6G6Z1705

Artificial Intelligence

Scenario 2

14032908
Joshua Michael Ephraim Bridge
joshua.m.bridge@stu.mmu.ac.uk

April 10, 2018

1 Introduction

In this report an AI classifier will be put forward which maps mamographical data to desired outputs (diagnoses). In order to do this two types of AI classifiers will be evaluated on their performance in this task, along with relevant pre-processing of the attributes to enhance classifier performance. The two classifier types will be a Decision Tree (J.48) and an Artificial Neural Network (Multilayer Perceptron, Minsky et al. (2017)). In order to evaluate their performance, considerations of both learning time & classification accuracy will be taken into account.

2 AI classifiers

In this section a brief study will be conducted into the two classifier types mentioned previously.

2.1 Decision Trees

A decision tree is a type of classifier (specifically a hierarchical variant of a multistage classifier, as defined by Safavian & Landgrebe (1991)) which uses a tree-like structure to test values on different attributes in a format similar to a flow chart. The tree structure itself could be described as a single root node with 0 to many connected children, each themselves with 0 to many connected children. Any node in a decision tree with no children is known as a leaf node and has a direct relationship with a class label. At each node in the tree a test is carried out on an attribute and the result of that test decides on which of the child nodes the process should continue onto. The process of completing each test from the root node to a leaf node should result in a classification of the data provided.

Self-learning decision trees are often very useful because they explicitly define how the instances are classified within the tree, simplifying the process into a set of simple rules. This is different to ANN's (see section 2.2) where the classification process is mostly hidden and can often be a very complex set of rules which would be very hard to follow.

2.1.1 C4.5 Decision tree

There are several parameters within the C4.5 algorithm that will affect the classification performance on a dataset.

Confidence. The confidence parameter is a way of controlling the amount of error-based pruning (Quinlan 1987) within the decision tree. More specifically, post-pruning is the process of estimating the error rate (probability of mis-classification) at each node in the tree, and deciding whether or not to remove the node. Lower values of the confidence factor will result in the post-pruning becoming much more aggressive with removing nodes (Beck et al. 2008).

Minimum number of objects. Within weka the Minimum number of objects parameter controls the Minimum number of instances per leaf. This means that each leaf within the tree must have at least the specified amount of classified instances for it not to be pruned. This parameter is good for data-sets which are particularly noisy which could introduce some leaf nodes which are not very stable classifiers. With a higher minimum number of objects, the tree will likely become much more pruned.

2.2 Artificial Neural Networks

An Artificial Neural Network is a mathematical system which is able to classify data by performing a series of mathematical functions (activation functions) which take weightings for each of their inputs and summise them into a single output. ANN's are designed in light of the way human/animal brains process information, via a series of neurons which are connected (in biology these connections are called synapses). Within an ANN each neuron is connected to either input attributes or the output of neruron(s) in another layer of the network. The connections between the neurons contain weightings which is the main principal behind how the network can emphasise some data over others.

The neurons within an ANN can be split up into a series of 'layers', where the outputs from one layer of neurons will become the inputs for the next layer of neurons. Within a Multilayer Perceptron (see section 2.2.1) the layers consist of 1 input layer, 1 output layer, and at least 1 'hidden layer' where each neuron in the hidden layer(s) and the output layer are neurons which perform an activation function.

Within an ANN there must be a process of 'learning' which enables it to find the most optimal values for the weights which are used in the activation functions. This is done via backpropagation which enables the algorithm to modify the weights based on the error rate of the output, compared to the expected output.

2.2.1 Multilayer Perceptron

(Minsky et al. 2017)

Hidden Layers. The hidden layers parameter allows the user to define the strcuture of the network they would like to train. Introducing more layers & neurons introduces more complexity which is good for more complex datasets with attributes which are not lineraly

seperable, however for simpler datasets this may introduce unwanted complexity within the network. What hidden layers are and how they relate to the ANN is explained in more deatail in section 2.2.

Learning Rate. Learning rate applies to the backpropagation algorithm and more specifically the Gradient Descent. It concerns the speed at which the minimum squared error is reached. A low learning rate would mean that many updates (a high training time) would be needed in order to find the global minimum - which is not desirable. If the learning rate is too high however, then this can lead to divergent behaviour where the backpropagation algorithm is not able to correctly settle on an optimal minimum.

Momentum. Once again momentum relates to the gradient descent for squared error, however momentum defines the way in which the minima is reached. As there may be several local minimas within the descent path, it would not be desirable to end in a minima which is not actually the global minima. In order to avoid this, the momentum value is linked to the learning rate in that increasing the momentum allows the descent path to continue past local minimas in search of lower squared errors.

Training Time. The main factor in the learning process of an ANN is the amount of time it has to train. There is no point in time which the network will be 'done' learning therefore the most optimal amount of learning time must be chosen. Training time is measured in 'epochs', where 1 epoch is the completion of a single training iteration. A training iteration includes the inputs passing through every layer, providing an output, and then the backpropagation algorithm updating the weights and biases for each applicable neuron. If the training time is too high, this can lead to something called 'overtraining' where the ANN becomes too dependent on the training data and will start giving worse

results when presented with unseen testing data. Therefore it is necessary to find the optimal training time, in combination with the optimal learning rate and momentum.

3 Data set analysis

In this section, the mamographical dataset will be analysed as a preparatory step to pre-processing of the data.

3.1 BI-RADS

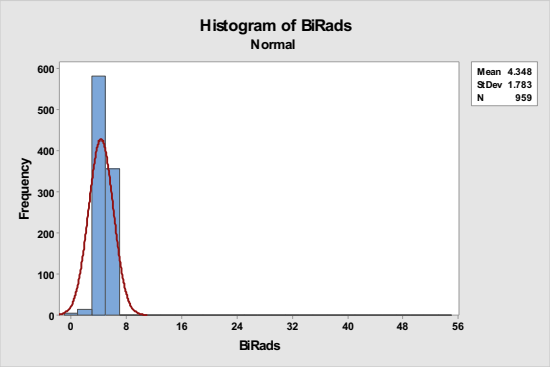
3.1.1 Measurement

The acronym BI-RADS stands for "Breast Imaging Reporting and Data System" (American College of Radiology 1998). It is a system which was designed to introduce some standardisation into the field of diagnosing breast cancer. The score for BI-RADS is on an ordinal scale from 1 to 5, with 1 being benign and 5 being very likely malignant. Below are the 5 definitions of the BI-RADS scale as defined by American College of Radiology (1998).

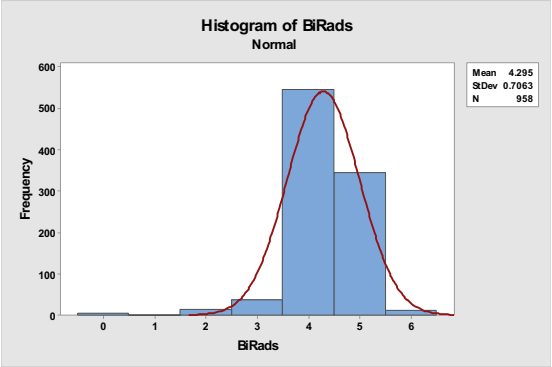
1) Negative
2) Benign findings
3) Probably benign
4) Suspicious abnormality
5) Highly suggesting of malignancy

3.1.2 Distribution

As shown in figure 1, the distribution of BI-RADS scores is non-normal and is skewed to the right.



(a) Initial



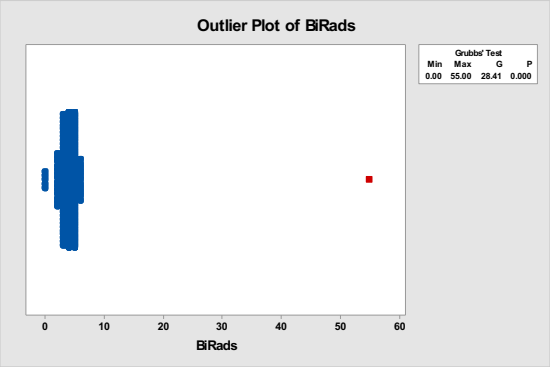
(b) Outlier replaced with median

Figure 1: BI-RADS histogram

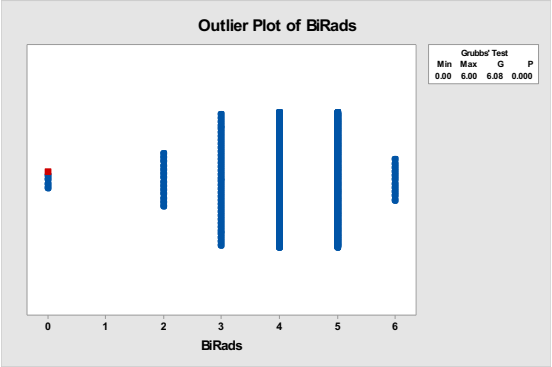
3.1.3 Outliers

Within the BI-RADS attribute there is at least one major outlier as shown in figure 2a. This outlier value is '55' and while it could be guessed that this is due to human error of entering a score of '5' twice accidentally, this outlier must be replaced by the central tendancy. This outlier plot has been repeated in figure 2b, with the outlier replaced with the median so as to get a clearer indication of other outliers.

Figure 2b shows that there are several instances with BI-RADS scores of both '0' and '6' which do not exist in the scale. These scores do exist in edition 4 of the BI-RADS scale (D'orsi et al. 2003) however the data descriptors provided with the dataset make no mention of these categories or which edition of the BI-RADS scale it refers to.



(a) Initial



(b) Outlier replaced with median

Figure 2: BI-RADS outlier plots

3.1.4 Predictive

With the BI-RADS score being a predictive scale, this would mean that if the score of BI-RADS was high then the chance of that instance having a melignant severity would be much higher. This can be shown in figure 3 where it is shown that there is a slight correlation between a higher BI-RADS score and % of severity classification.

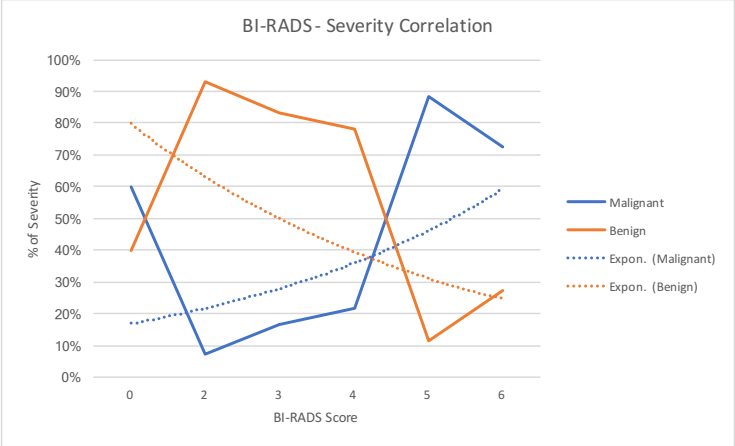
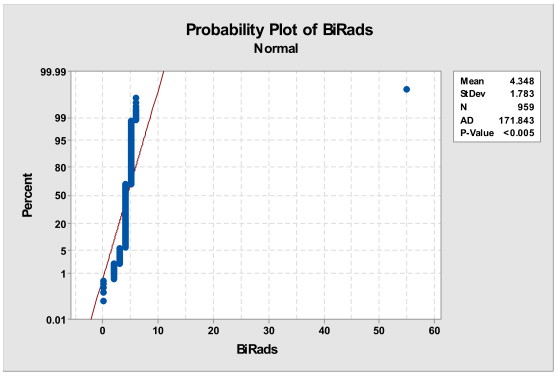
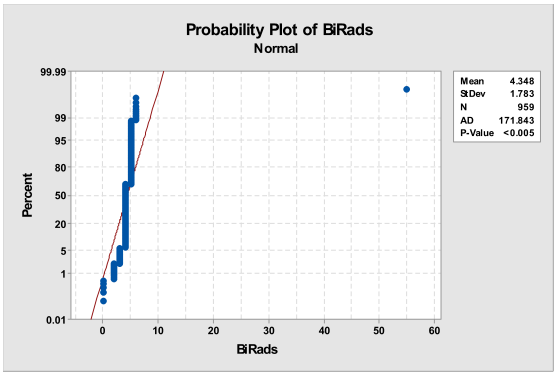


Figure 3: BI-RADS - Severity correlation



(a) Initial



(b) Outlier replaced with median

Figure 4: BI-RADS probability.

3.2 Age

3.2.1 Measurement

This is the only non-predictive attribute provided in the dataset, and is a simple ratio scale of the patients age.

3.2.2 Distribution

As shown in figure 5, the distribution of age within the dataset is non-normal and is skewed very slightly to the right.

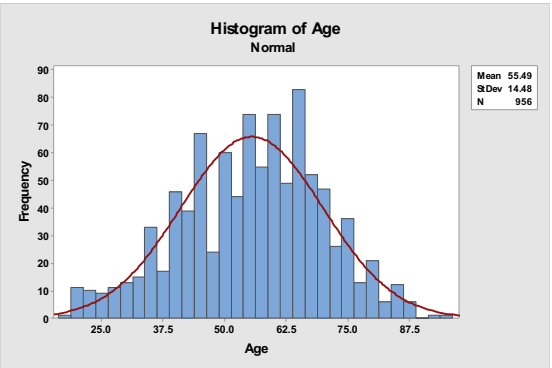


Figure 5: Age histogram

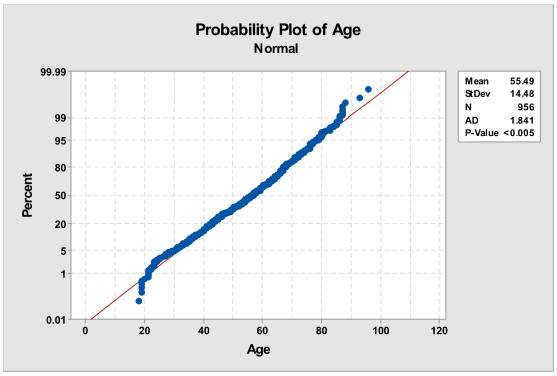


Figure 6: Age probability

3.2.3 Outliers

There does not appear to be any outliers in the age attribute as shown by figure 7. There are, however, 2 missing values within the attribute. The central tendancy for this attribute should be the median due to the fact that the data type is ratio and that the distribution is non-normal.

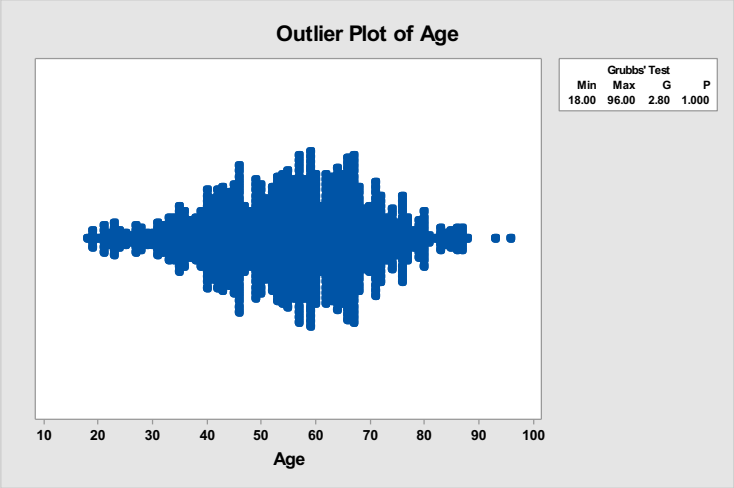


Figure 7: Age outlier plot

3.2.4 Predictive

While age itself is not a predictive attribute, it has been shown that women diagnosed with breast cancer are much more likely to be above 50 years of age (Kerlikowske et al. 1993). This can be backed up by figure 8 which shows a clear correlation between age and the % of severity classifications per age group in the provided dataset

Due to this and the findings with the BI-RADS attribute, it could be inferred that those with an age above 50 and a high BI-RADS score have a very high change of having a malignant severity.

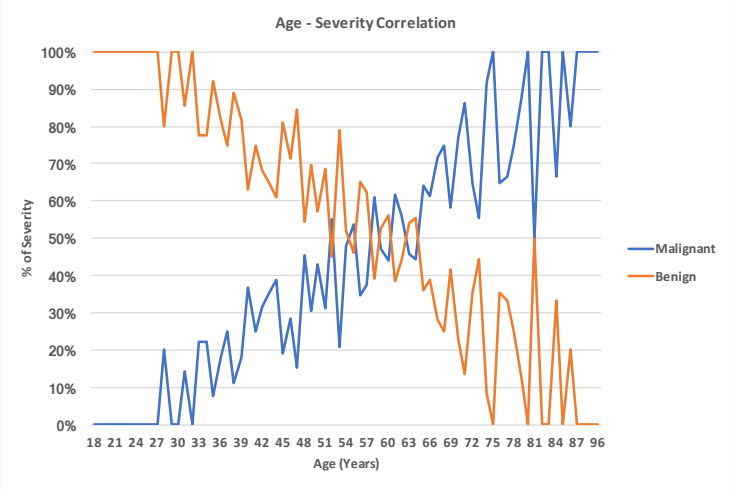


Figure 8: Age - Severity correlation

3.3 Shape

3.3.1 Measurement

This attribute is used to describe the shape of the mass being investigated. It is a nominal attribute as the number scale has no bearing on its meaning. Below are the definitions for each of the shape scores as defined by the dataset.

1) Round
2) Oval
3) Lobular
4) Irregular

3.3.2 Distribution

As shown in figure 9, the shape attribute is non-normal and is skewed to the right.

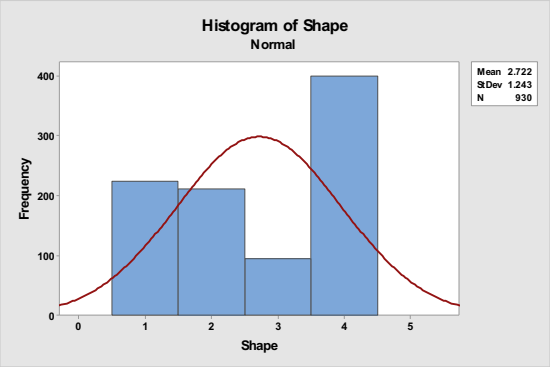


Figure 9: Shape histogram

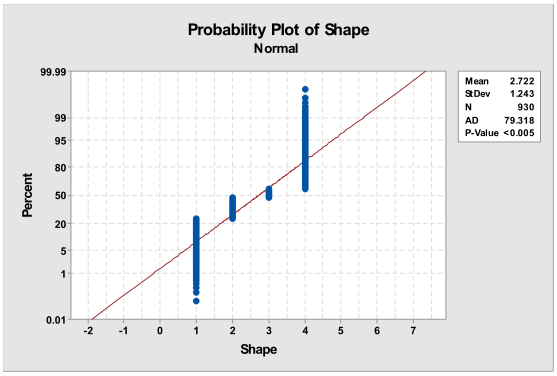


Figure 10: Shape probability

3.3.3 Outliers

While there do not appear to be any outliers within this attribute, it does contain 31 missing values. Due to it being nominal the central tendancy for these values should be the mode.

3.3.4 Predictive

In figure 11 below it can be seen that while shape is not on an ordinal scale, a higher score of shape is correlated with a higher chance of a malignant severity.

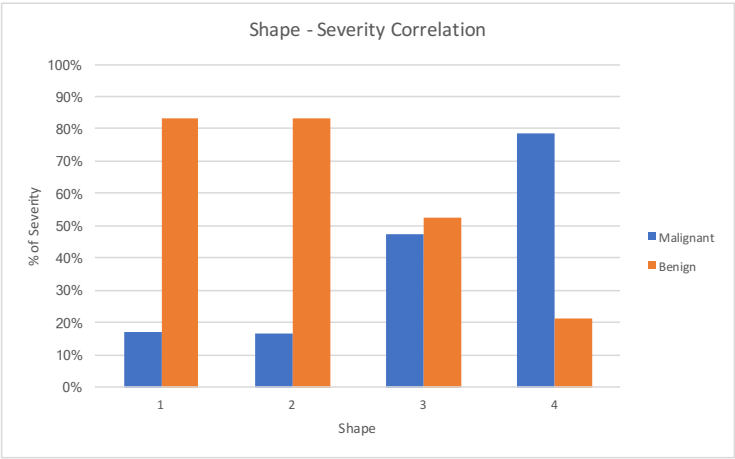


Figure 11: Shape - Severity correlation

3.4 Margin

3.4.1 Measurement

The margin attribute is used to describe the edges of the mass and their distinctiveness from the rest of the breast tissue within the scan. This is also a nominal attribute as there is no scale for the margin descriptors. Below are the 5 different definitions for each of the margin scores.

1) Circumscribed
2) Microlobulated
3) Obscured
4) Ill-defined
5) Spiculated

3.4.2 Distribution

The distribution of values in the margin attribute appears to be non-normal and is skewed to the right.

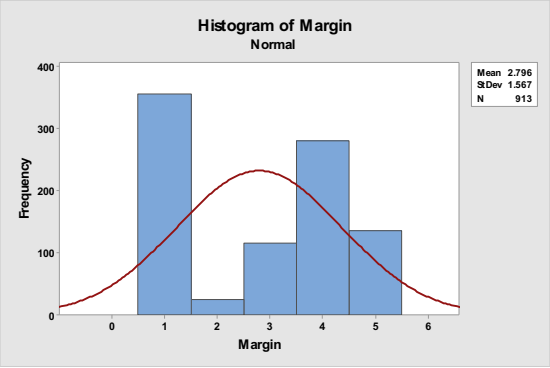


Figure 12: Margin histogram

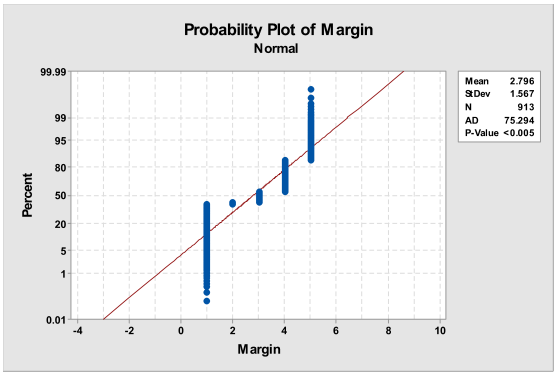


Figure 13: Margin probability

3.4.3 Outliers

There does not appear to be any outliers within this attribute however it does contain 48 missing values. Due to it again being a nominal value the central tendancy for these values should be the mode.

3.4.4 Predictive

Once again, it can be observed in figure 14 that a higher score of margin will correlate with a higher chance of a malignant severity.

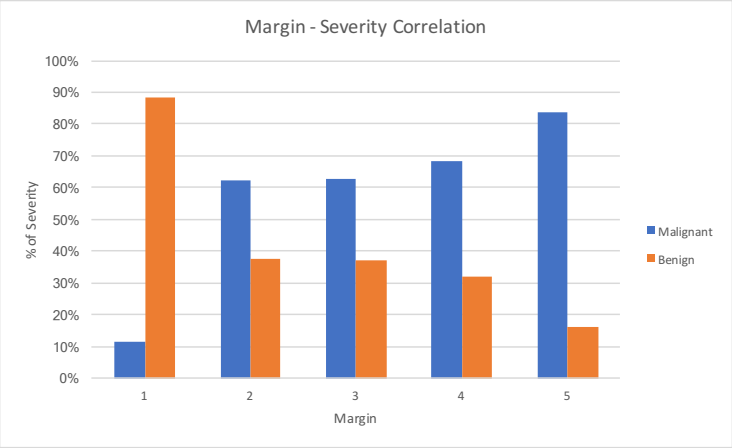


Figure 14: Margin - Severity correlation

3.5 Density

3.5.1 Measurement

This attribute is used to describe the density of the mass on an ordinal scale. A lower score indicates a higher density, which has been shown to be a significant predictor of breast cancer (Woods et al. 2011). Below are the descriptors used to score masses onto the ordinal scale.

1) High

2) ISO (Isodense)

3) Low

4) Fat-containing

3.5.2 Distribution

As shown in figure 12, the distribution of values in the density attribute appear to be non-normal and skewed to the right.

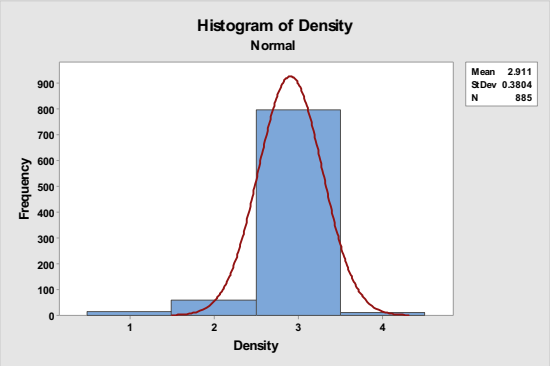


Figure 15: Density histogram

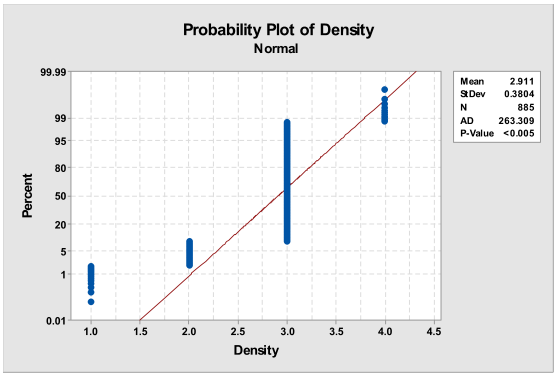


Figure 16: Density probability

3.5.3 Outliers

While there is no outliers within this attribute, there are 76 missing values. The central tendancy for replacing these values should be the median due to the data type being ordinal.

3.5.4 Predictive

Figure 17 below shows that in contrary to Woods et al. (2011), the dataset provided does not seem to have a significant correlation between high-density masses and a malignant severity.

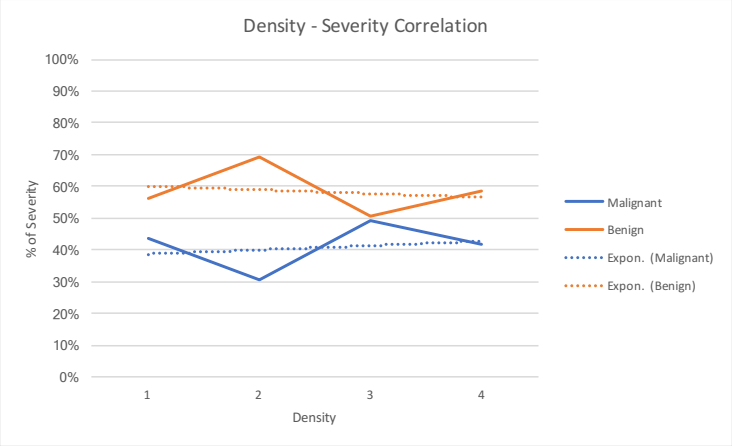


Figure 17: Density - Severity correlation

3.6 Severity

This field is known as a goal field and is a binomial type. This field contains the classification of the instances. In order to make the dataset more compatible with WEKA processing, the original values of '0' and '1' have been replaced by '−' and '+' respectively.

4 Classifier Evaluation

In this section the two different classifier types will be evaluated for their suitability in the task of classifying the mammographic dataset.

4.1 Decision Tree

4.1.1 Strengths

Speed. Decision trees are often a lot faster than neural networks (once trained), this is due to the fact that the algorithm creates a set of simple rules which are very quick to execute, and can be implemented in any programming language.

Interpretability. Again due to the nature of decision trees, the resulting classifier is very easy to interpret due to the creation of simple rules. It then provides insights on the relationships between the attributes and the classes.

Feature selection. Decision trees are good with noisy data due to the fact that they are able to completely ignore certain attributes if they are not deemed important enough as a contributor to classification.

4.1.2 Weaknesses

Nonlinear interactions. Decision trees are not as good at handling data with nonlinear relationships due to their structure. If the relationship between two attributes is not definable in a simple rule then a decision tree will not be able to

4.2 Artificial Neural Network

4.2.1 Strengths

Nonlinear interactions. ANNs are very good at handling abstract relationships in datasets, therefore if there is any information contained in the data which can be inferred, then it is likely that a neural network could make use of this inferrance.

4.2.2 Weaknesses

No feature Selection. ANNs automatically use every piece of input data given to them. If one or more of the attributes contain noisy data or do not provide much value, then these could hinder the performance of the classifier unless they are removed manually as a pre-processing step.

Black box. ANNs do not provide a systematical view of how a classification is made (it can be done given the proper functions and weight/bias values, however calculating it manually could take a very long time). This is not useful when trying to determine the relationships between different attributes in the data.

Speed. The time taken for both training and classification are often higher in neural netwoks especially when the complexity of the network structure is increased (i.e. by adding more neurons and/or hidden layers).

4.3 Prediction

In order to predict the outcome of which classifier is most suited for this task, it must be made clear how the dataset will interact with the classifier. In section 3 it was shown that the dataset has some clear links between attributes and classifications (such as a high correlation between age and severity shown in figure 8). Due to the nature of decision trees which are more able to deal with data that have a clear link between attributes and classification, it could be presumed that the decision tree classifier would be more suited for this task than an artificial neural network.

Table 1: Confidence (MO=2)

Confidence	Accuracy
0.05	82.2
0.1	82.16
0.15	82.19
0.2	82.27
0.25	82.19
0.3	**82.33**
0.35	82.31
0.4	82.12

Table 2: Minimum number of objects highest classification 1 (C=0.3)

Min Objects	Accuracy
2	82.33
5	82.3
10	82.24
15	82.54
18	82.83
19	82.99
20	83
21	83.04
22	82.98
30	83.16
35	83.53
40	83.73
45	83.8
50	**83.82**
60	83.04
70	82.82

Table 3: Learning rate accuracy from 200-3000 epochs

	Learning Rate				
	0.1	0.3	0.5	0.7	0.9
Epochs	Accuracy (%)				
200	**81.19**	80.72	80.49	80.21	79.93
250	80.19	**80.99**	80.58	80.39	79.97
350	81.27	**81.36**	81.01	80.5	80.2
450	81.49	**81.52**	80.87	80.55	80.39
550	**81.72**	81.56	81.13	81.08	80.74
650	**82.05**	81.7	81.34	81.27	80.92
750	**82.04**	81.68	81.59	81.37	81.14
850	**82.16**	81.77	81.76	81.52	81.16
950	**82.12**	81.9	81.71	81.66	81.32
1050	**82.13**	81.96	81.84	81.77	81.28
1150	81.99	**82**	81.8	81.81	81.34
1500	82.08	**82.22**	81.84	81.92	81.5
2000	82.23	**82.32**	81.88	81.94	81.74
3000	**82.25**	82.24	81.84	81.86	81.69

Table 4: Momentum accuracy from 200-3000 epochs (LR=0.4, HL=A)

	Momentum				
Epochs	0.1	0.3	0.5	0.7	0.9
200	80.59	**80.66**	80.24	79.92	79.31
300	**81.07**	80.96	80.55	80.46	79.34
400	**81.31**	81.17	80.68	80.55	79.49
500	81.33	**81.36**	80.96	80.72	79.39
600	81.45	**81.51**	81.22	80.98	79.5
700	81.55	**81.64**	81.38	81.15	79.48
800	81.52	**81.71**	81.52	81.41	79.52
900	81.69	**81.72**	81.53	81.46	79.38
1000	**81.85**	81.78	81.55	81.42	79.46
1100	81.89	**81.98**	81.61	81.63	79.66
1500	**82.07**	82.04	81.69	81.61	79.9
2000	**82.06**	82.03	81.78	81.66	80.05
3000	81.98	**82.05**	81.81	81.64	80

Table 5: Two hidden layer ANN structure (LR=0.4, M=0.2, E=950)

	Second Layer Neurons				
First Layer Neurons	1	2	3	4	5
1	82.34	82.25	82.43	82.64	**82.67**
2	81.78	81.89	82.29	82.06	**82.38**
3	81.02	81.32	81.79	81.96	**82.01**
4	80.66	81.38	80.92	80.99	**81.07**
5	80.55	80.53	**81.29**	81.02	80.7

Table 6: Learning rate impact on accuracy from 250-3000 epochs (M=0.2, HL=1)

	Learning Rate				
Epochs	0.1	0.3	0.5	0.7	0.9
250	81.73	**81.88**	81.85	81.87	81.8
350	**82.27**	82.24	82.21	82.19	82.23
450	**82.58**	82.49	82.39	82.33	82.35
550	82.55	**82.61**	82.58	82.41	82.33
650	**82.73**	82.7	82.72	82.66	82.45
750	82.77	82.8	**82.83**	82.74	82.5
850	82.79	**82.93**	82.88	82.66	82.5
950	82.89	**83**	82.87	82.78	82.55
1050	82.9	**83**	82.89	82.83	82.61
1150	82.89	**83**	82.87	82.88	82.69
1500	82.98	82.99	**83.09**	83	82.87
2000	83.08	83.15	**83.19**	83.14	83
3000	**83.25**	83.19	83.21	83.18	83.01

5 Initial Experiments

6 Main Experiments

7 Advanced Pre-processing

8 Conclusions

References

American College of Radiology (1998), *Breast imaging reporting and data system*, American College of Radiology.

Beck, J. R., Garcia, M., Zhong, M., Georgiopoulos, M. & Anagnostopoulos, G. C. (2008), A backward adjusting strategy and optimization of the c4. 5 parameters to improve c4. 5's performance.

D'orsi, C., Bassett, L., Berg, W., Feig, S., Jackson, V., Kopans, D. et al. (2003), 'Breast imaging reporting and data system: Acr bi-rads-mammography', *American College of Radiology* **4**.

Kerlikowske, K., Grady, M. & Barclay, M. (1993), 'Mammography by age', *Jama* **270**, 2444–2450.

Minsky, M., Papert, S. A. & Bottou, L. (2017), *Perceptrons: An introduction to computational geometry*, MIT press.

Quinlan, J. R. (1987), 'Simplifying decision trees', *International journal of man-machine studies* **27**(3), 221–234.

Safavian, S. R. & Landgrebe, D. (1991), 'A survey of decision tree classifier methodology', *IEEE transactions on systems, man, and cybernetics* **21**(3), 660–674.

Woods, R. W., Sisney, G. S., Salkowski, L. R., Shinki, K., Lin, Y. & Burnside, E. S. (2011), 'The mammographic density of a mass is a significant predictor of breast cancer', *Radiology* **258**(2), 417–425.