BREAST CANCER is the most common invasive cancer in female worldwide. It accounts for 16% of all females’ cancer and 22.9% of invasive cancer in women and 18.2% of all cancer deaths worldwide including both males and females. According to National Cancer Institute (NCI), it is estimated that 232,340 women will be diagnosed with breast cancer and 39,620 women will die of cancer of the breast in 2013. An accurate prediction of breast cancer survivability is very important as it often drives the treatment process and provides valuable indicators for the public health authorities. Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques can be used.

In this study we investigate the use of several data mining techniques to automatically predict breast cancer survivability based on specific features. Several measurements were used to evaluate the accuracy and the performance of the selected data mining methods. Furthermore, by implementing feature selection analysis we were able to identify the set of attributes that contributed the most in breast cancer survivability prediction. This study is based on the Surveillance, Epidemiology, and End Results (SEER) breast cancer database.

PREDICTION MODELS: In this study, the selected data mining techniques for predicting breast cancer survivability are:

- Naive Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes’ theorem with strong independence assumptions.
- Decision tree (C4.5) algorithm is a predictive model that maps observations about an item to conclusions about that item’s target value.
- Artificial neural network (ANN) is an analytical system inspired by the structure of biological neural networks and their way of encoding and solving problems. For the ANN implementations, two types of ANN were used: Multilayer perceptron and Radial Basis Function (RBF) Network.

RESULTS

In this study, the goal is to have high accuracy, as well as high sensitivity and specificity metrics. When running the selected classifiers, the final version of the dataset, with missing values and detected outliers removed, was used. In comparison with the other selected models, as shown in Figure 3, the decision tree (C4.5) algorithm outperformed the other classifiers with a classification accuracy of 0.8930, sensitivity of 0.891, and specificity of 0.985. The experimental results of the used performance measurements for each prediction algorithm are listed in the following table:

<table>
<thead>
<tr>
<th>Prediction Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>87.34%</td>
<td>0.874</td>
<td>0.927</td>
</tr>
<tr>
<td>C4.5</td>
<td>89.30%</td>
<td>0.891</td>
<td>0.985</td>
</tr>
<tr>
<td>ANN: Multilayer Perceptron Network</td>
<td>87.420%</td>
<td>0.874</td>
<td>0.973</td>
</tr>
<tr>
<td>RBF Network</td>
<td>88.8734%</td>
<td>0.889</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Based on the conducted feature selection analysis, an evaluation of the accuracy of the selected model was performed using the top nine attributes (Histologic type, Tumor size, Number of nodes, Grade, Site specific surgery code, Extension, Number of positive nodes, and Lymph node involvement). As shown in the table below, the achieved accuracy by the prediction models when using the top nine features is very similar to that of the accuracy when using the full set of features. Based on these results, we can infer that this set of top ranked features contributed more in prediction accuracy.

PREDICTION TECHNIQUE | Full attributes | Top 9 attributes |
----------------------|----------------|-----------------|
Naive Bayes           | 87.34%         | 87.4826%        |
C4.5                  | 89.30%         | 89.0896%        |
RBF Network           | 88.8734%       | 88.6313%        |

CONCLUSION

In this work, the use of several data mining techniques for predicting breast cancer survivability was studied. The obtained results are promising for applying the data mining methods into the survivability prediction problem. These results suggest that among the prediction algorithms tested, the C4.5 decision tree achieved the highest results confirming published results of related work. Moreover, based on feature selection analysis and the evaluation of the accuracy of the different prediction models, this study shows that the set of the top ranked features contributes more in the prediction accuracy.

FUTURE WORK

Due to the time constrains this work has limitations and could be improved in several ways. First, we need all records in the SEER 1973–2010 dataset to be in the same electronic format. Second, acquire more powerful resources to run different classification algorithms. Third, use different strategies when dealing with missing values. Finally, the integration of statistical functions should be examined to provide more reliable results to predict survivability.

REFERENCES

[3] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov), Research Data Center (RDC) (contracted to the University of Minnesota Cancer Epidemiology and Genetics Research Center (CEGRC), Minneapolis, MN), 2005.