

**American University of Sharjah**

**College of Engineering**

**Department of Computer Science & Engineering**

**Spring 2021**

**CMP 466 – Machine Learning & Data Mining**

**Assignment 1**

**Due: Mar 7th, 2021**

**Farooq Mirza 80205**

**Hamad AlMaqoodi 76307**

**Omar Fayed 69702**

**Introduction**

Breast cancer is the most common type of cancer among women and causes deaths of many around the world. This disease can be detected by binary classification of breast cancer tumors as either malignant or benign tumors. It takes a lot of experience for doctors to differentiate between malignant and benign tumors as this takes lot of careful observation and practice to predict the type correctly. Therefore, the automation of this diagnosing process is necessitated to recognize tumors. Therefore, our primary aim is to create a prediction model using the Wisconsin Breast Cancer dataset to classify breast cancer tumors as either malignant or benign that it can be utilized to predict the tumor cells early and avoid any false diagnosis..

**Literature Review**

1. This paper performs classification and predictions based on 4 different machine learning classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), C4.5 and k-Nearest Neighbor (KNN) and provides the performance comparison between them based on efficiency and effectiveness, accuracy, sensitivity, specificity, and precision. These algorithms are thoroughly tested both experimentally and theoretically on the Wisconsin Breast Cancer datasets and they concluded that the SVM classifier which has the accuracy of 97.13% outperforms other algorithms in other performance metrics as well.

Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science,* *83*, 1064-1069. doi: 10.1016/j.procs.2016.04.224

1. In this paper, the authors display characteristics of two different classifiers: Naïve Bayes (NB) classifier and the k nearest neighbor (KNN) for breast cancer classification. The paper uses the Wisconsin Breast Cancer Dataset to offer an insight to the relatively new machine-learning algorithms and provides their accuracy measures using cross validation. The paper concludes that KNN gives the highest accuracy (97.51%) with lowest error rate than the NB classifier (96.19 %).

Amrane M., Oukid S., Gagaoua I. and Ensarİ T., "Breast cancer classification using machine learning," *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, Turkey, 2018, pp. 1-4, doi: 10.1109/ebbt.2018.8391453.

1. This research paper is similar to the two former listed papers also to evaluate several classifiers to obtain the best classifier which provides the best accuracy of results. However, the experiments in this paper were performed completely in MATLAB using the Wisconsin Breast Cancer Dataset after converting it from a .csv to a .mat file. The paper arrives at the result that the quadratic support vector machine (SVM) grants the largest accuracy of 98.1% with the lowest false discovery rates.

Omar I., Mohammed A., Abd Ghani K., Mostafa S., Al-Dhief. F. (2018). “Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer*.” International Journal of Engineering and Technology. 7,* 160-166. doi: 10.14419/ijet.v7i4.36.23737

1. This research looks into the relationship between breast cancer and different attributes in the aim of reducing death rates in breast cancer patients. Five different methods of data classification are used to on two data different datasets of breast cancer. The two datasets used are: Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Breast Cancer Database (WBCD). The five classification methods used are: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Logistics Regression (LR). The results from the five classification methods are compared using prediction accuracy values, F-measure metric and AUC values. The results showed that the Random Forest (RF) model is the most optimal for predicting the nature of breast cancer cells.

Y. Li, Z. Chen, “Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction”, *Applied and Computational Mathematics, vol. 7,* pp. 212-216, 2018. Available: 10.11648/j.acm.20180704.15.

1. This research proposes the development of machine learning tools to diagnose breast tumors and determine whether they are benign or malignant tumors. The study proposes developing “automated proliferative breast lesion diagnosis”. The study used two datasets, BIDMC-MGH and Wisconsin Diagnostic Breast Cancer. The researchers used Tabu search to select the most prominent features in the datasets. Then, the reduced features were categorized using five classification models and the final results of the study showed that AdaBoost and logistic regression had the best performance in diagnosing breast cancer. The effect of the Tabu search on accuracy is analyzed and compared with other works using other techniques.

H. Dhahri, I. Rahmany, A. Mahmood,E. Al Maghayreh, W. Elkilani, “Tabu Search and Machine-Learning Classification of Benign and Malignant Proliferative Breast Lesions”, *BioMed Research International*, vol. 2020, pp.1-10, 2020. Available: doi.org/10.1155/2020/4671349

1. This research studies the detection and classification of benign and malignant breast cancer cells using Extreme Learning Machine (ELM), which is a classification method. It was carried out on 9 different features in the dataset. The data is based on image segmentation from the Breast Cancer Wisconsin (Diagnostic) dataset. The results from the ELM showed a highest accuracy performance, which was 98.99%, compared to Naïve Bayes, Support Vector Machine, and Artificial Neural Network methods. As compared to these methods, ELM has a good accuracy and speed.

A. Toprak, “Extreme Learning Machine (ELM)-Based Classification of Benign and Malignant Cells in Breast Cancer”, *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research,* pp. 6537-6543, 2018. Available: 10.12659/MSM.910520

1. This research proposes a semi-supervised diagnoses method for breast cancer. This would result in a cheaper process over the normal supervised method because the supervised method requires more medical labels, which at times are expensive to obtain. The datasets used are well known benchmarks obtained from the UCI machine learning repository. The results show that the semi-supervised method has a more promising efficiency than the traditional supervised method.

M. Hosni, I. Abnane, A. Idri, J. Carrillo de Gea and J. Fernández Alemán, "Reviewing ensemble classification methods in breast cancer", *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 89-112, 2019. Available: 10.1016/j.cmpb.2019.05.019.

1. This research studies and compares 2 well known machine learning methods. These methods were used on a dataset of multi-classified histopathological images of breast cancer. The study shows that the use of pre-existing networks to extract features proved to be more fruitful than using handcrafted features as a baseline. The research also states that benign and malignant classes seem to be the most complex classes for the magnification factors.

L. Peng, W. Chen, W. Zhou, F. Li, J. Yang and J. Zhang, "An immune-inspired semi-supervised algorithm for breast cancer diagnosis", *Computer Methods and Programs in Biomedicine*, vol. 134, pp. 259-265, 2016. Available: 10.1016/j.cmpb.2016.07.020.

1. This research paper is a literature review of the classification methods for breast cancer. It studies various papers and the classification techniques they employed based on many aspects such as medical tasks tackled and the research types adopted. The study reports that the most frequently tackled medical task was diagnosis. The study also points out many flaws and optimization issues with many research papers and provides recommendations for such issues.

S. Sharma and R. Mehra, "Conventional Machine Learning and Deep Learning Approach for Multi-Classification of Breast Cancer Histopathology Images—a Comparative Insight", *Journal of Digital Imaging*, vol. 33, no. 3, pp. 632-654, 2020. Available: 10.1007/s10278-019-00307-y.

**Description of Dataset**

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

Type of data is previous records of patients from Wisconsin.

The data includes: 569 different cases of 357 benign records, 212 malignant records

Ten real-valued features are computed for each cell nucleus:

1. Radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (Perimeter2 / Area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("Coastline Approximation" - 1)

All features are of type float and they lie within their specific ranges. The diagnosis is an object that contains a character, ‘M’ for Malignant or ‘B’ for Benign. The id number is an int.

The labels are either given as Benign or Malignant.

**Features of the Data from Python Environment**

Table

Description automatically generated

The dataset contains 33 columns none of which have any null values. The first one for id which is of type int. The second column is for the diagnosis meaning whether the record was benign or malignant which is of type object. The object contains a char (‘B’ or ‘M’). The rest of the columns are various features which are all of type float. The base is 64 bits for both int and float, ensuring high precision.

Table

Description automatically generated

Table

Description automatically generated