BERKELEY PUBLICATION

03.28.2022 AND RESEARCH INTERNATIONAL

Journal of Health, Metabolism & Nutrition Studies Vol. 19 No. 3

COMPARATIVE ANALYSIS OF RANDOM FOREST-NEURAL NETWORK CLASSIFICATION TECHNIQUES FOR LUNG CANCER DISEASE DIAGNOSIS

AHMED, ABIODUN TAOFIK

School of Information and Communication Technology, Department of Computer Science, Kwara State College of Arabic and Islamic Legal Studies, Ilorin. PMB 1579

ABSTRACT

Cancer is term to be the most disastrous and threatening disease universally among different diseases. The statistics shows that around more than fifty million cases were diagnosed around 10.2 million demise occurred in 2020.

The paper therefore, adopts two classification algorithms Multilayer Perceptron (MLP) and Random Forest (RF) for comparison on analysis of lung cancer microarray dataset. The experimental analysis was performed on weka Version 3.6.10 enviroment. It was observed that the multilayer perceptron (MLP) gives an accuracy of 68%, sensitivity of 69% and total time to build model of 992.64seconds while random forest (RF) gives an accuracy of 86%, sensitivity of 86% and total time to build

Introduction:

Cancer is term to be the most disastrous and threatening disease universally among different diseases. The statistics shows that around more than forty million cases were diagnosed more than 10.2 million demise occurred in 2020, by cancer which is high than that of 2016 was millions present cases are million which as kill people due to cancer. GLOBOCAN 2020 shows that cancer (lung) is the actual cause of mortality worldwide. This is the fact that cancer is not easily detected in the beginning but difficult to overcome at

model of 94seconds. The results obtained show that random forest (RF) outperforms multilayer Perceptron (MLP) classifier.

the end stage. If the identified lung nodules can be accurate at initial stage, the patient's survival rate can actually be increased by a certain percentage. Now days, automated diagnostic systems play important role of detecting any kind of disease. Diagnostic machine built especially for medical diagnosis leads to solution which will help in decreasing mortality rate and these diagnostics medical systems or machine helps in discovering cancer at early stage which is mostly important in bioinformatics.

Artificial intelligence's (A I) is used in different field like cancer diseases diagnosis. It was reveal by Cancer Society of American (2020) that cancer begins when a particular cell in segment of the body begins to increase out of proportion. Cancer has different kinds, but they all begins to grow of uncommon cells. There are different in cancer growth than that of the normal cell growth. Instead of dying, there's formation of abnormal cells. Cells can also intrude other tissues, where the normal cells cannot intrude. The invading of other tissues can cause disorder to the cell of the body which is a cancer cell. Cancer Society in American (2018). recent cases of cancer is more than 4,548 per 200,000 men and women per year statistic 2013 til 2018 cases, the statistics of cancer death is more than 1,712 per 200,000 men and women per year statistics 2003 til 2018 death. In 2019more than 15,780 of babies and adolescent between age of 0 - 19 who are diagnosed with the disease and 1,960 were lost to the disease. In 2019, more than fourteen (14) million recent cases and Eight (8) million cancer – related death worldwide. The statistics of recent cases of cancer will rise to twenty two million within the next twenty years. Above 60 percent (%) of the world recent cancer cases happened in Africa, Asia, and Central South America, 70 percent (%) of cancer deaths also happen in this region. Detecting of cancer at initial stage is the key to its cure. The disease detection is crucial, real-global medical problem; this disease is deadly diseases in the world (Ganesan, et al., 2021).

Problem Statement

Detecting of cancer at initial stage is the key to its cure. The disease detection is crucial, real-global medical problem; this disease is deadly diseases in the world (Ganesan, et al., 2018).

The competent of the physician to effectively diagnose and cure cancer is directly dependent on their cancer detection at the initial stages before it is too late. According to 'World Health Organization'WHO (2020), cancer is a group of diseases that can destroy the body. Other terms used are malignant and neoplasms. Cancer feature is fast creation of strange cells that grow above their usual proportions, and which can flow into adjoining area and flow to other segment in body system. This is called metastasis. This is the cause death majorly from cancer.

Aim and objective of the study

The aim of this research is for comparison of two classification algorithms Multilayer Perceptron" (MLP) and Random forest (RF) algorithm for lung cancer disease diagnosis respectively using microarray data.

The objectives are:

1. Employ develop a classification model of lung cancer disease using multilayer perceptron and Random forest model.
2. Implement the bid model on WEKA environment based on performance metrics
3. Compare the precision of two classification algorithms.

Scope of the Study

The scope is to analysis two classification algorithms Random Forest and Multilayer Perceptron in solving Health care problems, specifically for the detecting lung cancer patients. Both RF and NN will be implemented on WEKA. Random Forest and Multilayer perceptron' models built on a system for identifying lung cancer using microarray data. A comprehensive learning of neural network such as (learning process, transfer function, back-propagation algorithm, feed-forward networks, network layers, perceptron, selection of weights, data description and training of data) and main ingredients of RF algorithm such as bootstrap, bagging, split, fitness, selection and crossover are implemented. The time complexity, Specificity, accuracy and sensitivity will be checked to see if there are changes

Literature Review

According to Cancer Research UK (2020), bodies receive oxygen through the lungs and pass it into the bloodstream so that it can circulate to everybody cell. The muscles of our chest and a large flat muscle under the lungs (the diaphragm – pronounced di-a-fram) are used to draw air into the lungs. The diaphragm is at the base of the chest cavity, just above the stomach. The chest cavity is sealed so that when you breathe in and the muscles make it bigger, this creates a vacuum inside, which draws air in through your nose and down into the lungs.

According to American Canser Society (2021), cancer can be described as a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Cancer is caused by both external factors (tobacco, infectious organisms, chemicals, and radiation) and internal factors (inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These causal factors may act together or in sequence to initiate or promote the development of cancer. Ten or more years often pass between exposure to external factors and detectable cancer.

History of Random Forest Algorithm

Random decision forest was first proposed by Tin Kam Ho of Bell Labs in 1995. This method combined Breiman's idea on bagging and the random selection of features. This classifier using different decision tree models it is also used for classification and it accuracy, variable important information is provided with the results. Random forest is a learning method implemented by increasing many classification trees and having them “vote” for a decision according to a majority role. Breiman (2019) initiate random forests, which add an additional layer of randomness to bagging. In addition to building each tree using various bootstrap samples of the data, random forests replace how the classification or regression trees are constructed. In standard trees, each node is divided using the best split among all variables.

History of Artificial Neural Network

Neural networks have had a unique history in the realm of technology. Unlike many technologies now a days which either immediately fail or are immediately popular, neural networks for popular for a short time, took about 20 years hiatus, and have been popular ever since (Eric, 2019). The researcher led the first effort to simulate a neural network. The perceptron analysed added sum of the inputs, subtracts a threshold, and passes one of two possible numbers out as the output. Unfortunately, the perceptron is limited and was proven as such during the "disillusioned years" in Marvin Minsky and Seymour Papert's in 1969 book Perceptrons. Bernard Widrow and Marcian Hoff of Stanford in 1959, the models were built called ADALINE and MADALINE. MADALINE was the first neural network to solve real world problem.

Review of Related Works

Abraham (2020), compared several machine algorithms for learning cancer disease diagnosis. This research shows the usefulness of learning algorithms like RF, Majority, Nearest Neighbors and Best Z-Score were compared on diagnosing cancer from gene showing level of data. Two dataset were used, a breast cancer dataset which classified cancers into basal and luminal, and a colorectal cancer dataset which determine if a cancer has a mutation in the gene. Best Z-Score and Nearest Neighbors algorithms were the better among others and they used all features in the classification. Decision Tree used only 13 features for classifying a sample and gave mediocre results while the worst algorithm was Majority because the algorithm did not look at any feature. Also, the researcher expressed that all algorithms were fast to train and test except decision tree which was slow because it had to help each feature in turn, calculating the information achieve of every possible choice of outpoint. However, the researcher does not pre-process the data to remove features (reduce dimensionality of the gene dataset feature selection) that are unlikely to be useful. The feature selection helps many algorithms by removing noise and speed up the training. Also, the researcher applied these techniques on breast cancer dataset and colorectal cancer dataset.

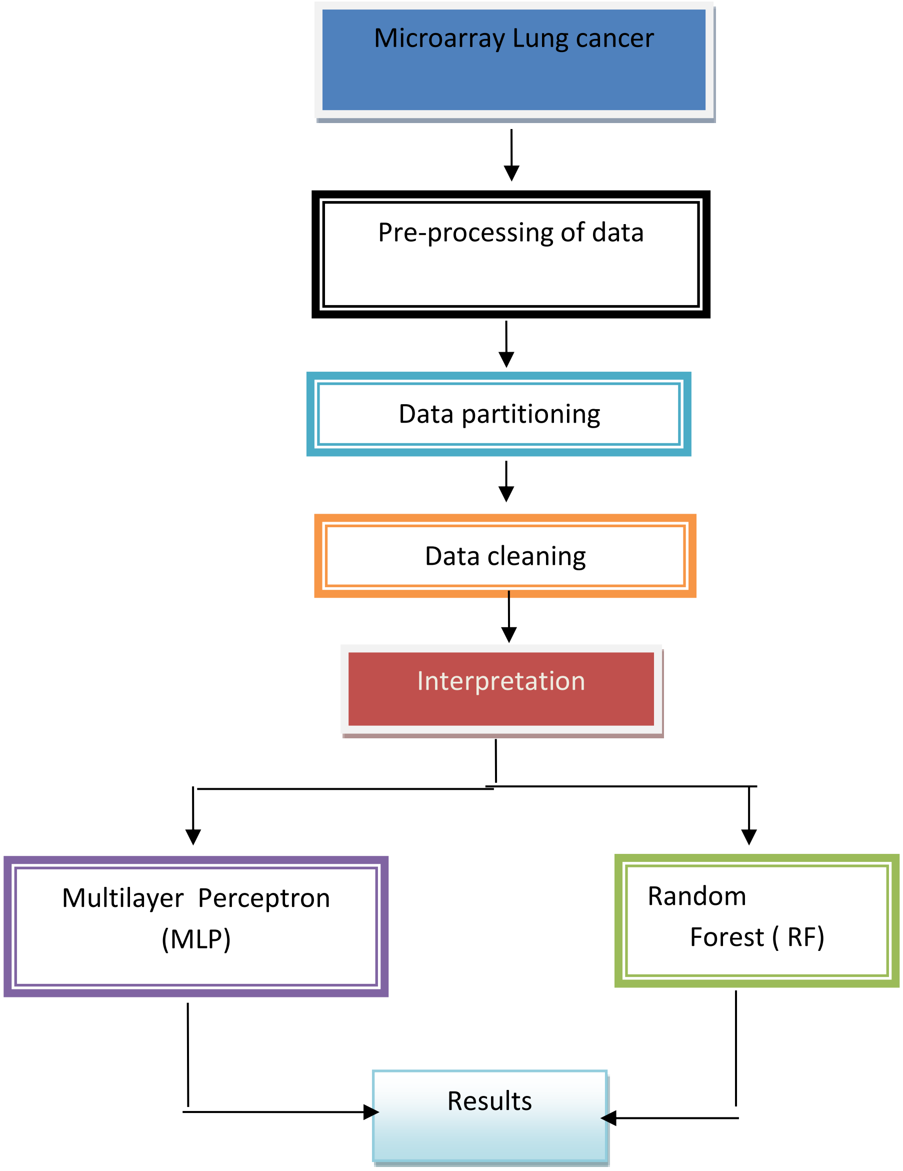
Khalid and Atif (2019) presented an evaluation of learning techniques for cancer prediction based on microarray data,various techniques were implied on prostate cancer dataset in oder to accuratly predict cancer class. The reserchers applied combination of statistical method like inter-quartile range and t-test, which has been effective in fitering significant genes and minimizing noices from data. However, each technique were handle monolthically on the prostate cancer dataset and this approch does not uselung cancer dataset.

Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood (2020) they compared classification results of two models. RF and the J48 for classifying 20 versatile datasets they took data sets containing instances varying from 148 to 20000 respectively. They compared the performance metric obtained from methods. RFs and Decision Tree (J48). The classification parameters consist of Performance metric e.g recall, TP rate, precision and accuracy. They discussed everything about the models for large and small data sets. The classification results show that RF gives better analysis for the same number of attributes and large data sets. The breast cancer analysis of data set show that when the number of instances increased from 286 to 699, the percentage% of correctly classified instances increased from 69.23% to 96.13% for RF dataset with same number of attributes but having more instances, the Random Forest accuracy increased.

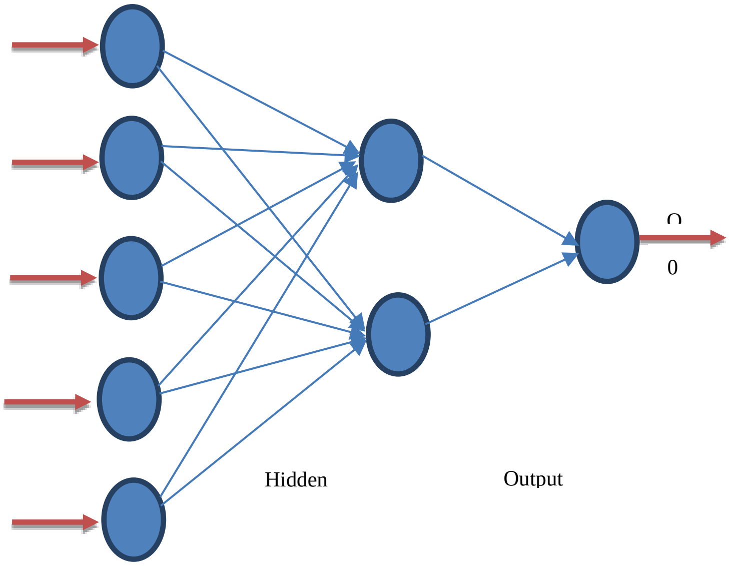
RESEARCH METHODOLOGY

The proposed system starts with the microarray lung cancer after which pre-processing of data is performed, data cleaning, data partitioning after which data interpretation is performed then MLP and RF algorithms is applied on the dataset. Afterward, the analyses of the RF and MLP are noted and the performance metrics, accuracy, specificity, precision and time of building the model is compared to evaluate the best algorithm.

Architecture of the proposed system



Neural Network for the classification of cancer disease



Artificial Neural Network

Classification is a task that is often encountered. The process classification involves assigning objects into predefined groups or classes based on a number of observed attributes related to those objects. Although there are some more traditional tools for classification, such as certain statistical procedures, neural networks have shown an effective solution to the problems. There are advantages of using neural networks, they are data driven, they are self-adaptive, they can approximate any function - structure as well as non-structure (which is quite important in this case because groups often cannot be divided by linear functions). Neural networks classify objects rather simply - they take data as input, derive rules based on those data, and make decisions.

Results and Discussion

The goal of this research is comparison of "multilayer perceptron and Random forest algorithm" for lung cancer disease diagnosis based on microarray data.

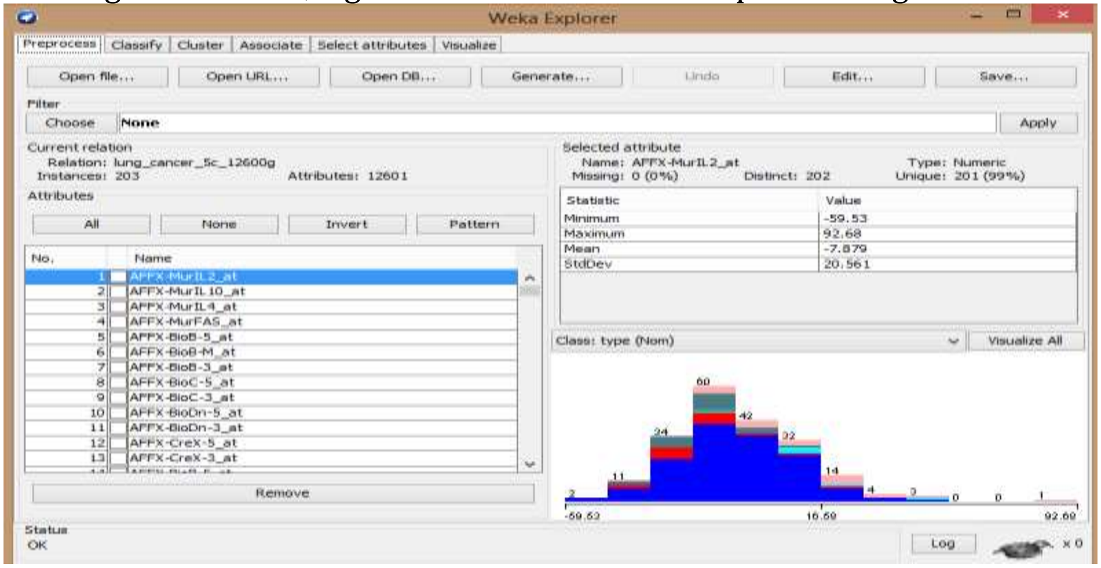
The implementation was done on Weka (Waikato Environment for Knowledge Analysis) which was developed at the University of Waikato, New Zealand. Weka is written in Java and commonly used for machine learning. Its features include pre-processing, classification, clustering, visualization, feature selection and association among others.

Proposed approach and System Specifications

The experiments were carried out on a 64-bit operating system with Windows 8.0, Intel(R) Core(Tm) i7- 3632QM CPU @ 2.20GHz and 8 GB of RAM. Expected to the iterative nature of the experiments and resultant processing power required, the Java heap size for WEKA version 3.6.12 was set to 2024MB to assess the efficiency of these algorithms.

Result of Neural Network multilayer Perceptron

The performance of a good classifer become more pronouced when subjected to a number of salient features. Figure 4.1 shows the preprocessing stage. Multilayer perceptron model was trained with 203 instances consisting of 12601 reduced attributes. Also, figure 4.2 illustrates the instances and attributes of the neural network classifer during training with 100% correctly classified instances. The figure also shows the number of incorrectly classified instances, Kappa statistic, Mean absolute error, Root mean squared error, Root relative squared error and total number of instances used for the training. The detailed accuracy by classifier model such sigmoid node o, sigmoid node 1 and 2 are depicted in figure 4.3.



Filter
None

Current relation
Relation: lung_cancer_5c_12600g
Instances: 203 Attributes: 12601

Attributes

No.		Name
1	□	AFFX-MurIL2_at
2	□	AFFX-MurIL10_at
3	□	AFFX-MurIL4_at
4	□	AFFX-MurFAS_at
5	□	AFFX-BioB-5_at
6	□	AFFX-BioB-M_at
7	□	AFFX-BioB-3_at
8	□	AFFX-BioC-5_at
9	□	AFFX-BioC-3_at
10	□	AFFX-BioDn-5_at
11	□	AFFX-BioDn-3_at
12	□	AFFX-CreX-5_at
13	□	AFFX-CreX-3_at

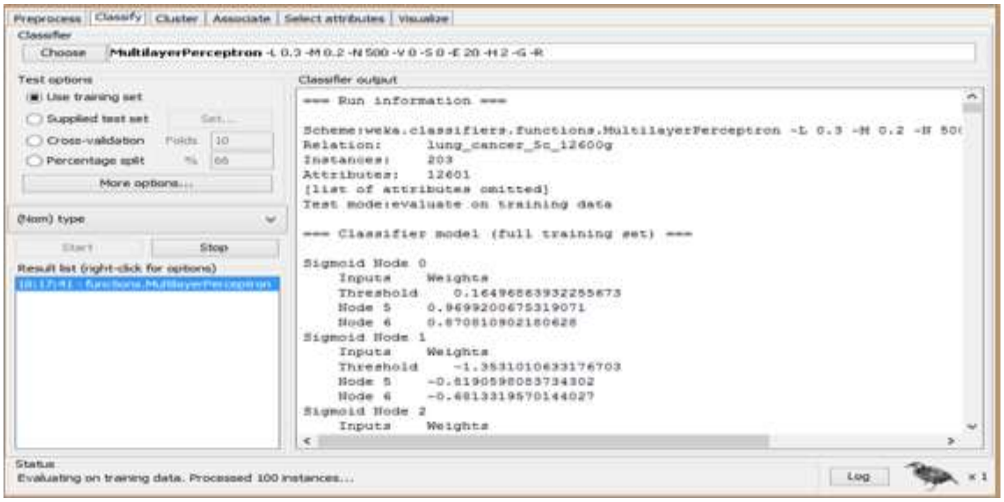
Selected attribute
Name: AFFX-MurIL2_at Type: Numeric
Missing: 0 (0%) Distinct: 202 Unique: 201 (99%)

Statistic	Value
Minimum	-59.53
Maximum	92.68
Mean	-7.879
StdDev	20.561

Class: type (Nom)

Status
OK

: Preprocess stage



Classifier: MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 2 -G -R

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

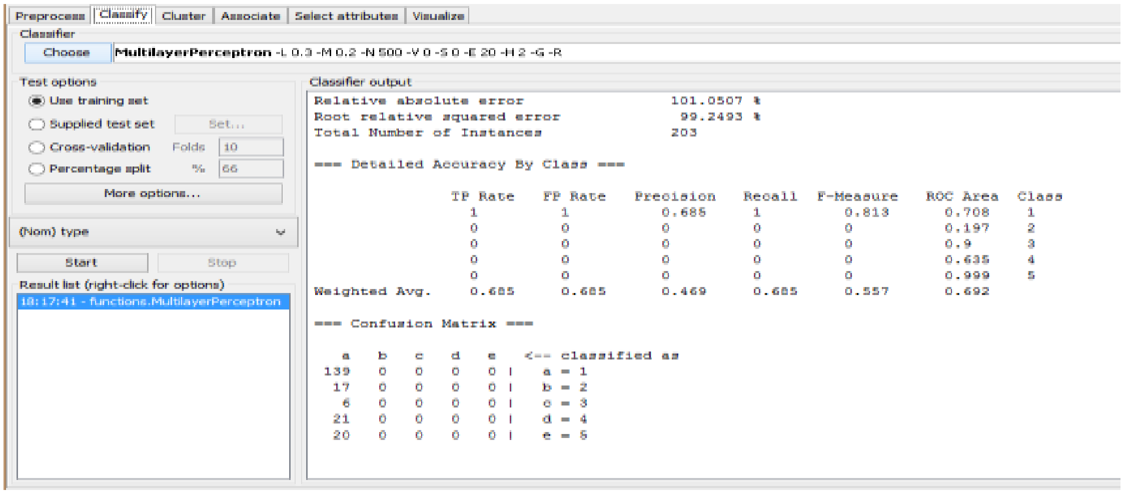
Classifier output

```
=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 50(
Relation:     lung_cancer_5c_12600g
Instances:    203
Attributes:   12601
[list of attributes omitted]
Test mode:evaluate on training data

=== Classifier model (full training set) ===

Sigmoid Node 0
    Inputs    Weights
    Threshold    0.16496863932255873
    Node 5    0.9699200675319071
    Node 6    0.870810902180628
Sigmoid Node 1
    Inputs    Weights
    Threshold    -1.3531010633176703
    Node 5    -0.8190598083734302
    Node 6    -0.6813319570144027
Sigmoid Node 2
    Inputs    Weights
```

Status
Evaluating on training data. Processed 100 instances...

MultilayerPerceptron instances and attributes

This shows the accuracy of the neural network classifier when subjected to testing dataset. The classifer acheived accuracy of 69% with false positive (FP) rate 68.5%. The confusion marix is shown in Figure 4.3.One hundred and thirty nine (139) instances were correctly classified out of the 203 instances used for testing and sixty four (64) incorrectly classified instances was observed from the experiment as depicted in figure 4.4. The model also achieved good true positive (TP) rate, FP rate,precision, recall, f-measure and receiver operating characteristic Receiver Operative Curve (ROC) area.



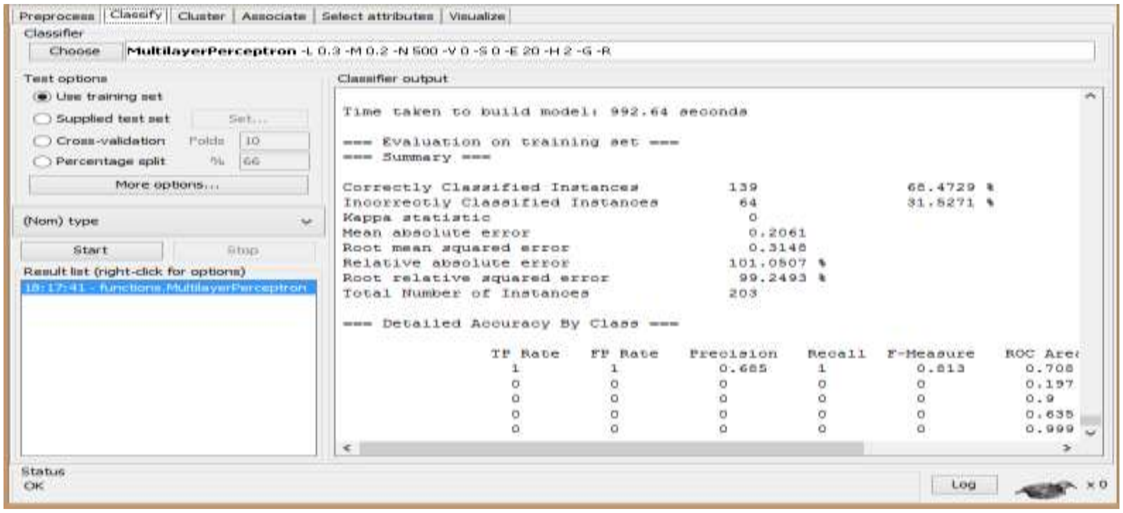
Classifier: MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 2 -G -R

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

Classifier output

```
Relative absolute error                101.0507 %
Root relative squared error             99.2493 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1         0.685     1         0.813     0.708     1
                 0         0         0         0         0         0.197     2
                 0         0         0         0         0         0.9       3
                 0         0         0         0         0         0.635     4
                 0         0         0         0         0         0.999     5
Weighted Avg.    0.685     0.685     0.469     0.685     0.557     0.692

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 139   0   0   0   0 |   a = 1
  17   0   0   0   0 |   b = 2
   6   0   0   0   0 |   c = 3
  21   0   0   0   0 |   d = 4
  20   0   0   0   0 |   e = 5
```

Performance of neural network classifier



Classifier: MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 2 -G -R

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

Result list (right-click for options): 18:17:41 - functions.MultilayerPerceptron

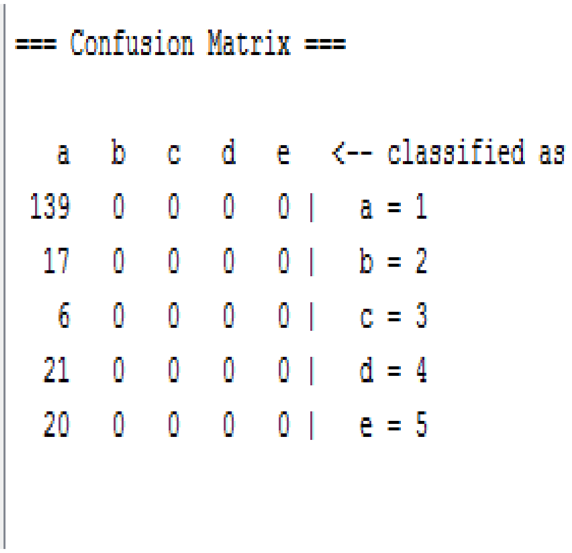
Status: OK

Classifier output

```
Time taken to build model: 992.64 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         139               68.4729 %
Incorrectly Classified Instances        64               31.5271 %
Kappa statistic                          0
Mean absolute error                      0.2061
Root mean squared error                  0.3148
Relative absolute error                101.0507 %
Root relative squared error             99.2493 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Are
                 1         1         0.685       1        0.813      0.708
                 0         0         0           0        0          0.197
                 0         0         0           0        0          0.9
                 0         0         0           0        0          0.635
                 0         0         0           0        0          0.999
```

Correctly classified and incorrectly classified instances

Indicates the confusion matrix which entails the TP rate, FP rate and precision.



```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 139   0   0   0   0 |   a = 1
  17   0   0   0   0 |   b = 2
   6   0   0   0   0 |   c = 3
  21   0   0   0   0 |   d = 4
  20   0   0   0   0 |   e = 5
```

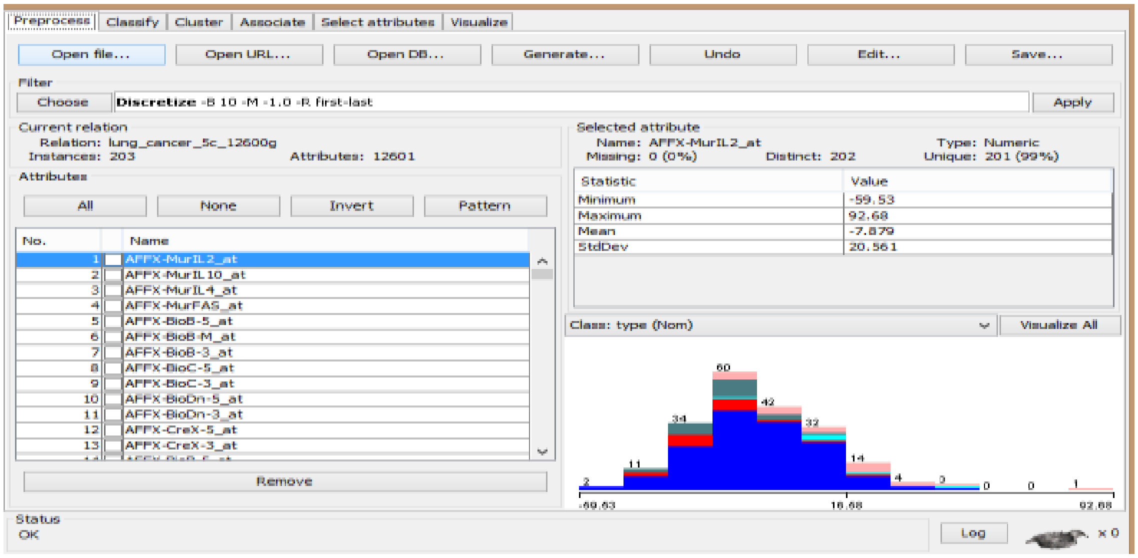
Confusion matrix of MLP

Random forest algorithm

RF is the combination of tree predictors i.e every tree relies on the number of a random vector sampled independently, selection of features randomly split each node which yields error rates that compare favourably to

Adaboost (Freund and Schapire,(1996), but are more robust with respect to noise. estimate of internal error is monitored, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Below figure shows the pre processing stage of the random forest algorithm, this stage is the initial start up of the algorithm.



Filter: Discretize -B 10 -M -1.0 -R first-last

Current relation
Relation: lung_cancer_5c_12600g
Instances: 203 Attributes: 12601

Selected attribute
Name: AFFX-MurIL2_at Type: Numeric
Missing: 0 (0%) Distinct: 202 Unique: 201 (99%)

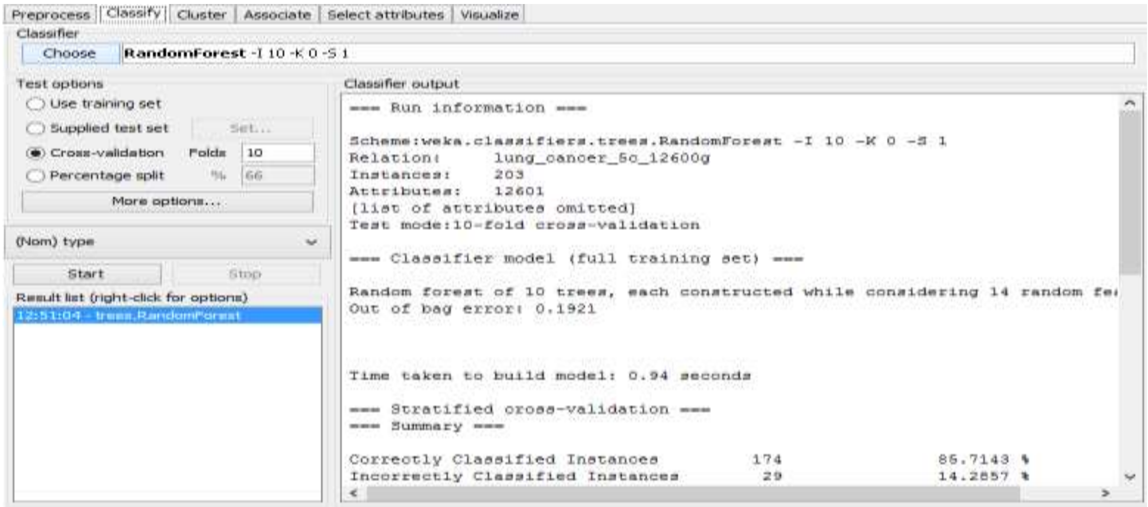
Statistic	Value
Minimum	-59.53
Maximum	92.68
Mean	-7.879
StdDev	20.561

No.	Name
1	AFFX-MurIL2_at
2	AFFX-MurIL10_at
3	AFFX-MurIL4_at
4	AFFX-MurFAS_at
5	AFFX-BioB-5_at
6	AFFX-BioB-M_at
7	AFFX-BioB-3_at
8	AFFX-BioC-5_at
9	AFFX-BioC-3_at
10	AFFX-BioDn-5_at
11	AFFX-BioDn-3_at
12	AFFX-CreX-5_at
13	AFFX-CreX-3_at

Class: type (Nom)

Status
OK

Preprocessing stage



Classifier: RandomForest -I 10 -K 0 -S 1

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) type

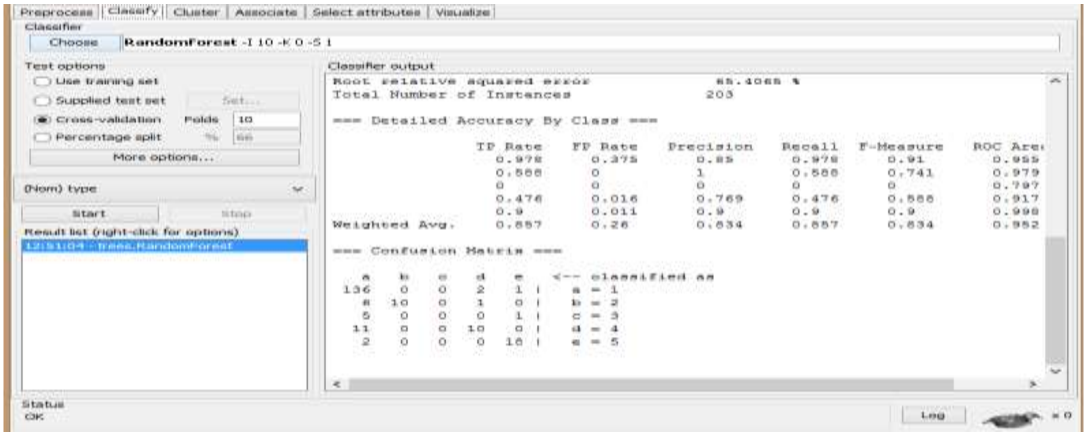
Result list (right-click for options)
12:51:04 - trees.RandomForest

Classifier output

```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1
Relation:     lung_cancer_5c_12600g
Instances:    203
Attributes:   12601
[list of attributes omitted]
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 14 random fe
Out of bag error: 0.1921



Time taken to build model: 0.94 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         174               85.7143 %
Incorrectly Classified Instances        29               14.2857 %
```

Random forest classifier model

Illustrates the time taken to build the model for the classifier. Its depicts the total time to build up the model.



Classifier: RandomForest -I 10 -K 0 -S 1

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) type

Result list (right-click for options)
12:51:04 - trees.RandomForest

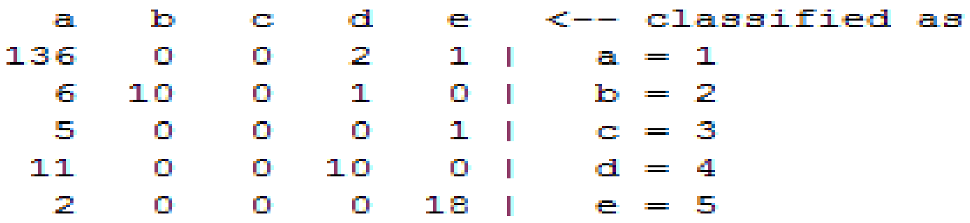
Status
OK

Classifier output

```
Root relative squared error              85.4065 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 0.978     0.375     0.85      0.978     0.91       0.955
                 0.588     0         1         0.588     0.741      0.979
                 0         0         0         0         0          0.797
                 0.476     0.016     0.769     0.476     0.588      0.917
                 0.9       0.011     0.9       0.9       0.9        0.998
Weighted Avg.    0.857     0.26      0.834     0.857     0.834      0.952

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 136   0   0   2   1 |   a = 1
   6  10   0   1   0 |   b = 2
   5   0   0   0   1 |   c = 3
  11   0   0  10   0 |   d = 4
   2   0   0   0  18 |   e = 5
```

Performance of Random forest classifier

This shows the accuracy of the Random forest classifier when subjected to testing dataset. The classifer acheived accuracy of 86% with false positive (FP) rate 0.26%. The confusion marix is shown in Figure 4.9. The model also achieved good true positive (TP) rate, FP rate,precision, recall, f-measure and receiver operating curve (ROC) area.



```
   a   b   c   d   e   <-- classified as
 136   0   0   2   1 |   a = 1
   6  10   0   1   0 |   b = 2
   5   0   0   0   1 |   c = 3
  11   0   0  10   0 |   d = 4
   2   0   0   0  18 |   e = 5
```

Confusion matrix of Random forest classifier

Performance Metrix

In the information retrieved senario, intances are documents and the task into return set of relevant documents given a search term or equivalent to :

i. Sensitivity : Also known as True Postive (TF)

$$\frac{TPR}{TP + FN} = TP$$

ii. Specificity : Also known as True Negeative rate

$$\frac{TNR = TP}{TP + FN}$$

iii. Accuracy: $\frac{TP + TN}{TP+ TN+ FP+FN} = \frac{TP + TN}{N}$

iv. Precision: Postive Prediction Rate

$$\frac{TP}{TP + FP} = p$$

v. Recall: TPR = Sensitivity

$$\frac{TP}{TP+ FN} = r$$

vi. F –Measure : $\frac{2 Pr}{P + r}$

Table 1: Evaluation of MLP and RF performance

Performance metrics		
	MLP	**RF**
Accuracy (%)	**69**	**86**
TP rate/ Sensitivity(%)	**69**	**86**
FP rate(%)	**69**	**26**
Precision	**47**	**83**
Recall	**69**	**86**
F- Measure	**56**	**83**
TTBM (in Sec.)	992.64	94

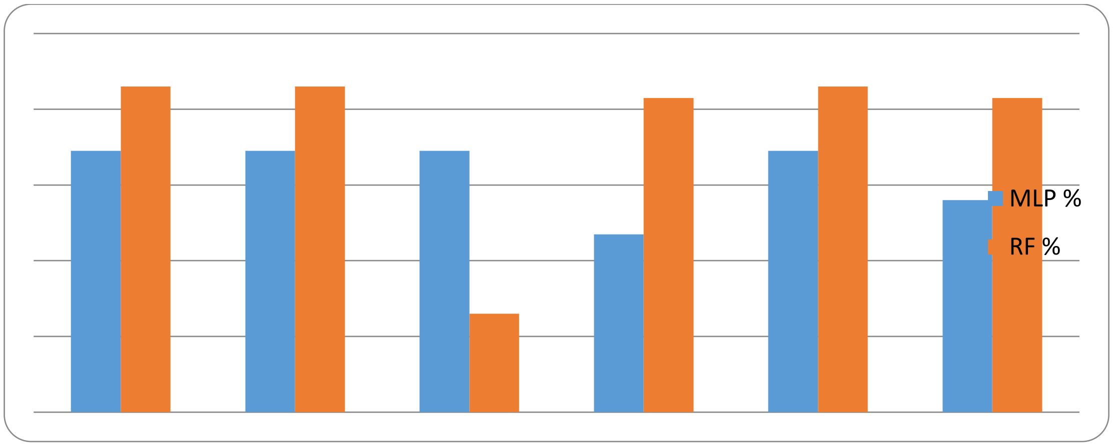


Chart of Performance Metrics of classifier algorithms

From the table 1, two Algorithms (MLP, RF) were used for lung cancer classification; it took MLP algorithm 992.64secs to build the execution model, with correctly classification accuracy of 69%, TP rate 69%, sensitivity 69%, FP rate 69% and Precision 47%. RF algorithm builds its model within 94secs, with correctly classification accuracy of 86%, TP rate 86%, sensitivity 86% ,FP rate 26 % and Precision 83%.

SUMMARY

Classification models in microarray technology play useful role in diagnosing and predicting diseases in medical research. There are extensive researches seeking the best methods of improving the classification precision of lung cancer. Different methods are statistical based methods. However, these methods result accuracy still need more attention. In this dissertation, two algorithms are proposed. These are multilayer Perceptron model (MLP) and Random forest (RF) algorithms. The two algorithms were investigated for the classification of lung cancer disease. The analysis was performed on WEKA environment. Experimental result indicates that the RF outperforms the MLP algorithms in accuracy.

CONCLUSION

This research work studies the different methods for classifying Microarray data analysis (lung cancer as a case study). Two classification methods were used, namely multilayer perceptron (MLP) and Random forest (RF) on a lung cancer dataset. This shows that the Random Forest (RF) algorithm is significantly more accurate than Multilayer Perceptron

(MLP). Data pre-processing significantly improves accuracies of the two algorithms. It was also found out that there is much evidence to support the performance difference between the MLP and the RF method. The average precision/ accuracy of RF are much higher than that of MLP algorithm. A possible explanation is that they are two different classification schemes, and hence one may be able to suits for a data set whereas the other does not.

RECOMMENDATION

This study investigates the classification of microarray dataset. These algorithms are used MLP and RF. The future work can be done to develop a system to work on the accuracy of RF classifier on a pre-processed microarray dataset. Another area that can be of interest is to compare the result between already pre-processed classified data and the classification of raw microarray dataset. Another area that can be of interest is to compare the result between already pre-processed classified data and the classification of raw microarray dataset. It helps the future researchers to know which dataset is best and suitable for the two algorithms compared and also it helps the future researchers to know the exact area to improve on as to help in solving health problem by using different algorithm

REFERENCES

America Cancer Society (2020) America cancer society

America Cancer Society (2021) America cancer society

Abraham, K. (2020). *Machine Learning Algorithms for Cancer Diagnosis.*

Breiman Leo (2019) *Random Forests--Random Features.Technical Report 567 September Statistics Department University of California Berkeley, CA 94720.*

Cancer Research UK. (2014). *The Lung.*

Eric, R. S. (2019). *Neural Networks.*

Ganesan, N., Venkatesh, K., Rama, M. A., & Malathi, P. A. (2021). *Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data. International Journal of Computer Applications, Volume 1 - No. 2*

Jehad Ali, et- el (2020): *Random Forest and Decision Tree IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694-0814* www.IJCSI.org*274*

Khalid, R., & Atif, N. H. (2019). *A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction Based on Microaray Data. International Journal of Bioinformatics Research and Applications .*

World Health Organization (2020)