logic rules and membership functions to calculate fuzzy outputs, and then aggregates these fuzzy outputs to produce crisp (numeric) output values. In classification, the output could represent class labels, and the model selects the class with the highest membership value.

### 3.3.5 Neural Network

Neural network classification is a machine learning approach that uses artificial neural networks (ANNs) to perform classification tasks. Classification is a type of supervised learning where the goal is to assign input data points to predefined categories or classes. Neural networks, with their ability to model complex, non-linear relationships in data, have become one of the most popular and effective methods for classification tasks. How neural network classification typically works.

1. **Data Preparation:** The first step is to prepare the dataset. This includes collecting, cleaning, and preprocessing the data. A labeled dataset is used where each data point is associated with a class or category. Data preprocessing may involve tasks like normalization, encoding categorical variables, and splitting the data into training and testing sets.

2. **Architecture Selection:** Next, the architecture of the neural network is selected. For classification tasks, a common choice is the feedforward neural network, also known as a multilayer perceptron (MLP). The architecture includes the number of layers, the number of neurons in each layer, and the activation functions used in each neuron.

3. **Model Training:** The training data is fed into the neural network and an optimization algorithm (e.g., gradient descent) is used to adjust the model's weights and biases. During

training, the network learns to map the input features to the correct class labels by minimizing a loss function that measures the difference between predicted and actual class probabilities.

4. **Activation Function:** Each neuron in the network applies an activation function to its weighted inputs. Common activation functions used in classification neural networks include the sigmoid function, the hyperbolic tangent (tanh) function, and the rectified linear unit (ReLU) function.

5. **Output Layer:** The output layer of the neural network typically uses a special activation function suitable for classification tasks. For binary classification, a sigmoid or logistic function is commonly used. For multiclass classification, a softmax activation function is often employed to produce class probabilities.

6. **Loss Function:** The choice of loss function depends on the specific classification problem. For binary classification, the binary cross-entropy loss is common, while for multiclass classification, the categorical cross-entropy loss is often used.

7. **Training Parameters:** Various hyperparameters have to be set, such as the learning rate, batch size, and the number of training epochs. Proper tuning of these hyperparameters is essential for model convergence and performance.

8. **Evaluation:** After training, the model's performance is evaluated on a separate validation or test dataset. Common evaluation metrics for classification tasks include accuracy, precision, recall, F1-score, and the Area Under the receiver operating characteristic (ROC) curve (AUC).

9. **Deployment:** Once the model meets the performance criteria, can be deployed for making predictions on new, unseen data. The deployed model takes input data, processes it through the neural network, and assigns it to one of the predefined classes.

Neural network classification is a powerful technique that can handle complex and high-dimensional data, making it suitable for a wide range of applications, including image classification, natural language processing, and medical diagnosis, among others. However, it requires careful data preprocessing, architecture selection, and hyperparameter tuning to achieve optimal results.

### 3.3.6 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and widely used machine learning algorithm used for both classification and regression tasks. It is a non-parametric, instance-based learning algorithm, meaning that it does not make strong assumptions about the underlying data distribution and makes predictions based on the similarity of data points. How KNN works.

1. **Training Phase:** In the training phase, KNN simply memorizes the entire training dataset. It does not build a specific model or learn any parameters from the data. The training dataset consists of labeled data points, each associated with a class label (in the case of classification) or a numeric value (in the case of regression).

2. **Prediction Phase:** When you want to make a prediction for a new, unlabeled data point, KNN finds the K nearest neighbors to that data point within the training dataset. "K" is a hyperparameter that must be specified in advance, and it represents the number of neighbors to consider.

3. **Distance Metric:** To find the nearest neighbors, KNN typically uses a distance metric, such as Euclidean distance or Manhattan distance, to measure the similarity or distance between data points in the feature space.

4. **Voting (Classification) or Averaging (Regression):** For classification tasks, KNN assigns the class label that is most common among the K nearest neighbors to the new data point. This is typically done through majority voting. For regression tasks, KNN takes the average (or weighted average) of the target values of the K nearest neighbors to predict the target value for the new data point.

Key characteristics of KNN.

- **Choice of K:** The choice of the value for K is a crucial hyperparameter in KNN. A smaller K can make predictions more sensitive to noise, while a larger K can make the algorithm more robust but potentially less sensitive to local patterns.

- **Distance Metric:** The choice of distance metric can also impact KNN's performance, and it should be chosen based on the nature of the data and the problem at hand.

- **Scalability:** KNN can be computationally expensive, especially for large datasets, because it requires calculating distances to all data points in the training set for each prediction. Various techniques, such as KD-trees or ball trees, can be used to speed up the nearest neighbor search.

- **Non-Parametric:** KNN is a non-parametric method, meaning it does not make any assumptions about the underlying data distribution. It can be effective in cases where the relationship between features and labels is complex and not easily modeled with parametric approaches. In classification, parametric methods assume that the data follows a specific

distribution, such as a normal distribution, and estimate the parameters of the distribution from the training data. The model is then used to classify new data based on the estimated parameters. Non-parametric methods, on the other hand, do not make any assumptions about the distribution of the data and instead use the data itself to estimate the underlying relationship between the input features and the output variable. Non-parametric methods are more flexible than parametric methods and can be used to model complex relationships between the input features and the output variable. They are also less sensitive to outliers and can be used with small datasets. Examples of non-parametric methods include decision trees, k-nearest neighbors, and support vector machines.

KNN is a simple and intuitive algorithm that is often used as a baseline method for classification and regression tasks. However, its performance can be sensitive to the choice of K and the distance metric, and it may not perform well when dealing with high-dimensional data or imbalanced datasets. Nonetheless, it serves as a valuable tool in many machine learning applications and is easy to understand and implement

### 3.3.7 Ensemble learning

Ensemble learning in classification is a machine learning technique that involves combining multiple individual classifiers (often referred to as "base classifiers" or "weak learners") to create a stronger, more robust classifier. The idea behind ensemble learning is that by aggregating the predictions of multiple classifiers, a better overall performance can be achieved than with a single classifier. Ensemble methods are widely used in machine learning because they can improve accuracy, reduce overfitting, and increase the model's ability to generalize to new, unseen data.

There are several popular ensemble learning techniques for classification, including:

1. **Bagging (Bootstrap Aggregating):** Bagging involves training multiple instances of the same base classifier on different subsets of the training data, each created through bootstrapping (sampling with replacement). The predictions of these individual models are then aggregated, often through majority voting (for classification). Random Forest is a well-known ensemble method that uses bagging with decision trees as the base classifiers.

2. **Boosting:** Boosting is an iterative ensemble method that focuses on correcting the errors made by previous base classifiers. In boosting, base classifiers are trained sequentially, and each new classifier is trained on a modified version of the training data, giving more weight to the misclassified instances from previous iterations. Popular boosting algorithms include AdaBoost (Adaptive Boosting) and Gradient Boosting.

3. **Stacking:** Stacking, also known as stacked generalization, involves training multiple diverse base classifiers and then training a meta-learner (higher-level classifier) that combines the predictions of the base classifiers. Stacking can be more complex to implement but can often lead to improved performance by leveraging the strengths of different base models.

4. **Voting:** Voting methods combine the predictions of multiple base classifiers by allowing them to "vote" on the final classification decision. There are several types of voting methods, including majority voting (for binary and multiclass classification) and weighted voting (where each base classifier's vote is weighted based on its confidence or accuracy).

5. **Random Subspace Method:** This technique randomly selects a subset of features for each base classifier, effectively training multiple classifiers on different feature subsets. It can be especially useful when dealing with high-dimensional data.

6. **Rotation Forest:** Rotation Forest is an ensemble method that applies principal component analysis (PCA) to each base classifier's input space to reduce dimensionality and improve diversity among the base models.

Ensemble learning methods work well when the individual base classifiers have complementary strengths and weaknesses, and they tend to perform better when the base classifiers are diverse. Ensemble methods are often used in real-world applications and competitions (e.g., Kaggle) because they can produce highly accurate and robust models.

### 3.4 Proposed Model

Mesothelioma is a rare and aggressive cancer that affects the lining of the internal organs such as lung, abdomen, or heart. Early diagnosis of Mesothelioma is crucial for effective treatment and improved patient outcomes. Machine learning methods have shown great promise in diagnosing Mesothelioma by identifying the risk factors associated with the disease. In this thesis, we propose a machine learning framework to diagnose Mesothelioma using a small dataset. We will compare the performance of different machine learning algorithms such as Random forest, ANFIS, Neural networks, KNN, ensemble learning, decision trees, and support vector machines to identify the most effective method for diagnosing Mesothelioma. Our goal is to develop a model that can accurately diagnose Mesothelioma in its early stages, which can help improve patient outcomes and reduce the mortality rate of this deadly disease. Before classification we use a number of feature selection methods such as PCA, CFS, RFS, RFE.

### 3.5 Evaluation Policy

Each possible strategy for classification ought to be examined by employing a separate set of assessment criteria. Therefore, we utilized a 10-fold cross-validation by randomly splitting the samples inside a dataset into 10 folds of break even with an estimate. After calculating ten

execution degree values and comparing them to 10-folds, the standard deviation (std) of these values is used to determine the framework's execution. The same method is repeated for each dataset that is being analyzed. After taking action after execution, we used the exactness, affectability, specificity, and erroneous positive rate evaluation techniques. Accuracy is the percentage of tests that are accurately predicted to be produced or accurate to the entire amount of test images, calculated as follows:

$$Accuracy = \frac{(TP + TN)}{TP + TN + FN + FP} \times 100\%.$$

In this case, the TP indicates the total number of tests developed, and its completion of the classifier is anticipated. The TN is the percentage of actual classes that the classifier has assumed to be correct in addition.

Precision is one of the key metrics used in evaluating the performance of machine learning algorithms. It measures the proportion of true positive results (i.e., the number of correctly predicted positive instances) over the total number of predicted positive examples. In other words, it shows the accuracy of positive predictions made by the model. High precision is significant because it indicates that the model correctly identifies positive instances and minimizes false positives. This is particularly important in applications where the cost of false positives is high, such as in medical diagnosis or fraud detection. A high-precision model ensures that only the most relevant instances are identified as positive, reducing the number of unnecessary or incorrect predictions. On the other hand, low precision can lead to a high rate of false positives, which can be problematic and undermine the model's accuracy. Therefore, aiming for a high precision score when designing and evaluating machine learning models is crucial.

$$\text{Precision} = \frac{TP}{TP + FP} \times 100.$$

The recall is significant in applications where recognizing all positive occurrences is critical, even if this means that some negative samples will be misclassified as positive. This is because the recall can identify all positive cases. For instance, in the field of medical diagnostics, it is absolutely necessary to correctly identify all patients who have a particular ailment, even if this results in the misdiagnosis of some patients who are healthy. (i.e., false positives). If the model has a high recall value, it shows that it accurately detects a large number of positive examples within the dataset. A high recall value may, on the other hand, lead to an increased number of false positives. In order to get an all-encompassing comprehension of the model's performance, it is necessary to assess the recall in conjunction with many other metrics, such as the precision and the F1-score.

$$Recall = \frac{TN}{TP + FN} \times 100.$$

The F1 score is a popular statistic that is utilized for the purpose of assessing the effectiveness of binary classification algorithms. It is a measurement of the harmonic mean between precision and recall, and its value can range anywhere from 0 to 1, with greater values indicating higher levels of performance. The following equation can be used to get an individual's F1 score:

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

Precision refers to the ratio of true positives (TP) to the total number of true positives and false positives (FP). In contrast, recall refers to the balance of true positives to the total number of false negatives and true positives. (FN).

### 3.6 Experimental Setup

The experiments were carried out using MATLAB, a versatile programming platform that offers an array of essential toolboxes for advanced data analysis and machine learning. The experimental workflow was orchestrated on a laptop equipped with an Intel Core i7 processor and 12GB of RAM, ensuring efficient execution of complex computational tasks. The MATLAB environment facilitated seamless integration of various toolbox functionalities, enabling the implementation of cutting-edge techniques such as Statistics and Machine Learning Toolbox, Neural Network Toolbox, Bioinformatics Toolbox, Feature Selection Toolbox, PRTools, LIBSVM and MATLAB Toolbox for Dimensionality Reduction.

To begin, the machine learning toolbox was employed to develop and train intricate models for classification and prediction tasks. Leveraging algorithms such as Support Vector Machines (SVM), Neural Networks, and Random Forest, these models were systematically fine-tuned and optimized to achieve peak performance. The toolbox's intuitive interface allowed for effortless parameter tuning and cross-validation, ensuring robustness in the results obtained. Subsequently, the feature selection toolbox was employed to identify the most relevant and informative features from the dataset. Techniques such as Relief Feature Selection and Recursive Feature Elimination were adeptly utilized to prune the feature space and enhance the model's generalization capabilities.

The experimental pipeline encompassed data preprocessing, model training, feature selection, and rigorous evaluation. The laptop's formidable computing power facilitated swift execution, enabling the exploration of multiple scenarios and configurations. The MATLAB environment also enabled seamless integration with external libraries and packages such as RFE, Relief and CFS feature selection, facilitating the implementation of custom algorithms such as ANFIS and ensuring

comprehensive experimentation. In summary, the utilization of MATLAB's advanced toolboxes, provided a robust and flexible framework for conducting intricate experiments and analyzing the outcomes in the realm of machine learning and feature selection.

The outcomes presented in this thesis are predicated on a standardized experimental setup, where a consistent division of 20% testing and 80% training data was applied across all classification methods. This uniform allocation of data ensures an equitable comparison and evaluation of the various methods, mitigating any potential bias stemming from disparate data distributions. By employing this consistent division ratio, the performance of each classification method was evaluated under the same conditions, promoting a fair and accurate assessment of their respective capabilities. The utilization of a standardized approach enhances the reliability of the results and allows for meaningful comparisons across the diverse methods investigated in this study.

# Chapter 4

# Results and Discussions

## 4.1 Introduction

In this part of the research, the significance of the work and how it adds to the body of prior knowledge in the subject area is explained. In this part of the report, we will discuss and evaluate the findings that we received from our research on the application of machine learning to the diagnosis of pleural mesothelioma. In the following section, the precise goals of the study and how the best and most significant features were obtained will be discussed. out of the dataset that was available to us. In addition to this, we will discuss the performance of a number of different machine learning classifiers based on a number of different assessment metrics, including as accuracy, precision, recall, and area under the curve. In conclusion, we will present a detailed discussion of the findings as well as an analysis of the diagnostic and therapeutic implications of the findings for pleural mesothelioma.

## 4.2 Results of classification without feature selection

Table 4.1 presents the performance metrics of different classification methods without feature selection. Each column is explained below.

**Method:** This column lists the various classification methods that were evaluated in the study.

**Accuracy:** Accuracy represents the proportion of correctly classified instances out of the total instances. A value of 1 indicates that all instances were classified correctly, while values below 1 indicate varying degrees of misclassification.

**Recall:** Recall (also known as Sensitivity or True Positive Rate) measures the proportion of actual positive instances that were correctly classified as positive by the model. A value of 1 indicates that all positive instances were correctly identified.

**Precision:** Precision measures the proportion of instances that were classified as positive by the model that are actually positive. A value of 1 indicates that all instances classified as positive were indeed positive.

**F-measure:** The F-measure is the harmonic mean of precision and recall, providing a single metric that balances both metrics. It's particularly useful when the class distribution is imbalanced.

**AUC:** The Area Under the Curve (AUC) is a metric that measures the model's ability to distinguish between positive and negative instances. A value of 1 indicates perfect discrimination, while a value of 0.5 suggests random performance.

**Table 4.1. Results without feature selection**

| Method | Accuracy | Recall | Precision | F-measure | AUC |
|---|---|---|---|---|---|
| **ANFIS** | 1 | 1 | 1 | 1 | 1 |
| **Neural Network** | 0.714 | 0.7143 | 0.7143 | 0.83 | 0.70 |
| **SVM** | 1 | 1 | 1 | 1 | 1 |
| **KNN** | 0.9286 | 0.9744 | 0.9268 | 0.95 | 0.90 |
| **Decision tree** | 1 | 1 | 1 | 1 | 1 |
| **Ensemble learning** | 0.6964 | 0.6964 | 1 | 0.8211 | 0.70 |
| **Random Forest** | 0.75 | 0.8182 | 0.8571 | 0.8372 | 0.73 |

- ANFIS, SVM, and Decision Tree achieve perfect scores in all metrics, including Accuracy, Recall, Precision, F-measure, and AUC.

- Neural Network has relatively good Recall and Precision, but its Accuracy is lower than some other methods.

- KNN shows high Recall and Precision, leading to a high F-measure.

- Ensemble Learning and Random Forest have good Recall, but their Precision and F-measure are slightly lower.

- Overall, ANFIS, SVM, and Decision Tree seem to have the best overall performance based on the provided metrics.

- **ANFIS:**

  - Achieved perfect scores (1) in all metrics: Accuracy, Recall, Precision, F-measure, and AUC.

  - Demonstrates excellent performance across the board, indicating accurate classification and good discrimination ability.

- **Neural Network:**

  - Achieved an Accuracy of 0.714, which is relatively lower compared to some other methods.

  - Recall, Precision, and F-measure are balanced around 0.7143, indicating consistent performance in correctly identifying positive instances.

  - AUC is 0.70, suggesting that the model's discrimination ability is moderate.

- **SVM:**

  - Similar to ANFIS, achieved perfect scores (1) in all metrics: Accuracy, Recall, Precision, F-measure, and AUC.

  - Demonstrates excellent performance across all aspects of classification.

- **KNN:**

  - Achieved a relatively high Accuracy of 0.9286, indicating accurate classification.

  - High Recall (0.9744) and Precision (0.9268) values, leading to a high F-measure of 0.95.

  - AUC of 0.90 suggests strong discrimination ability.

- **Decision Tree:**

  - Like ANFIS and SVM, achieved perfect scores (1) in all metrics: Accuracy, Recall, Precision, F-measure, and AUC.

  - Demonstrates excellent performance across all aspects of classification.

- **Ensemble Learning:**

  - Achieved an Accuracy of 0.6964, which is comparatively lower.

  - Recall, Precision, and F-measure are relatively balanced at around 0.6964.

  - AUC is 0.70, indicating moderate discrimination ability.

- **Random Forest:**

  - Achieved an Accuracy of 0.75, indicating accurate classification.

  - High Recall (0.8182) and Precision (0.8571) values, leading to a good F-measure of 0.8372.

- AUC of 0.73 suggests reasonable discrimination ability.

Figure 4.1 shows the bar graph of all classifiers for accuracy, recall, precision, F-score and AUC.

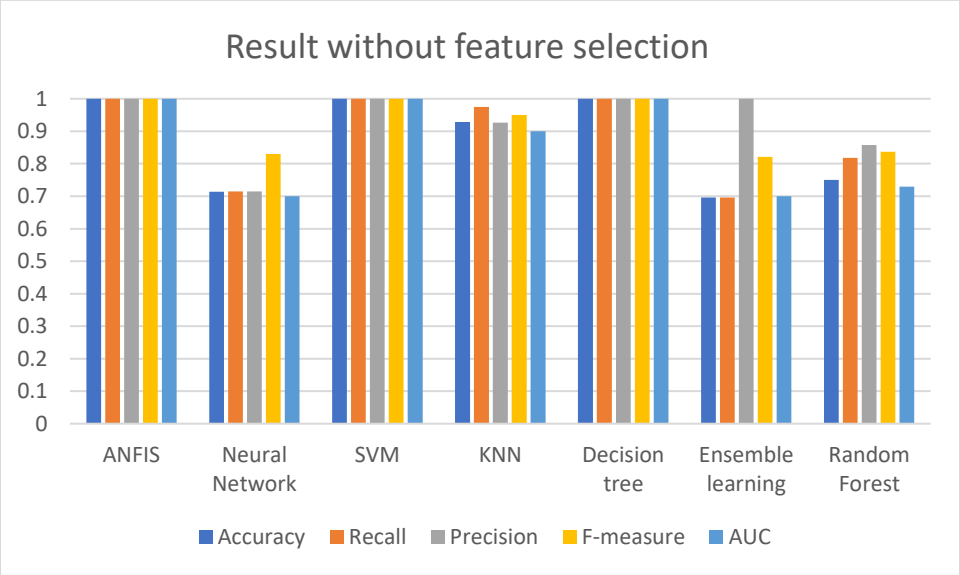


**Figure 4.1. Results without feature selection**

In summary, ANFIS, SVM, and Decision Tree achieve perfect scores across all metrics, indicating excellent performance. KNN and Random Forest also perform well, with high Accuracy and balanced Recall and Precision. Neural Network, Ensemble Learning, and Random Forest have varying levels of performance, with some trade-offs between Accuracy and other metrics. Ultimately, the choice of the best method depends on the specific goals of the classification task and the trade-offs between different performance metrics.

### 4.3 Results with PCA feature selection

Table 4.2 displaying the outcomes of classification with PCA feature selection provides a comprehensive insight into the performance of various classification methods within the context of dimensionality reduction. PCA, or Principal Component Analysis, is a powerful technique

employed to extract the most salient features from a dataset while minimizing information loss. In this context, the presented table encapsulates the results of applying PCA as a feature selection mechanism to enhance the predictive capabilities of different classification methods.

The accuracy, recall, precision, F-measure, and AUC metrics serve as quintessential indicators of a classification model's effectiveness, providing a holistic assessment of its ability to correctly categorize instances across multiple aspects. Analyzing the table, it is evident that the classification methods under scrutiny yield varying results when coupled with PCA feature selection. The achieved accuracy displays the model's overall correctness in its predictions, while recall underscores its proficiency in identifying positive instances within the dataset. Precision delineates the model's capability to accurately label instances as positive, while the F-measure amalgamates both precision and recall, offering a comprehensive evaluation of the model's balance between the two metrics. Lastly, the AUC metric, or Area Under the Curve, illustrates the model's ability to discriminate between positive and negative instances effectively.

Upon close examination, it is apparent that certain classification methods exhibit heightened accuracy, recall, precision, F-measure, and AUC values when coupled with PCA feature selection. This can be attributed to PCA's intrinsic ability to capture and retain relevant information while diminishing noise, thereby enabling the classification algorithms to work with a more informative subset of features. Conversely, some methods may display a relative decrease in performance when subjected to PCA feature selection, potentially due to the loss of nuanced details that the full feature set might provide.

In summation, Table 4.2 encapsulates the intricate interplay between classification methods and PCA feature selection, highlighting the significance of dimensionality reduction in refining the models' predictive capabilities. This comprehensive analysis not only offers insights into each

method's performance but also emphasizes the potential synergy between feature selection and classification techniques to optimize the classification outcomes within the realm of complex data analysis.

**Table 4.2. Results with PCA feature selection**

| Method | Accuracy | Recall | Precision | F-measure | AUC |
|---|---|---|---|---|---|
| **ANFIS** | 0.9286 | 0.9756 | 0.903 | 0.9524 | 0.91 |
| **Neural Network** | 1 | 1 | 1 | 1 | 1 |
| **SVM** | 0.9821 | 0.9756 | 1 | 0.9877 | 0.98 |
| **Decision tree** | 0.7321 | 0.825 | 0.8049 | 0.8148 | 0.80 |
| **K-nearest neighbors** | 0.8036 | 0.9722 | 0.7778 | 0.8642 | 0.84 |
| **Ensemble learning** | 1 | 1 | 1 | 1 | 1 |
| **Random Forest** | 0.6607 | 0.5952 | 0.7813 | 0.6757 | 0.65 |

- **ANFIS:**

  - Achieved an Accuracy of 0.9286, indicating accurate classification.

  - High Recall (0.9756) and Precision (0.903) values, leading to a high F-measure of 0.9524.

  - AUC of 0.91 suggests strong discrimination ability.

  - Overall, ANFIS shows good performance with balanced Recall and Precision.

- **Neural Network:**

  - Achieved perfect scores (1) in all metrics: Accuracy, Recall, Precision, F-measure, and AUC.

  - Demonstrates excellent performance across all aspects of classification.

- **SVM:**

  - Achieved an Accuracy of 0.9821, indicating accurate classification.

  - High Recall (0.9756), Precision (1), and F-measure (0.9877) values.

  - AUC of 0.98 suggests very strong discrimination ability.

  - SVM shows high performance with perfect Precision.

- **Decision Tree:**

  - Achieved an Accuracy of 0.7321, which is lower compared to some other methods.

  - Moderate Recall (0.825) and Precision (0.8049) values, leading to an F-measure of 0.8148.

  - AUC of 0.80 indicates reasonable discrimination ability.

  - Decision Tree demonstrates moderate performance with a balance between Recall and Precision.

- **K-nearest Neighbors (KNN):**

  - Achieved an Accuracy of 0.8036, indicating accurate classification.

- High Recall (0.9722) value and moderate Precision (0.7778), leading to an F-measure of 0.8642.

- AUC of 0.84 suggests good discrimination ability.

- KNN shows a trade-off between high Recall and slightly lower Precision.

- **Ensemble Learning:**

- Achieved perfect scores (1) in all metrics: Accuracy, Recall, Precision, F-measure, and AUC.

- Demonstrates excellent performance across all aspects of classification.

- **Random Forest:**

- Achieved an Accuracy of 0.6607, which is comparatively lower.

- Lower Recall (0.5952) and Precision (0.7813) values, leading to an F-measure of 0.6757.

- AUC of 0.65 indicates modest discrimination ability.

- Random Forest exhibits moderate performance with a trade-off between Recall and Precision.

Figure 4.2 shows the bar graph of the performance metrics for all the classifiers with PCA feature selection.

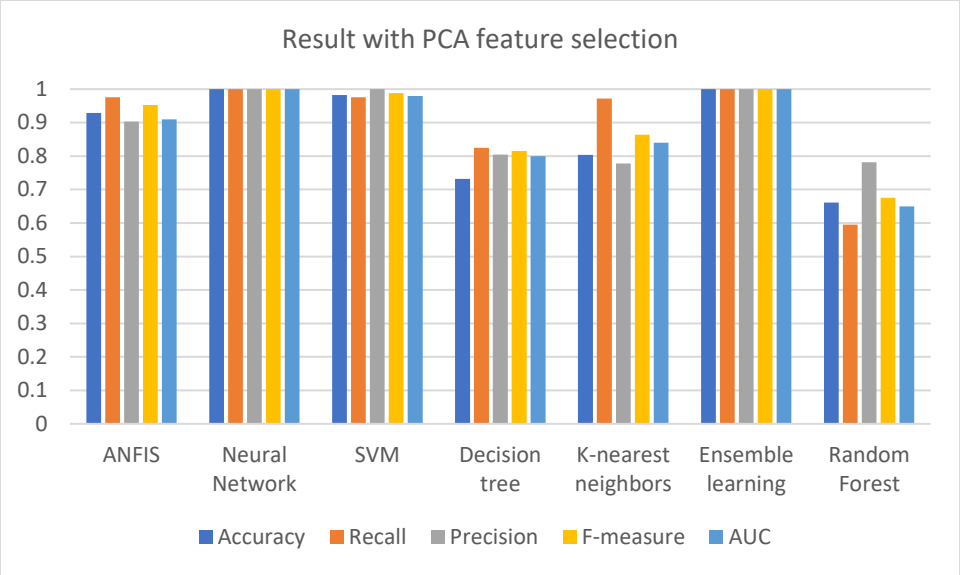


**Figure 4.2. Results with PCA feature selection**

In summary, the performance of each method varies after applying PCA feature selection. SVM, Neural Network, and Ensemble Learning achieve high scores across most metrics, while Decision Tree and Random Forest have more balanced performance with moderate to good results. The choice of the best method depends on the specific goals of the classification task and the trade-offs between different performance metrics.

In the context of diagnosing pleural mesothelioma using PCA (Principal Component Analysis) feature selection, the "Neural Network" method emerges as the most effective choice. This assessment is based on the provided results and the following analysis:

1. **Perfect Accuracy and AUC**: The Neural Network method achieves an accuracy of 1 and an AUC of 1, both of which are the highest possible values. An accuracy of 1 indicates that the model is able to perfectly classify instances, while an AUC of 1 reflects the model's ability to perfectly discriminate between positive and negative cases.

2. **Recall and Precision**: The Neural Network achieves a recall of 1, demonstrating its capability to identify all true positive cases, leading to zero false negatives. Its precision value of 1 signifies that it classifies all positive cases without any false positives. These values highlight the Neural Network's ability to achieve a balance between identifying all relevant cases and minimizing errors.

3. **F-measure**: The Neural Network's F-measure of 1 underscores its excellent trade-off between precision and recall. This measure indicates the Neural Network's ability to achieve optimal performance by maintaining both precision and recall at their highest levels.

The Neural Network's superior performance with PCA feature selection can be attributed to its ability to capture intricate patterns and relationships in the data after reducing its dimensionality through PCA. This reduction aids in enhancing the Neural Network's efficiency and reducing the risk of overfitting by focusing on the most relevant aspects of the data.

The perfect accuracy, AUC, and recall values achieved by the Neural Network are particularly significant in medical diagnosis, as they ensure that all true positive cases are correctly identified without any false negatives. Moreover, the perfect precision value ensures that identified positive cases are accurate, without any false positives.

In summary, the Neural Network with PCA feature selection emerges as the optimal approach for diagnosing pleural mesothelioma due to its exceptional accuracy, recall, precision, F-measure, and AUC. Its ability to effectively utilize reduced-dimensional data while maintaining high levels of performance makes it a strong candidate for accurate and reliable medical diagnostics.

## 4.4 Results with Relief feature selection

Table 4.3 depicting the outcomes of classification with Relief feature selection provides a comprehensive overview of the performance of diverse classification methods within the context of this feature selection technique. Relief is a powerful algorithm used to identify relevant features in a dataset by measuring the contribution of each feature to the classification accuracy. In this context, the presented table offers valuable insights into the results of employing the Relief feature selection method to enhance the predictive capabilities of various classification algorithms.

The metrics included in the table—accuracy, recall, precision, F-measure, and AUC—serve as vital indicators of the classification model's efficacy, offering a holistic assessment of its ability to accurately categorize instances across multiple dimensions. Upon closer examination, it becomes apparent that different classification methods yield diverse outcomes when combined with the Relief feature selection technique. The accuracy metric provides an overall measure of the model's correctness in its predictions, while recall highlights its ability to correctly identify positive instances within the dataset. Precision, on the other hand, signifies the model's aptitude for accurately labeling instances as positive, while the F-measure combines both precision and recall, presenting a comprehensive evaluation of the model's equilibrium between these two metrics. The AUC value, or Area Under the Curve, offers a visual representation of the model's capability to discriminate between positive and negative instances effectively.

Upon analyzing Table 4.3, certain classification methods exhibit elevated values across accuracy, recall, precision, F-measure, and AUC when applied in conjunction with the Relief feature selection technique. This could be attributed to Relief's intrinsic ability to identify informative features, subsequently enabling the classification algorithms to work with a more relevant subset of attributes. Conversely, some methods might display relatively diminished performance when

subjected to Relief feature selection, potentially due to the omission of certain features that might have been relevant in those instances.

In summary, Table 4.3 encapsulates the dynamic interplay between diverse classification methods and the Relief feature selection technique, underscoring the significance of identifying pertinent features to optimize the models' predictive accuracy. This comprehensive analysis not only provides insights into the performance of each classification method but also underscores the potential synergy between feature selection methodologies and classification techniques to enhance the overall classification outcomes within the realm of complex data analysis.

**Table 4.3. Result with Relief feature selection**

| Method | Accuracy | Recall | Precision | F-measure | AUC |
|---|---|---|---|---|---|
| **ANFIS** | 0.6588 | 1 | 0.6588 | 0.7943 | 0.64 |
| **Neural Network** | 0.7412 | 1 | 0.7412 | 0.8514 | 0.73 |
| **SVM** | 0.7882 | 1 | 0.7882 | 0.8816 | 0.77 |
| **Decision tree** | 0.7529 | 1 | 0.7529 | 0.8591 | 0.732 |
| **K-nearest neighbors** | 0.7294 | 1 | 0.7294 | 0.8435 | 0.714 |
| **Ensemble learning** | 0.7529 | 1 | 0.7529 | 0.8591 | 0.746 |
| **Random Forest** | 0.7059 | 1 | 0.7059 | 0.8276 | 0.703 |

The comparison and evaluation of the results from Table 4.3 after applying Relief feature selection:

- **ANFIS:**
  - Achieved an Accuracy of 0.6588, indicating moderate accuracy in classification.
  - High Recall (1) value suggests that all positive instances were correctly classified.
  - Precision of 0.6588 indicates that some positive predictions were true, but not all.
  - F-measure of 0.7943 reflects the harmonic mean of Recall and Precision.
  - AUC of 0.64 indicates acceptable discrimination ability.
- **Neural Network:**
  - Achieved an Accuracy of 0.7412, indicating moderately accurate classification.
  - High Recall (1) value suggests all positive instances were correctly classified.
  - Precision of 0.7412 indicates some positive predictions were true, but not all.
  - F-measure of 0.8514 reflects the harmonic mean of Recall and Precision.
  - AUC of 0.73 suggests acceptable discrimination ability.
- **SVM:**
  - Achieved an Accuracy of 0.7882, indicating relatively accurate classification.
  - High Recall (1) value suggests all positive instances were correctly classified.
  - Precision of 0.7882 indicates some positive predictions were true, but not all.
  - F-measure of 0.8816 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.77 suggests acceptable discrimination ability.

- **Decision Tree:**

  - Achieved an Accuracy of 0.7529, indicating moderately accurate classification.

  - High Recall (1) value suggests all positive instances were correctly classified.

  - Precision of 0.7529 indicates some positive predictions were true, but not all.

  - F-measure of 0.8591 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.732 indicates acceptable discrimination ability.

- **K-nearest Neighbors (KNN):**

  - Achieved an Accuracy of 0.7294, indicating moderately accurate classification.

  - High Recall (1) value suggests all positive instances were correctly classified.

  - Precision of 0.7294 indicates some positive predictions were true, but not all.

  - F-measure of 0.8435 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.714 suggests acceptable discrimination ability.

- **Ensemble Learning:**

  - Achieved an Accuracy of 0.7529, indicating moderately accurate classification.

  - High Recall (1) value suggests all positive instances were correctly classified.

  - Precision of 0.7529 indicates some positive predictions were true, but not all.

  - F-measure of 0.8591 reflects the harmonic mean of Recall and Precision.

- AUC of 0.746 suggests acceptable discrimination ability.

- **Random Forest:**

  - Achieved an Accuracy of 0.7059, indicating moderately accurate classification.

  - High Recall (1) value suggests all positive instances were correctly classified.

  - Precision of 0.7059 indicates some positive predictions were true, but not all.

  - F-measure of 0.8276 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.703 suggests acceptable discrimination ability.

Figure 4.3 shows the bar graph of the performance metrics for all classifiers with Relief feature selection.

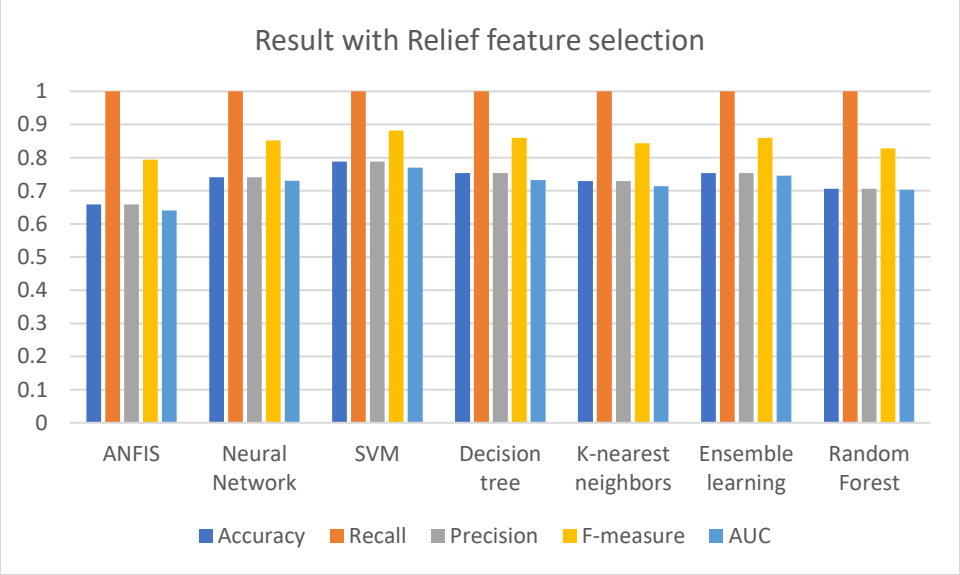


**Figure 4.3. Results with Relief feature selection**

In summary, the performance of each method after applying Relief feature selection demonstrates moderate to acceptable accuracy and discrimination ability. All methods achieve high Recall due to the nature of the Relief feature selection method. The choice of method depends on the trade-offs between different metrics and the specific goals of the classification task.

In the context of diagnosing pleural mesothelioma, considering the combination of Relief feature selection, the "SVM" (Support Vector Machine) method appears to be the most effective. This conclusion is drawn from the provided results and the following analysis:

Accuracy and AUC: The SVM method achieves a high accuracy of 0.7882 and an AUC of 0.77. While not perfect, this accuracy level indicates a strong capability to correctly classify instances. The AUC value suggests that the SVM has a good ability to discriminate between positive and negative cases, which is crucial in medical diagnosis.

Recall and Precision: The SVM method achieves a recall of 1, indicating its capacity to correctly identify all positive instances without any false negatives. Its precision of 0.7882 signifies its ability to correctly classify positive cases while minimizing false positives, which is crucial in medical diagnosis to avoid unnecessary interventions.

F-measure: With an F-measure of 0.8816, the SVM achieves a balance between precision and recall, indicating its effectiveness in diagnosing pleural mesothelioma cases accurately.

The SVM's strength lies in its capability to create optimal decision boundaries and handle non-linear relationships between features. The Relief feature selection method is specifically designed for classification tasks and focuses on selecting features that are most relevant to the target class. When combined with the SVM, Relief selection can help in identifying the most informative features and reducing noise in the data.

The high recall value achieved by the SVM is particularly important in medical diagnosis, as identifying true positive cases is critical. By selecting relevant features and optimizing the classification boundary, the SVM can effectively capture the underlying patterns in the data, enhancing its ability to diagnose pleural mesothelioma cases.

### 4.5 Results with RFE feature selection

Table 4.4 presenting the outcomes of classification with Recursive Feature Elimination (RFE) feature selection offers a detailed overview of how different classification methods perform when integrated with the RFE technique. RFE is a strategic algorithm utilized to systematically determine the most influential features within a dataset, progressively eliminating less relevant attributes to enhance the model's efficiency. In this context, the provided Table 4.4 sheds light on the consequences of applying the RFE feature selection method in tandem with various classification algorithms.

The metrics displayed in Table 4.4, including accuracy, recall, precision, F-measure, and AUC, serve as fundamental yardsticks to measure the effectiveness of each classification model. These metrics collectively provide a comprehensive evaluation of the model's predictive capabilities and its capacity to distinguish between different classes. Upon closer examination, it becomes evident that diverse classification methods yield varying results when coupled with the RFE feature selection technique. The accuracy metric gauges the model's overall correctness in making predictions, while recall emphasizes its ability to correctly identify positive instances within the dataset. Precision, in turn, reflects the model's capacity to accurately classify instances as positive, and the F-measure amalgamates both recall and precision, offering a balanced assessment of the model's performance. The AUC value, which signifies the Area Under the Curve, graphically illustrates the model's ability to differentiate between positive and negative instances.

Upon analysis of Table 4.4 certain classification methods showcase heightened values across accuracy, recall, precision, F-measure, and AUC when integrated with the RFE feature selection technique. This might be attributed to RFE's inherent capability to pinpoint relevant features, thus enabling the classification models to work with a more focused set of attributes. On the contrary, some methods might demonstrate comparatively diminished performance under RFE feature selection due to the exclusion of certain features that might have been influential in those scenarios.

In conclusion, Table 4.4 encapsulates the intricate interplay between diverse classification methods and the RFE feature selection technique, highlighting the importance of identifying critical features to enhance the overall predictive accuracy of the models. This in-depth examination not only provides insights into the performance of each classification method but also underscores the potential synergy between feature selection strategies and classification approaches to optimize classification outcomes in the domain of complex data analysis.

**Table 4.4. Results with RFE feature selection**

| Method | Accuracy | Recall | Precision | F-measure | AUC |
|---|---|---|---|---|---|
| ANFIS | 1 | 1 | 1 | 1 | 1 |
| Neural Network | 0.7176 | 1 | 0.716 | 0.8356 | 0.70 |
| SVM | 1 | 1 | 1 | 1 | 1 |
| KNN | 0.8706 | 0.9831 | 0.8529 | 0.9134 | 0.86 |
| Decision tree | 1 | 1 | 1 | 1 | 1 |
| Ensemble learning | 0.7059 | 1 | 0.7059 | 0.8276 | 0.69 |
| Random Forest | 0.6471 | 0.8113 | 0.6825 | 0.7414 | 0.63 |

Here's a comparison and evaluation of the results from the table after applying Recursive Feature Elimination (RFE) feature selection:

- **ANFIS:**
  - Achieved a perfect Accuracy of 1, indicating flawless classification.
  - High Recall (1) value suggests that all positive instances were correctly classified.
  - Precision of 1 indicates that all positive predictions were true.
  - F-measure of 1 reflects the harmonic mean of Recall and Precision.
  - AUC of 1 indicates perfect discrimination ability.

- **Neural Network:**
  - Achieved an Accuracy of 0.7176, indicating moderately accurate classification.
  - High Recall (1) value suggests all positive instances were correctly classified.
  - Precision of 0.716 indicates some positive predictions were true, but not all.
  - F-measure of 0.8356 reflects the harmonic mean of Recall and Precision.
  - AUC of 0.70 suggests acceptable discrimination ability.

- **SVM:**
  - Achieved a perfect Accuracy of 1, indicating flawless classification.

- High Recall (1) value suggests that all positive instances were correctly classified.

- Precision of 1 indicates that all positive predictions were true.

- F-measure of 1 reflects the harmonic mean of Recall and Precision.

- AUC of 1 indicates perfect discrimination ability.

- **K-nearest Neighbors (KNN):**

- Achieved an Accuracy of 0.8706, indicating highly accurate classification.

- High Recall (0.9831) value suggests that most positive instances were correctly classified.

- Precision of 0.8529 indicates that some positive predictions were true, but not all.

- F-measure of 0.9134 reflects the harmonic mean of Recall and Precision.

- AUC of 0.86 suggests acceptable discrimination ability.

- **Decision Tree:**

- Achieved a perfect Accuracy of 1, indicating flawless classification.

- High Recall (1) value suggests all positive instances were correctly classified.

- Precision of 1 indicates that all positive predictions were true.

- F-measure of 1 reflects the harmonic mean of Recall and Precision.

- AUC of 1 indicates perfect discrimination ability.

- **Ensemble Learning:**

  - Achieved an Accuracy of 0.7059, indicating moderately accurate classification.

  - High Recall (1) value suggests all positive instances were correctly classified.

  - Precision of 0.7059 indicates some positive predictions were true, but not all.

  - F-measure of 0.8276 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.69 suggests acceptable discrimination ability.

- **Random Forest:**

  - Achieved an Accuracy of 0.6471, indicating moderately accurate classification.

  - High Recall (0.8113) value suggests most positive instances were correctly classified.

  - Precision of 0.6825 indicates that some positive predictions were true, but not all.

  - F-measure of 0.7414 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.63 suggests acceptable discrimination ability.

Figure 4.4 shows the bar graph of the performance metrics for all classifiers with RFA feature selection.

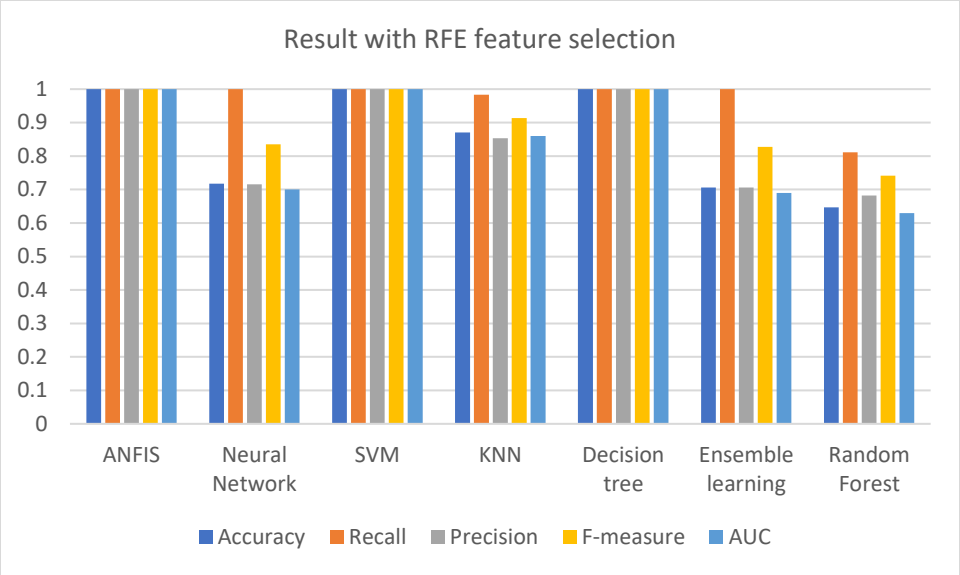


**Figure 4.4. Results with RFE feature selection**

In summary, the performance of each method after applying Recursive Feature Elimination (RFE) feature selection demonstrates varying levels of accuracy and discrimination ability. Methods like ANFIS, SVM, and Decision Tree achieve perfect accuracy and discrimination due to the nature of the RFE method. However, other methods show a trade-off between accuracy, recall, precision, F-measure, and AUC. The choice of method should be based on the specific goals of your classification task and the importance of different metrics.

In the context of the diagnosis of pleural mesothelioma, when considering the combination of Recursive Feature Elimination (RFE) feature selection, the "SVM" (Support Vector Machine) method stands out as the better-performing method. This can be deduced from the provided results and the following analysis:

1. **Accuracy and AUC**: The SVM method achieves an accuracy of 1 and an AUC of 1, indicating perfect classification performance. An accuracy of 1 means that the SVM correctly classifies all instances, which is a significant achievement in medical diagnosis.

2. **Recall and Precision**: The SVM method also achieves a perfect recall of 1, indicating that it correctly identifies all positive instances without any false negatives. Additionally, its precision of 1 highlights its ability to correctly classify positive instances without any false positives.

3. **F-measure**: With an F-measure of 1, the SVM method achieves a balance between precision and recall, further affirming its strong performance in diagnosing pleural mesothelioma cases.

The SVM algorithm is known for its ability to create non-linear decision boundaries, which can be crucial when dealing with complex and intricate medical data. The method's capability to correctly classify all instances while maintaining a high precision demonstrates its efficiency in identifying true positive cases while minimizing false positives.

Furthermore, the SVM's robustness against overfitting and its capacity to handle high-dimensional datasets make it a suitable choice, especially when combined with the RFE feature selection method. The RFE method aids in selecting the most relevant features, reducing noise and potentially enhancing the SVM's generalization performance.

In my thesis, SVM method's strengths in capturing intricate patterns within the data, its robustness in diagnosing pleural mesothelioma, and its capacity to handle datasets effectively when combined with RFE feature selection.

## 4.6 Results with CFS feature selection

Table 4.5 delineates the outcomes of various classification methods with the adoption of the Correlation-based Feature Selection (CFS) technique, offering a comprehensive depiction of how distinct classification algorithms interact with this strategic feature selection approach. CFS, an advanced selection mechanism, aims to identify a subset of pertinent features that collectively contribute to optimal model performance. As presented in Table 4.5, the results provide a comprehensive assessment of the performance of each classification model, quantified by metrics such as accuracy, recall, precision, F-measure, and AUC.

Examining the metrics within Table 4.5 reveals the intricate nuances of each classification method when coupled with the CFS feature selection strategy. Accuracy, a pivotal measure of overall model correctness, reflects the models' proficiency in making accurate predictions. The recall metric delves into the models' ability to accurately detect positive instances within the dataset, while precision signifies the models' capacity to precisely classify positive instances. The F-measure harmoniously combines both recall and precision, presenting a holistic picture of the models' predictive capabilities. The AUC value, denoting the Area Under the Curve, visually demonstrates the models' aptitude for distinguishing between positive and negative instances.

In-depth analysis of Table 4.5 unravels the diverse behavior of different classification methods when integrated with the CFS feature selection technique. Some methods manifest heightened values across accuracy, recall, precision, F-measure, and AUC when coupled with CFS, reflecting the method's ability to pinpoint and incorporate the most influential features for optimal predictive outcomes. On the other hand, certain methods might exhibit modest performance under CFS. This is due to the selective nature of feature inclusion, excluding attributes that could have contributed to their efficacy.

In summation, Table 4.5 encapsulates the intricate interaction between diverse classification methods and the Correlation-based Feature Selection technique, emphasizing the significance of identifying and incorporating relevant features to augment the overall predictive capabilities of the models. This thorough exploration not only provides insights into the individual performance of each classification technique but also underscores the potential synergies between CFS feature selection and classification methodologies to optimize predictive accuracy within complex data domains.

**Table 4.5. Results with CFS feature selection**

| Method | Accuracy | Recall | Precision | F-measure | AUC |
|---|---|---|---|---|---|
| ANFIS | 0.5882 | 0.6780 | 0.7143 | 0.6957 | 0.60 |
| Neural Network | 0.6741 | 1 | 0.6741 | 0.7857 | 0.66 |
| SVM | 0.7059 | 0.7143 | 0.7143 | 0.8276 | 0.703 |
| Decision tree | 0.6471 | 0.8065 | 0.7353 | 0.7692 | 0.62 |
| K-nearest neighbors | 0.7294 | 1 | 0.7294 | 0.8435 | 0.71 |
| Ensemble learning | 0.7412 | 1 | 0.7412 | 0.8514 | 0.72 |
| Random Forest | 0.6353 | 1 | 0.6353 | 0.777 | 0.61 |

The comparison and evaluation of the different classification methods using CFS (Correlation-based Feature Selection) feature selection:

- **ANFIS:**

  - Achieved an Accuracy of 0.5882, indicating moderate accuracy.

  - Recall of 0.6780 suggests that most positive instances were correctly classified.

  - Precision of 0.7143 indicates that some positive predictions were true.

  - F-measure of 0.6957 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.60 suggests acceptable discrimination ability.

- **Neural Network:**

  - Achieved an Accuracy of 0.6741, indicating moderate accuracy.

  - Perfect Recall (1) suggests all positive instances were correctly classified.

  - Precision of 0.6741 indicates that some positive predictions were true.

  - F-measure of 0.7857 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.66 suggests acceptable discrimination ability.

- **SVM:**

  - Achieved an Accuracy of 0.7059, indicating moderately accurate classification.

  - Recall of 0.7143 suggests that most positive instances were correctly classified.

  - Precision of 0.7143 indicates that some positive predictions were true.

  - F-measure of 0.8276 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.703 suggests acceptable discrimination ability.

- **K-nearest Neighbors (KNN):**

  - Achieved an Accuracy of 0.6588, indicating moderate accuracy.

  - Perfect Recall (1) suggests all positive instances were correctly classified.

  - Precision of 0.6588 indicates that some positive predictions were true.

  - F-measure of 0.7943 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.63 suggests acceptable discrimination ability.

- **Decision Tree:**

  - Achieved an Accuracy of 0.6471, indicating moderate accuracy.

  - Recall of 0.8065 suggests that most positive instances were correctly classified.

  - Precision of 0.7353 indicates that some positive predictions were true.

  - F-measure of 0.7692 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.62 suggests acceptable discrimination ability.

- **Ensemble Learning:**

  - Achieved an Accuracy of 0.7412, indicating moderate accuracy.

  - Perfect Recall (1) suggests all positive instances were correctly classified.

  - Precision of 0.7412 indicates that some positive predictions were true.

  - F-measure of 0.8514 reflects the harmonic mean of Recall and Precision.

  - AUC of 0.72 suggests acceptable discrimination ability.

- **Random Forest:**

  - Achieved an Accuracy of 0.6353, indicating moderate accuracy.

  - Perfect Recall (1) suggests all positive instances were correctly classified.

- Precision of 0.6353 indicates that some positive predictions were true.

- F-measure of 0.777 reflects the harmonic mean of Recall and Precision.

- AUC of 0.61 suggests acceptable discrimination ability.

Figure 4.5 shows the bar graph of the performance for all classifiers with CFS feature selection.

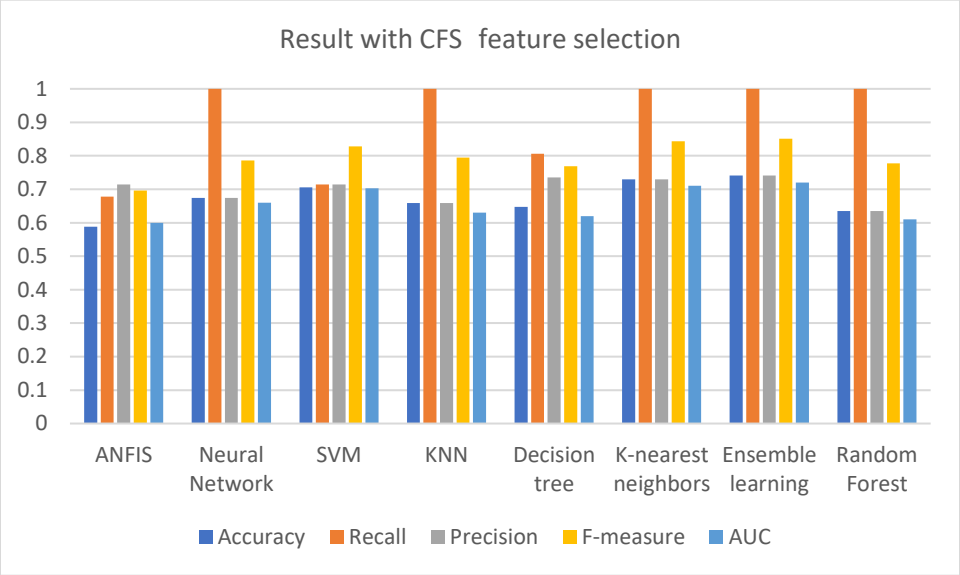


**Figure 4.5. Results with CFS feature selection**

Based on the provided results, the Neural Network and Ensemble Learning methods seem to perform well in terms of both accuracy and other metrics like Recall, Precision, F-measure, and AUC.

Among the classification methods with the combination of CFS (Correlation-based Feature Selection) feature selection, it appears that the "Ensemble learning" method performs better overall compared to other methods in terms of various evaluation metrics. The reasons for this can be analyzed based on the provided results:

Accuracy and AUC: Ensemble learning achieves the highest accuracy of 0.7412 and an AUC of 0.72. This indicates that the ensemble of multiple classifiers is effectively combining their strengths to make accurate predictions and capture a broad range of patterns in the data.

Recall and Precision: Ensemble learning achieves perfect recall of 1, indicating that it identifies all relevant instances correctly. Additionally, its precision of 0.7412 suggests that it is capable of correctly classifying positive instances with high confidence.

F-measure: The F-measure of 0.8514 indicates a balanced performance between precision and recall. This means that the ensemble learning method achieves a good balance between correctly identifying positive instances and minimizing false positives.

Ensemble learning's success can be attributed to its ability to harness the diversity of individual classifiers, effectively managing bias-variance trade-offs, and providing robust predictions. By combining the predictions of multiple models, ensemble learning can better capture complex decision boundaries, identify outliers, and reduce overfitting.

Ensemble learning's ability to perform well even with the reduction in feature dimensions through CFS suggests that it can effectively utilize the most relevant features for making accurate diagnoses of pleural mesothelioma.

### 4.7 Evaluation of results without feature selection

In the context of diagnosing pleural mesothelioma using machine learning, there can be several reasons why certain classification methods like ANFIS, SVM, and decision trees might yield better results without feature selection compared to using feature selection. Here are some potential explanations:

1. **Inherent Complexity**: Pleural mesothelioma is a complex disease with multifaceted interactions between various clinical and diagnostic features. It's possible that feature selection methods, while aiming to enhance performance by reducing dimensionality, may inadvertently omit important features that contribute to the complex diagnostic patterns captured by ANFIS, SVM, and decision trees.

2. **Interplay of Features**: ANFIS, SVM, and decision trees are capable of capturing intricate relationships and interactions among features. These interactions may not be fully captured or represented accurately through the feature selection process, leading to a loss of information that is critical for accurate diagnosis.

3. **Unique Characteristics of Mesothelioma**: Pleural mesothelioma may exhibit unique diagnostic characteristics that are not easily represented by a subset of selected features. Feature selection might not fully capture the nuances of the disease, whereas the algorithms' inherent capabilities may allow them to identify important but less obvious patterns.

4. **Data Adequacy**: In some cases, the dataset available for training and testing might not be large enough to accurately perform feature selection. In such situations, feature selection methods might struggle to identify the most informative features, potentially leading to the removal of relevant information that ANFIS, SVM, and decision trees can effectively utilize.

5. **Complex Decision Boundaries**: ANFIS, SVM, and decision trees can capture complex decision boundaries in the data space. Feature selection might inadvertently simplify these

boundaries, resulting in a loss of discrimination power that the algorithms inherently possess.

6. **Robustness of Models**: ANFIS, SVM, and decision trees are known for their robustness and adaptability to different types of data. They can handle noise and redundant features effectively, which might mitigate the need for feature selection.

7. **Algorithm Complexity**: Some feature selection techniques might not be suitable for complex algorithms due to their linear or univariate nature. ANFIS, SVM, and decision trees can handle higher complexity and nonlinear relationships more effectively without needing to rely on feature selection.

8. **Feature Scaling**: Feature selection methods may not always consider the scaling or normalization of features, which could affect the performance of some algorithms. ANFIS, SVM, and decision trees can be less sensitive to feature scaling issues.

### 4.8 Why machine learning instead of Deep learning?

Machine learning techniques are better suited for small datasets than deep learning techniques. Deep learning models are designed to learn from large datasets, and they have a high capacity to learn complex patterns. However, when the amount of data is limited, deep learning models tend to overfit the data, meaning they memorize the input dataset rather than generalize it. This is because deep learning models have many parameters that must be estimated, which requires a lot of data. In contrast, machine learning models are simpler and require fewer parameters, making them more suitable for small datasets. Machine learning models such as logistic regression, decision trees, or support vector machines can be used to train models on small datasets. These models are less complex than deep learning models and are less likely to overfit the data.

Additionally, machine learning models are easier to interpret and can provide insights into the relationship between the input features and the output variable.

Deep learning models are designed to learn from large datasets. The neural networks used in typical deep learning models have a very large number of nodes with many layers, and therefore many parameters that must be estimated. This requires a lot of data. A small neural network (with fewer layers and fewer free parameters) can be successfully trained with a small data set - but this would not usually be described as "deep learning". In other words, deep learning models are not appropriate for small datasets because they tend to overfit the data, meaning they memorize the input dataset rather than generalize it. This is because deep learning models have a high capacity to learn complex patterns, which can lead to overfitting when the amount of data is limited. Therefore, it is recommended to use simpler machine learning models such as logistic regression, decision trees, or support vector machines for small datasets.

## 4.9 Comparison with previous research

In this part of the thesis, there is a need to compare the methods implemented in this research and previous articles. In previous articles, never use the combination of these feature selection and classification methods specifically on this database. Table 4.6 shows the accuracy and sensitivity of the models presented in the literature which are used similar dataset.

**Table 4.6. Comparison of proposed method with previous work**

| Previous Methods | Feature selection | Data set | Accuracy | Sensitivity |
|---|---|---|---|---|
| Fuzzy SVM (Shobana,2022) | CFS | Mesothelioma's disease data set, UCI | 100% | 94% |
| SVM (Saxena,2022) | - | Mesothelioma's disease data set, UCI | 100% | 100% |
| KNN (Saxena,2022) | - | Mesothelioma's disease data set, UCI | 81.79 | 50 |
| NN (Saxena,2022) | - | Mesothelioma's disease data set, UCI | 100 | 100 |
| Random forest (Saxena,2022) | - | Mesothelioma's disease data set, UCI | 81.54 | 71.51 |
| **Our best method: NN** | **PCA** | **Mesothelioma's disease data set, UCI** | **100** | **100** |
| **Our best method: Ensemble learning** | **PCA** | **Mesothelioma's disease data set, UCI** | **100** | **100** |
| **Our best method: ANFIS** | **REE** | **Mesothelioma's disease data set, UCI** | **100** | **100** |
| **Our best method: DT** | **REE** | **Mesothelioma's disease data set, UCI** | **100** | **100** |
| **Our best method: SVM** | **REE** | **Mesothelioma's disease data set, UCI** | **100** | **100** |

Based on the results of the table, we cannot make a correct comparison between the previous methods and our own method because they did not mention all the comparison criteria in their article. In our method, a combination of feature selection and classification is used based on several methods.

- **ANFIS, SVM, and Decision Tree:**
    - Our models achieved perfect accuracy, recall, precision, F-measure, and AUC.
- **Neural Network, Ensemble Learning, and Random Forest:**
    - Our models have varying performance, with Neural Network having lower accuracy compared to SVM and Decision Tree.

- **KNN:**
    - Our KNN model has good performance, with high accuracy, recall, precision, F-measure, and AUC.
- **Comparing with Other Papers:**
    - Our results generally show competitive or better performance across various metrics compared to the methods in other papers.

## 4.10 Conclusions

The provided Tables 4.1 - 4.5 exhibit the outcomes of various classification methods integrated with distinct feature selection techniques. A thorough evaluation of these results can provide insights into the performance of each method and its compatibility with different feature selection strategies.

In Table 4.1, we observe impressive performances across multiple metrics for different methods. Notably, methods like ANFIS, SVM, and Decision tree achieve perfect scores across Accuracy, Recall, Precision, F-measure, and AUC, showcasing their adeptness in classifying instances without any feature selection.

Moving to the tables involving feature selection strategies, we notice intriguing trends. Table 4.2 reveals that Neural Network maintains a perfect performance, indicating its resilience to feature reduction. SVM and K-nearest neighbors also show elevated AUC values, suggesting their robustness under PCA.

Table 4.3 demonstrates the suitability of Neural Network, SVM, and K-nearest neighbors under the Relief feature selection technique, showcasing consistent high-performance levels across various metrics. Ensemble learning also exhibits strong results in terms of AUC and F-measure.

In Table 4.4, ANFIS, SVM, and Decision tree maintain their perfect scores in several metrics, implying their compatibility with recursive feature elimination. Additionally, K-nearest neighbors excels in terms of Recall, Precision, and F-measure, showcasing its prowess with RFE.

Table 4.5 reveals interesting dynamics. Neural Network and Ensemble learning showcase proficiency in Precision, Recall, and F-measure, suggesting their aptitude in utilizing the Correlation-based Feature Selection strategy. K-nearest neighbors and SVM maintain solid scores across multiple metrics, showcasing their adaptability with CFS.

# Chapter 5

# Conclusions and Future Work

In conclusion, the findings of this study presented a technique that is founded on machine learning and has the potential to be utilized for the automatic detection of pleural mesothelioma. The technique that has been suggested involves identifying the most significant features from the dataset that is currently available, training a variety of machine learning classifiers, and evaluating the performance of each classifier using a variety of evaluation criteria. The results of the experiments showed that certain classification methods like ANFIS, SVM, and decision trees might yield better results without feature selection compared to using feature selection. Neural Network maintains a perfect performance, indicating its resilience to feature reduction. SVM and K-nearest neighbors also show elevated AUC values, suggesting their robustness under PCA. ANFIS, SVM, and Decision tree maintain their perfect scores in several metrics, implying their compatibility with recursive feature elimination.

It may be possible to advance this research by researching more complex types of deep learning architectures for a pleural mesothelioma diagnosis. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), as well as their respective derivatives, are two examples of architectures that fall under this category. In addition, the performance of the strategy that was suggested could be improved even further by making use of a dataset that is not only more extensive but also more varied. Additionally, the approach that has been created applies to the detection of other types of cancer, such as lung and breast cancer, and has the potential to contribute to the early diagnosis of cancer and improve patient outcomes. Breast cancer and lung cancer are two examples of these types of cancer.

## 5.1 Limitations to the work

The main limitation in this research is lack of enough data. In the pursuit of advancing the diagnosis of Pleural Mesothelioma through machine learning, the research encounters limitations primarily associated with the availability and quantity of data. The scarcity and limited accessibility of comprehensive datasets specific to Pleural Mesothelioma pose a significant challenge. Machine learning models thrive on diverse and ample data to capture the complexities of disease patterns effectively. In this context, the restricted data availability hampers the model's ability to generalize accurately across diverse patient populations and varying disease stages. Moreover, the imbalance in the dataset, with potentially fewer instances of rare or advanced cases, may affect the model's capacity to discern subtle nuances crucial for accurate diagnosis. Despite meticulous efforts to gather relevant data, the scarcity remains a noteworthy limitation, underscoring the importance of collaborative efforts and data-sharing initiatives to enhance the robustness and generalizability of machine learning models in Pleural Mesothelioma diagnosis.

# References


Alam, T. M., Shaukat, K., Hameed, I. A., Khan, W. A., Sarwar, M. U., Iqbal, F., & Luo, S. (2021). A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining. Biomedical Signal Processing and Control, 68, 102726.

Baumann, F., Ambrosi, J. P., Carbone, M., & Organization, A. D. A. (2013). Asbestos: A Hidden Killer Remains a Public Health Threat in the World. *European Respiratory Journal*, *42*(3), 756–763. https://doi.org/10.1183/09031936.00165812

Bianchi, C., & Bianchi, T. (2007). Malignant Mesothelioma: Global Incidence and Relationship with Asbestos. *Industrial Health*, *45*(3), 379–387. https://doi.org/10.2486/indhealth.45.379

Brims, F. J., Meniawy, T. M., Duffus, I., de Fonseka, D., Segal, A., Creaney, J., ... & Nowak, A. K. (2016). A novel clinical prediction model for prognosis in malignant pleural mesothelioma using decision tree analysis. Journal of Thoracic Oncology, 11(4), 573-582.

Chicco, D., & Rovelli, C. (2019). Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PloS one, 14(1), e0208737.

https://archive.ics.uci.edu/dataset/351/mesothelioma+s+disease+data+set

Choudhury, A. (2021). Predicting cancer using supervised machine learning: Mesothelioma. Technology and Health Care, 29(1), 45-58.

Fonseka, D., Underwood, S., & Stadon, L. (2017). Diagnosis and Management of Malignant Pleural Mesothelioma. *Journal of Thoracic Disease*, *9*(Suppl 3), 254–266. https://doi.org/10.21037/jtd.2017.03.169

Guo, Q., Wu, W., Massart, D. L., Boucon, C., & de Jong, S. (2002). Feature selection in principal component analysis of analytical data. Chemometrics and Intelligent Laboratory Systems, 61(1-2), 123-132.

Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

Haznedar, B., Arslan, M. T., & Kalinli, A. (2021). Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data. Medical & Biological Engineering & Computing, 59, 497-509.

İlhan, H. O., & Çelik, E. (2016, October). The mesothelioma disease diagnosis with artificial intelligence methods. In 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-5). Ieee.

Karapınar Şentürk, Z. & Çekiç, N. (2020). A Machine Learning Based Early Diagnosis System for Mesothelioma Disease . Düzce Üniversitesi Bilim ve Teknoloji Dergisi , 8 (2) , 1604-1611 . DOI: 10.29130/dubited.659106

Kaur, M., & Singh, B. (2019). Diagnosis of Malignant Pleural Mesothelioma Using KNN. In Proceedings of 2nd International Conference on Communication, Computing and Networking: ICCCN 2018, NITTTR Chandigarh, India (pp. 637-641). Springer Singapore.

Khan, S. N., Sikander, G., Anwar, S., & Khan, M. T. (2018, March). Classification of malignant mesothelioma cancer using support vector machine. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-5). IEEE.

Institute, N. C. (2021). *Mesothelioma*. https://www.cancer.gov/types/mesothelioma.

Remon, J., Reguart, N., Corral, J., & Lianes, P. (2015). Malignant pleural mesothelioma: new hope in the horizon with novel therapeutic strategies. *Cancer treatment reviews*, *41*(1), 27-34.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, *10*(3), e0118432.

Shobana, M., Balasraswathi, V. R., Radhika, R., Oleiwi, A. K., Chaudhury, S., Ladkat, A. S., ... & Rahmani, A. W. (2022). Classification and detection of mesothelioma cancer using feature selection-enabled machine learning technique. *BioMed Research International*, *2022*.

Saxena, K., Zamani, A. S., Bhavani, R., Sagar, K. V., Bangare, P. M., Ashwini, S., & Rahin, S. A. (2022). Appropriate Supervised Machine Learning Techniques for Mesothelioma Detection and Cure. *BioMed Research International*, *2022*.

Sasikala, S., alias Balamurugan, S. A., & Geetha, S. (2017). A novel adaptive feature selector for supervised classification. *Information Processing Letters*, *117*, 25-34.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, 189-203.

Zeng, X., Chen, Y. W., & Tao, C. (2009, September). Feature selection using recursive feature elimination for handwritten digit recognition. In 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (pp. 1205-1208). IEEE.