**BENCH MARKING**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Paper** | **Dataset** | **Features** | **Missing values** | **Number of instances** | **Algorithm** | **Accuracy** | **Disadvantages** | **Advantages** |
| BREAST CANCER PREDICTION USING MACHINE LEARNING TECHNIQUES  (K. Varshini, Ram Kishore Sethuramamoorthy, Vipin Kumar, S. Abitha Shree, S. Deivarani) | Breast cancer Wisconsin diagnosis dataset  (WDBC) | 32 | No | 569 | KNN | 94.9% | Here, the research has been implemented only using KNN Algorithm. No other data mining technique was deployed which makes the research uncertain and ambiguity. | 5 different ML algorithms have been implemented in my paper. Thus, my paper looks very well researched and well implemented. |
| Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods (D. Delen, G.Walker, A. Kadam) | SEER dataset  With 10 -fold cross validation | More than 30 | No | More than 200,000 | Decision tree (C5) | 93.6% | In this paper, only three ML algorithms (Decision Tree, Artificial Neural Networks and Logistic Regression) have been researched and implemented. | I have researched and implemented five different ML algorithms. This makes my paper extensive and well researched. I have also achieved higher accuracy using Random forest algorithm. |
| Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model (K.Sivakami) | Wisconsin Breast cancer database (WBC) | 10 | Yes (16) | 699 | DT-SVM | 91% | This paper has used a hybrid model i.e. DT-SVM for higher accuracy rates. This method is solved using Net Beans and WEKA analytical tool. | The methods that I have used give more accuracy (i.e. 94.15%) and a deeper and well-oriented research. |
| Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data  (Muhammet Fatih Aslan\*, Yunus Celik, Kadir Sabanci, Akif Durdu) | Breast Cancer Coimbra Data Set  (Blood analysis data) | 10 | NO | 116 | ANN | 80% | The dataset used here is not commonly studied. So, there might be a possibility that the database show ambiguity and gaps which are yet to be discovered. This  dataset includes features which can be collected in routine blood analysis.  (MATLAB has been used in this research paper) | The dataset used in my paper is much common and reliable. Thus, the accuracy obtained from a proper dataset is precisely accurate and higher. Here, Python and Jupyter platform is used for the implementation whose data structures are superior than MATAB data structures. Also, Python is a preferable tool for ML algorithms and much easier to understand the code. |