The breast cancer is comprised of a total of 699 patient's records with cancer diagnosis information. The dataset is derived from cell images from which cell-specific

Features are extracted. Each data contains 10 features, namely:

* radius (mean of distances from center to points on the perimeter)
* texture (standard deviation of gray-scale values)
* perimeter
* area
* smoothness (local variation in radius lengths)
* compactness (perimeter^2 / area - 1.0)
* concavity (severity of concave portions of the contour)
* concave points (number of concave portions of the contour)
* symmetry
* fractal dimension ("coastline approximation" – 1)

Each data measurement also has associated diagnosis information, where the Diagnosis has only two values i.e. 2 or 4 with 2=B= benign, 4=M = malignant. Out of 699 instances, we selected 390 instances for inclusion into the Training Data Set and 309 instances into the Test Data Set. We employed a supervised classification technique with a Logistic Regression Classifier on Hadoop and a Decision Forest Classification Algorithm on MapReduce. Partial Decision Forests is a MapReduce implementation where each mapper builds a subset of the forest using only the data

available in its partition. This allows building forests using large datasets as long as each partition can be loaded in-memory.

Generate descriptor for the dataset using the command:

bin/hadoop jar $MAHOUT\_HOME/mahout-core-0.6-job.jar

org.apache.mahout.classifier.df.tools.Describe -p BC/BC\_Train.arff -f /

BC/Desc.info -d 9 N L

Generate the model using command:

bin/hadoop jar $MAHOUT\_HOME/mahout-examples-0.6-job.jar

org.apache.mahout.classifier.df.mapreduce.BuildForest -

Dmapred.max.split.size=1874231 -d BC/BC\_Train.arff -ds BC/Desc.info -t 1

–o BC-model

Classify the test data using command:

bin/hadoop jar $MAHOUT\_HOME/mahout-examples-0.6-job.jar

org.apache.mahout.classifier.df.mapreduce.TestForest -i BC/BC\_Test.arff -ds BC/Desc.info -m BC-model -a -mr -o BC-output