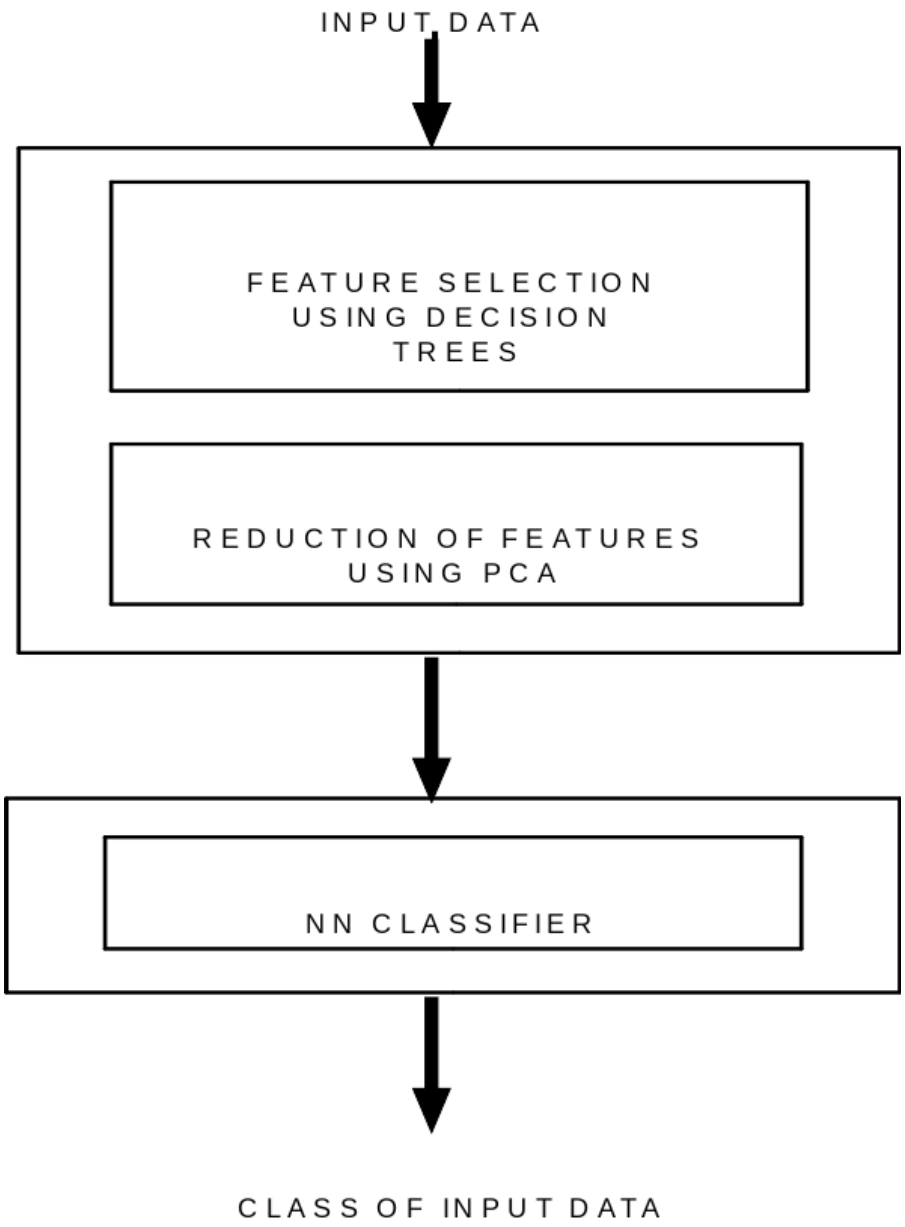
HYBRID APPROACH FOR CLASSIFICATION OF BREAST CANCER

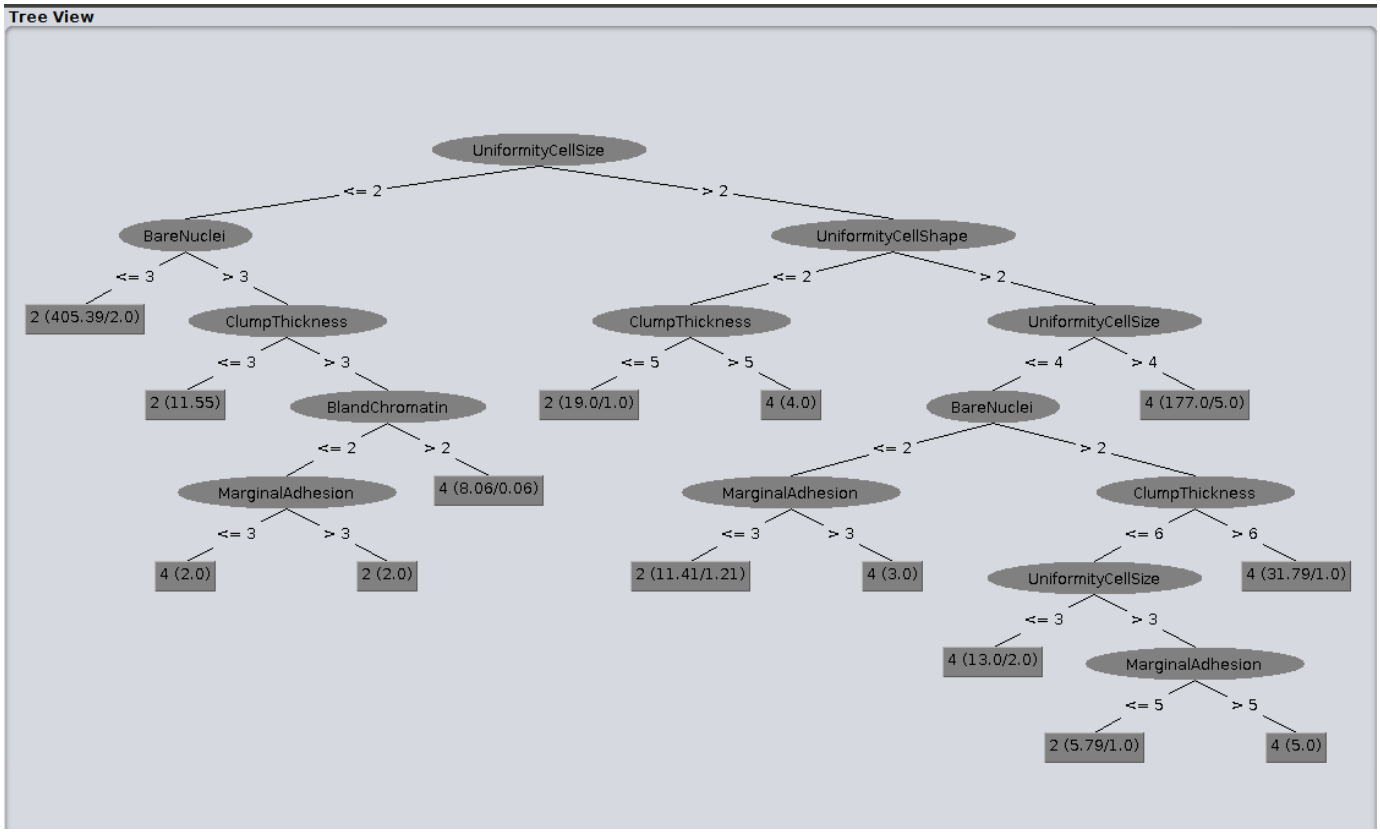
This method consists of 2 phases. In the first phase, for feature selection, namely DECISION TREES and PCA are used respectively. Therefore, the important features have been selected and feature vector size has been reduced using PCA. In the second phase, these reduced data are used for the artificial neural network and classification has been performed.
The block diagram of the process is as given below :



FEATURE SELECTION

If two or more attributes are highly correlated, they receive too much weight in the final decision as to which class an example belongs to. This leads to
a decline in accuracy of prediction in domains with correlated features. C4.5 does not suffer from this problem because if two attributes are
correlated, it will not be possible to use bothof them to split the training set, since this would lead to exactly the same split, which makes no difference to the existing tree.

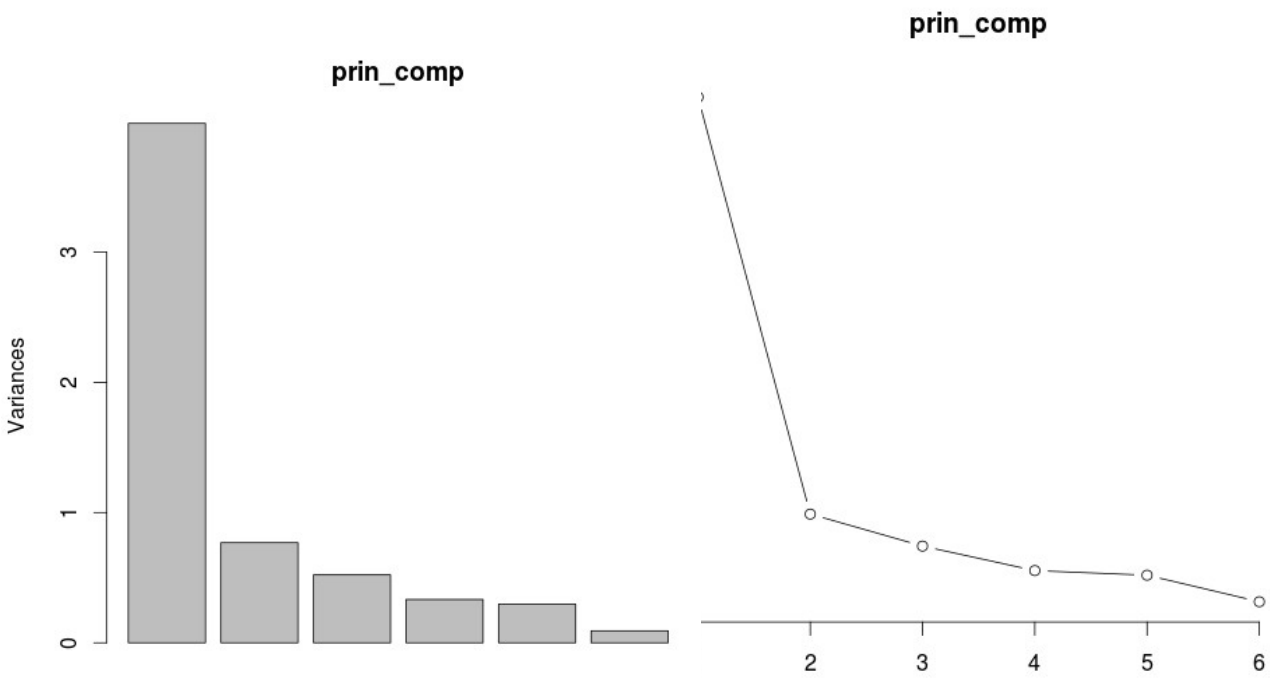
1. Shuffle the training data and take a 10% sample.
2. Run C4.5 on data from step 1.
3. Repeat 5 times (step 1-2)
4. Select a set of attributes that appear repeatedly in the simplified decision tree as relevant features.
5. Form a union of all the attributes from the 5 rounds.



Running C4.5 on 10% of the input data shows that out of total 9 features only 6 features were being repeatedly used.
These 6 features are -
Uniformity Cell Size, Bare Nuclei, Uniformity Cell Shape, Clump Thickness, Bland Chromatin, Marginal Adhesion.

These features were then reduced using Principal Component Analysis.

```
> prin_comp = princomp(newDecision.table, scores = TRUE, cor = TRUE)
> summary(prin_comp)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.9970491 0.8767980 0.72224445 0.57660917 0.54492963 0.30324365
Proportion of Variance 0.6647008 0.1281291 0.08693951 0.05541302 0.04949138 0.01532612
Cumulative Proportion  0.6647008 0.7928300 0.87976948 0.93518250 0.98467388 1.00000000
>
```

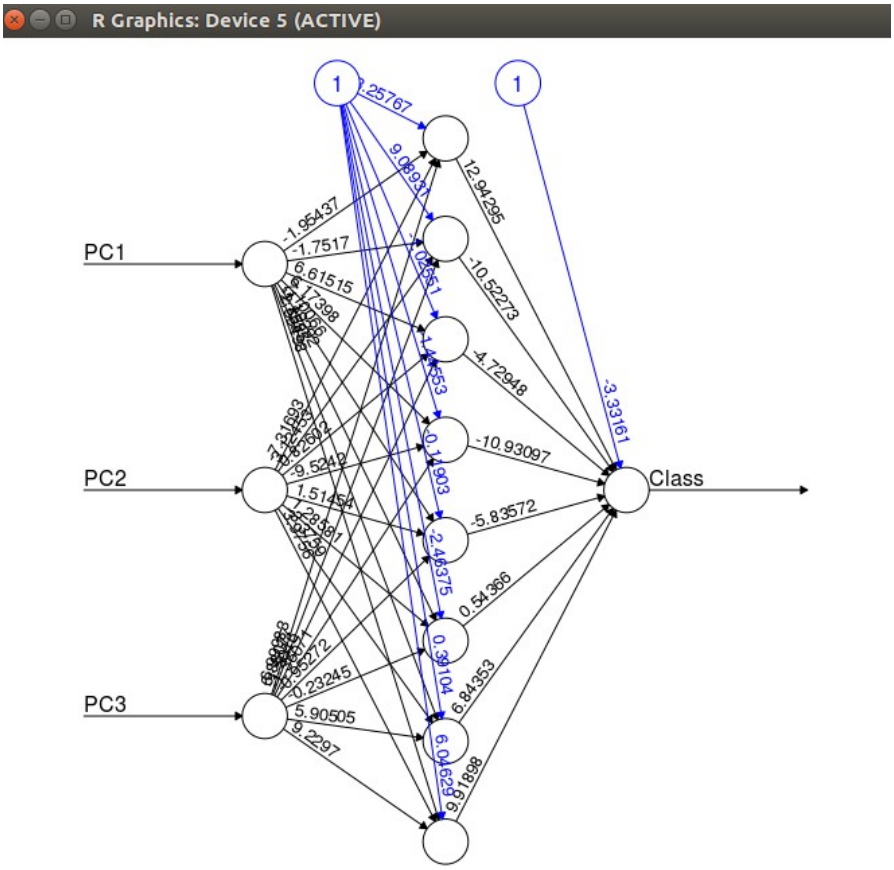


The first 3 components were then used for feeding into the neural network.

ARTIFICIAL NEURAL NETWORK

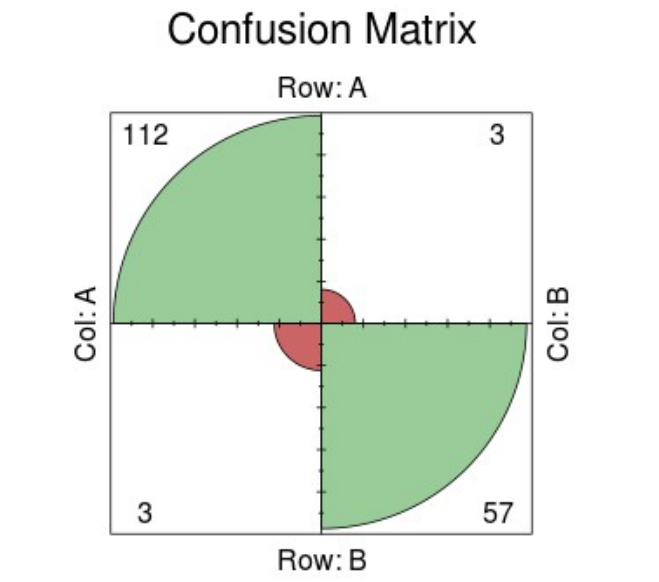
Artificial Neural Networks (NN) are biologically inspired, intelligent techniques and they have a number of simple and highly interconnected layers of neurons. Multilayered perceptron neural networks (MLPNNs) are the simplest NN architectures, and therefore most commonly used.

An MLPNN has mainly three layers: an input layer, an output layer, and an intermediate or hidden layer. The input layer neurons distribute the input signals x i to neurons in the hidden layer (s).
Training a network consists of adjusting weights of the network using a learning algorithm. The Back-Propagation learning algorithm is used in this study.



PERFORMANCE EVALUATION

```
> confusionMatrix(predict.nn1,test.pca$Class, positive = NULL,
+ dnn = c("Prediction", "Reference"))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 112   3
         1   3  57

               Accuracy : 0.9657143
                 95% CI : (0.9268712, 0.9873157)
    No Information Rate : 0.6571429
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.923913
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9739130
            Specificity : 0.9500000
         Pos Pred Value : 0.9739130
         Neg Pred Value : 0.9500000
             Prevalence : 0.6571429
         Detection Rate : 0.6400000
   Detection Prevalence : 0.6571429
      Balanced Accuracy : 0.9619565

       'Positive' Class : 0
```


Confusion Matrix
Row: A
112
3
Col: A
Col: B
3
57
Row: B