Description of the Wisconsin Breast Cancer Database

Our experimental study is based on the Wisconsin Breast Cancer database from the UC Irvine Machine Learning Repository.
The Breast Cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. It contains 699 instances, 458 (65.5%) benign and 241 (34.5%) malignant cases. Each case is characterized by 9 attributes as described by Table I and two classes (benign and malignant).

WISCONSIN BREAST CANCER DATASET ATTRIBUTES: TABLE I

	ATTRIBUTE	DOMAIN
1.	Clump Thickness	1 – 10
2.	Uniformity of Cell Size	1 – 10
3.	Uniformity of Cell shape	1 – 10
4.	Marginal Adhesion	1 – 10
5.	Single Epithelial Cell Size	1 – 10
6.	Bare Nuclei	1 – 10
7.	Bland Chromatin	1 – 10
8.	Normal Nucleoli	1 – 10
9.	Mitoses	1 – 10

Attributes 1 through 8 were computed from digital images of fine needle aspirates (FNA) of breast masses. These features describe the characteristics of the cell nuclei in the image, namely:

1. Clump thickness: Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers.
2. Uniformity of cell size/shape: Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.
3. Marginal adhesion: Normal cells tend to stick together. Cancer cells tend to lose this ability. So, loss of adhesion is a sign of malignancy.
4. Single epithelial cell size: It is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell.
5. Bare nuclei: This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.
6. Bland Chromatin: Describes a uniform texture of the nucleus seen in benign cells. In cancer cells, the chromatin tends to be coarser.
7. Normal nucleoli: Nucleoli are small structures seen in the nucleus. In normal cells, the nucleolus is usually very small, if visible at all. In cancer cells the nucleoli become more prominent, and sometimes, there are more of them.

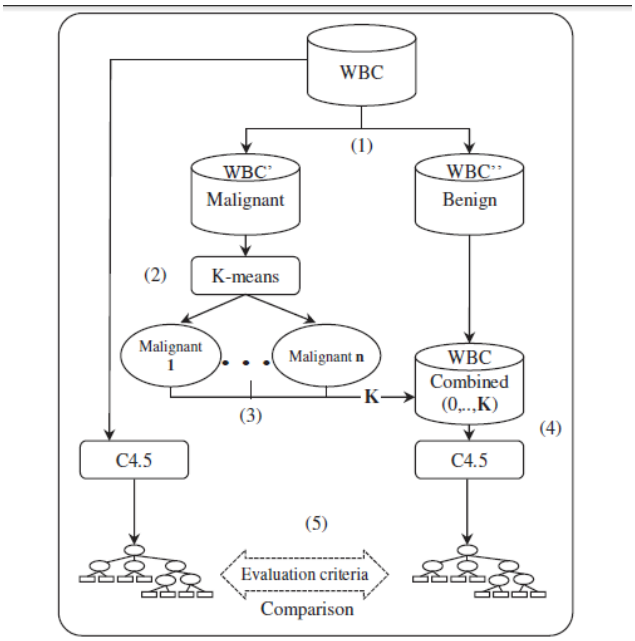
8. Mitoses: They describe cancer cells division. It gives an evaluation of the aggressiveness of the tumor. The class membership was established via subsequent biopsies or via long-term monitoring of the tumor.

A Hybrid Approach based on Decision Trees and Clustering for Breast Cancer Classification

Our report offers an experimental study on the Wisconsin Breast Cancer database using decision trees. It provides a refined treatment of malignant cases by showing that we can enhance the classification results by distinguishing different types of Breast Cancer using a clustering technique. Then to judge the quality of these clusters, we used C5.0 classifier and its accuracy which is commonly used.

The algorithm is as follows:

1. Extraction of the malignant instances from the original Wisconsin dataset.
2. Submission of the dataset obtained to the clustering method based on the K-means algorithm to split the malignant instances up into K = 2 clusters.
3. Combination of the result with the benign instances extracted from the original Wisconsin dataset, then we obtain a new dataset in which the instances are classified into (K+1 = 3) classes (benign, K types of malignant).
4. Submission of the new dataset to the C5.0 algorithm to find out the results of classification based on the confusion matrix and the global and detailed accuracy values.
5. Gathering the results to obtain a confusion matrix composed of only two classes benign and malignant (different classes of malignant combined), then computing the evaluation criteria to check the efficiency of this algorithm.



Decision Trees

Decision trees are considered as one of the most used machine learning techniques. A decision tree presents a decision procedure having the objective to find the class of an object. Three basic elements involve the representation of this technique:

1. A decision node, relative to a test attribute
2. A branch corresponding to the one of the possible attribute values.
3. A leaf including objects that, generally, belong to the same class, or at least are very similar. It is labeled by a unique class.

The attribute selection measure in C4.5 is the gain ratio based on the information theory. This criterion has the objective to determine the best attribute for each node measuring the discriminative power of each attribute A_k over classes C_i.

Different steps of the C4.5 algorithm can be summarized as follows:

1. If all objects belong to one class, then the decision tree is a leaf containing that class.
2. Otherwise,
    - Apply the gain ratio criterion and select the attribute maximizing this measure.
    - Split the training set into several training subsets, where each one corresponds to one value of the selected attribute.
    - Apply the same procedure for each subset using objects belonging only to them.

C5.0 is a more optimized version of the C4.4 algorithm.

Clustering

Clustering, or unsurprised classification partitions objects into clusters according to their similarity. The objective is to minimize the intra-cluster distance and to maximize the inter-cluster. K-means allows to partition large data sets into a predetermined number of clusters. It is based on an iterative process by randomly choosing K objects composing cluster centers. The K-means algorithm is composed of the following steps:

1. Choose a value for K, the number of clusters.
2. Randomly choose K objects as cluster centers.
3. Assign the remaining objects to their nearest cluster center based on a distance measure
4. Compute the new cluster center for each cluster.
5. Repeat steps 3-5 until no object has changed cluster

To evaluate the classification efficiency, we submit the original Wisconsin dataset to C5.0 algorithm.

Table II presents results relative to the original Wisconsin dataset.

RESULTS OF ORIGINAL WISCONSIN DATASET: TABLE II

Predicted Actual	Benign	Malignant
Benign	434	24
Malignant	6	235

Accuracy: 95.7%

Now, coming to the proposed strategy:
We extract the malignant instances from the original Wisconsin dataset and submit them to the K-means algorithm (with K = 2), to split up the malignant instances into 2 clusters.
Using the new dataset generated with the K-means algorithm (K=2), we combine it with the benign instances from the original Wisconsin dataset to obtain three values classification values (i.e. Benign, Malignant 1, Malignant 2). To evaluate the classification efficiency, we apply C4.5 algorithm, which gives results presented in Table III.

CONFUSION MATRIX OF THE NEW DATASET: TABLE III

Predicted Actual	Benign	Malignant I	Malignant II
Benign	447	8	3
Malignant I	4	95	6
Malignant II	0	4	132

By gathering the malignant instances from confusion matrix, we obtain the new confusion matrix presented in Table IV.

GATHERING MALIGNANT CLASSES: TABLE IV

Predicted Actual	Benign	Malignant
Benign	447	11
Malignant	4	237

Accuracy: 97.8%

Conclusion

Clearly, these results outperform those of the original Wisconsin dataset since we improve the evaluation criteria by using clustering before classification.