Lymphedema Detection Based on Self-Reported Symptoms for Breast Cancer Survivors

Chenge Li  
Polytechnic School of Engineering  
New York University  
New York, NY  
Email: cl2840@nyu.edu

Yao Wang  
Polytechnic School of Engineering  
New York University  
New York, NY  
Email: yw523@nyu.edu

Mei R Fu  
College of Nursing  
New York University  
New York, NY  
Email: mf67@nyu.edu

Abstract—With the increased breast cancer survival, more women are facing life-time risk of developing lymphedema. Using web-based technology to develop low cost and pragmatic patient-centered e-platform will enable lymphedema detection using real-time symptom report and greatly improve patient-centered care.

We have conducted a machine learning study using a dataset consisting of 25 symptom features and lymphedema status of 355 patients, aiming to classify patients into with and without lymphedema using the symptom features only. We implemented and compared 5 well-known supervised learning algorithms including Decision Tree variants (C4.5, C5.0), Gradient Boosting Model, Artificial Neural Network and Support Vector Machine. Furthermore, we have conducted feature selection on our dataset based on various methods. Several symptom importance rankings are provided for better understanding of association between the symptoms and the lymphedema status.

I. INTRODUCTION

Breast cancer related lymphedema (hereafter, lymphedema) is a progressive and chronic syndrome of abnormal swelling and multiple symptoms. It can occur immediately or even 20 years after cancer treatment [2]- [5]. Even with current advances in cancer treatment, lymphedema remains a major health problem that has affected 20-40% of 2.9 million breast cancer survivors in the US [4]- [7]. With the increased rate and length of survival from breast cancer, more survivors are facing life-time risk of developing lymphedema, exerting tremendous impact on breast cancer survivors’ quality of life and long-term survival [8]- [10].

It remains a great challenge to achieve timely detection of lymphedema. Current detection of lymphedema largely relies on clinicians’ observation of swelling [11]- [12]. In research, assessment often focuses on a single symptom of arm swelling since it is the only symptom that can be verified by objective measures [11]. Thus, lymphedema is usually defined as a 2-cm increase in limb girth or a >200 mL or > 10% limb volume increase comparing affected (i.e. the same body side in which breast surgery or radiation performed) and unaffected limbs using water displacement, sequential circumference limb tape measurement, and infra-red perometer [11]- [12]. Water displacement and tape measurements are time consuming with limited reliability and no published sensitivity and specificity [11]. Infra-red perometer is reliable, but very costly. There are fewer than 10 infra-red perometers in the US [11]. More importantly, when swelling can be observed by clinicians or measured in terms of limb girth or limb volume, the condition usually has been for a while, which leads to poor clinical outcomes and progression to chronic condition [11]- [13]. Bioelectrical impedance analysis (BIA) can directly measure lymph fluid changes [12], it has shown limited sensitivity in identifying true cases of lymphedema and is costly. Specifically, the current cutoff point of L-Dex (lymphedema index) ratio of > +10 misses 34% of true lymphedema cases (AUC=0.81 sensitivity=0.66 [95% CI: 0.51 - 0.79]) [12].

Timely detection paired with timely intervention can reduce the chance of lymphedema progressing to chronic or severe stage [14]. It has been known to clinicians that lymphedema is a syndrome of multiple symptoms in addition to arm swelling [13] [15] - [18]. Lymphedema symptoms may indicate a critical stage of lymphedema in which changes in limb volume or limb girth cannot be detected by objective measures [15], [17] - [18]. Research has shown that with increased number of symptoms reported, limb volume elevated [18]. It is possible that real-time symptom report, that is, symptoms that patients have experienced at the time of reporting, may help to achieve timely detection of lymphedema [15] , [18] and furthermore to estimate the risk for lymphedema in the future. In this work, we examine the potential of using machine learning tools for discovering the hidden association between lymphedema and relevant symptoms using data collected in a prior study [19].

II. DATASET DESCRIPTION

Prior to the study, we built a user-friendly e-platform and successfully collected the lymphedema symptom reports from patients. In this dataset, we have 355 patient records, including 208 positive observations (with Lymphedema) and 147 negative observations. Each record contains the patient’s demographic information and self-reported symptoms as shown in Table I. Patients are asked to rate the 25 categorical symptoms in scales from 0 to 4, where 0 = No symptom; 1 = A little; 2 = Somewhat; 3 = Quite a bit; 4 = Severe. These 25 categorical symptoms are chosen as the input training features.
A. Model performance assessment

1) 5 fold Cross Validation: We randomly split the entire samples into several data subsets, with the effort to keep the lymphedema and non-lymphedema sample ratio in each subset to reflect the entire sample ratio. Each subset is termed as a “fold”. We conducted the 5-fold cross-validation. Given a classifier, for each candidate parameter, we used 4 folds of data for training and then evaluated its performance on the remaining 5th fold of data. We repeated this process 5 times, each time using a different testing fold for evaluation. The average of the classification accuracies on the 5 testing folds of data was the cross validation accuracy, used as the performance evaluation metric. Note that with such 5-fold cross validation, each training set (leaving out one fold of data subset) contains approximately 284 samples, satisfying the recommended 5:1 to 10:1 sample-per-symptom feature ratio for training classifiers [30].

2) Estimation Stability with Cross Validation: One challenging problem in training a classifier with limited data is how to avoid overfitting. The optimal model parameters chosen by minimizing the cross validation error may still suffer from overfitting. Lim and Yu in [31] have introduced the “Estimation Stability” (ES) in the Cross Validation, measuring the closeness of the prediction results obtained by models trained using different training data. When the difference is measured by the $L_2$ error, this Estimation Stability is related with variance of the model predictions. This improved criterion combined “Estimation Stability” with Cross Validation(CV), referred as “ESCV”. When looking for the proper parameters for the model, we first find the local minima of cross validation error, and then we will further limit the parameter choice to a local minimum of ES score so that the corresponding model complexity is closest to and yet smaller than the model complexity corresponding to the global minimum of CV error. Hence the parameter choice is guided by both ES score and CV error, which guarantees a small bias, and it is also gifted with higher stability.

We use Cross Validation and ESCV when tuning for the best parameters for all five algorithms. For example as shown in Figure 1, when looking for the optimal tree depth (tree layer count -1) for the decision tree, we plot both the cross validation error rate and the ES score. But instead of choosing the global minimum for the cross validation error rate at depth of 8, we have chosen a local minimum at depth of 4, since this enabled us to have a local minimum of ES score, too. A less complex model is less likely to be overfitting.

B. Decision Tree Based Algorithms

1) C 4.5: C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [21]. It is an improved extension of Quinlan’s earlier ID3 algorithm. In the construction of the decision tree, at each internal tree branching node, the tree chooses one feature $Fea$ with a threshold $T$, so that the sample set $S$ can be split into two subsets $S1$ and $S2$ based on $Fea$ and $T$. The subset $S1$ will contain samples satisfying $Fea \leq T$, and subset $S2$ containing samples satisfying $Fea > T$. The
criterion for deciding which feature and the corresponding threshold to use is referred as the splitting criterion. The splitting criterion for C4.5 is Normalized Information Gain. It chooses the feature Fea and threshold T among all possible candidate pairs to maximize the Normalized Information Gain.

From the 5-fold Cross Validation, we have trained 5 trees. In order to get better understanding of the relative importance of the symptoms, one of these trees are shown in Figure 2.

2) C 5.0: C5.0 is an improved version of C4.5, and it is considered as the most advanced decision tree classifier. We have used the software C5.0 developed by RuleQuest Research [23]. It constructs a decision tree in two steps. It first grows a large tree to fit the data closely and then prunes the initial tree by removing parts that are predicted to have a relatively high error rate. The pruning is controlled by the parameter Pruning Confidence Factor (CF), which affects the way that error rates are estimated and hence controls the severity of pruning. Values smaller than the default (25%) cause more nodes in the initial tree to be pruned, while larger values result in less pruning. The other parameter is the Minimum case number, which is the minimal number of samples remaining in a tree branch to be considered for further branching of the tree. Minimum case number higher than the default (2) can forbid the initial tree from growing fully. Through cross validation, we found using 35% for CF and 8 for the minimum case number achieved the best average cross validation accuracy.

3) Gradient Boosting Model: The idea of boosting is to combine the power of weak classifiers to generate a strong one. Weak classifiers are those with accuracy only slightly higher than random guess. The procedure of boosting is to sequentially train weak classification models in a stage-wise additive way with modified versions of the data. Data points that are hard to train, i.e. those giving large loss function values in current iteration, will be emphasized more in the next iteration. Hence in the end of M iterations, boosting model will have a sequence of M weak classifiers. Higher weights will be given to those classifiers with higher accuracies, and the ultimate classification decision is sign of the sum of the weighted votes from them all.

C4.5 and C5.0 both produce a single tree model after training. The Gradient Boosting Model (GBM) [24] produces a family of weak trees and uses them together to produce a strong classifier. It is a generalized method in the sense that the loss function can be any differentiable loss function. Here the loss function we used for training GBM is the Binomial Deviance, as it has been shown to be more robust for classification problems. Let y denote the training label (ground truth) for a data sample and f the prediction result, the Binomial Deviance is defined as:

\[ L = \log(1 + \exp^{-2yf}) \]  

The GBM model was implemented using the GBM package in R [25]. There are 4 main tuning parameters: max iteration number M, variable interaction depth (tree depth) J, shrinkage factor (learning rate) V and subsampling ratio R. The shrinkage factor, also referred as the learning rate, controls the contribution of each tree when it is added to the current model. If the subsampling ratio R is less than 100%, the program only uses a fraction of the training data (without replacement) in each iteration.

C. Artificial Neural Network

We have implemented the single hidden layer back-propagation network using the MATLAB Neural Network Toolbox [26] [27]. The basic structure for ANN is shown in the Figure 3. In our case, the input layer has 25 symptoms and
TABLE II: Average Accuracy, Sensitivity, Specificity over 5 Cross Validation Folds

<table>
<thead>
<tr>
<th>Model</th>
<th>Decision Tree C 4.5</th>
<th>Decision Tree C 5.0</th>
<th>Gradient Boosting Model</th>
<th>Artificial Neural Network</th>
<th>Support Vector Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (SD)</td>
<td>76.31% (0.0393)</td>
<td>77.11% (0.0386)</td>
<td>80.50% (0.0497)</td>
<td>93.75% (0.0272)</td>
<td>81.65% (0.0404)</td>
</tr>
<tr>
<td>Sensitivity(SD)</td>
<td>89.89% (0.0206)</td>
<td>86.05% (0.0751)</td>
<td>81.68% (0.0562)</td>
<td>95.65% (0.0266)</td>
<td>85.52% (0.0835)</td>
</tr>
<tr>
<td>Specificity(SD)</td>
<td>57.00% (0.1208)</td>
<td>64.34% (0.1443)</td>
<td>71.97% (0.1076)</td>
<td>91.03% (0.0419)</td>
<td>76.14% (0.0558)</td>
</tr>
</tbody>
</table>

D. Support Vector Machine

Support Vector Machine [28] is a very popular method for classification problems. The SVM separates the sample set using a hyperplane that is maximally distant from both sides. A Support Vector Machine model using the Radial Basis Function was implemented using the package e1071 in R [29], and parameters Cost C and γ were tuned.

For nonlinear separable case, we are allowing samples to land within or even beyond the margin. The parameter C trades off misclassification of training examples against simplicity of the decision surface. Larger C could help classifying all training examples correctly by selecting more samples as support vectors. But if C is too large, the SVM model will overfit.

The parameter γ of the RBF kernel function controls the width of the kernel. The separating surface will be based on a combination of bell-shaped surfaces centered at each support vector. The parameter γ will be inversely proportional to the width of each bell-shaped surface. Hence γ can be seen as the inverse of the radius of influence of the support vectors. With small γ values, the influence of support vectors are “far” and could affect the whole training set, hence the complexity or “shape” of the decision surface is constrained. By exhaustive grid search, we found using 0.0019 for γ and 100 for cost C can lead to the best cross validation accuracy.

IV. COMPARISON OF FIVE ALGORITHMS

A. Accuracy, Sensitivity, and Specificity

Table II presents the average cross validation accuracy, sensitivity, and specificity with the standard deviation across different folds for each classifier under the optimal parameter setting. The optimal parameter setting is shown in Table III. The Artificial Neural Network achieved the best performance, with an average cross validation accuracy of 93.75%, sensitivity of 95.65%, and specificity of 91.03%. Other machine learning methods performances were also significantly higher than that achieved by using bio-impedance analysis [12].

B. Feature Selection

Feature selection (or variable selection), is to identify the features that are most relevant for the classification problem. In the lymphedema detection problem considered here, features are patient-reported symptoms. A good feature/symptom ranking could enable clinicians to have more insight about the hidden relationship between lymphedema and the symptoms.

Feature selection algorithms typically fall into three categories: filtering, wrapper, and embedded method [32]. Filtering uses a certain criterion to choose or order features independent of the classifier adopted. Wrapper based method evaluates the classification performance obtainable with different subsets of features for a chosen classifier and chooses the subset that yields the highest classification performance using either exhaustive search or some fast search strategies including forward selection, backward selection, and genetic algorithms. Finally, embedded methods embeds feature selection in the classifier training process. We have used each of these methods to rank the symptom features for lymphedema classification.

1) Feature ranking based on minimal Redundancy Maximum relevance criterion (mRMR) [33]: mRMR is a popular filtering method for feature ranking. It combines two constraints, maximizing the dependency/relevance between
2) Feature ranking based on GBM: This is a feature ranking method embedded in the GBM classifier training. From the $M$ iterations of the Gradient Boosting Model, $M$ regression trees are built to fit the gradient of the loss function. For each feature in one single tree, its relevance is defined as the weighted sum of the mutual information between the feature and the target, with the information gain it introduced in each internal node as weight. For multiple trees from $M$ iterations, the relevance of a feature is the average relevance among all the $M$ trees. The feature relevance determined in this manner is shown in Figure 4.

3) Feature ranking based on ANN: Backward feature selection progressively deletes the least useful feature, one at a time. We used this approach together with the neural net classifier. As shown in Figure 5, starting from including all features, in each new iteration, ANN deletes one feature, which hurts the cross validation accuracy the least (or even increase the performance). From this process, those features that contribute the least will be deleted first. In Figure 5, the first point from left is the accuracy after deleting feature No. 9. The last point is the accuracy without using any feature, ANN just predicts based on the ratio of positive samples ($208/355 = 58.60\%$, i.e. random guess). The second to the last one is the accuracy just using feature No. 6, 77.1%. The feature ranking order is the reverse of the feature deletion order.

4) Comparison of different ranking methods: Ranking orders given by the three feature selection methods are summarized in Table IV. In this table, we also show the ranking based on the bivariate Odds Ratio [34]. The Odds Ratio (OR) quantifies the association between the presence or absence of a variable $A$ with the presence or absence of another variable $B$. Higher OR value indicates that a feature is highly correlated with the presence of lymphedema.

Although these different methods provide different ranking orders, the symptom No. 6 “Arm Swelling” is regarded as important by all methods. Other features such as 8 “Chest Wall Swelling” and 25 “Pocket of fluid develop (Seroma Formation)” are also ranked in the beginning of the first three feature selection methods. The relative importance of these symptoms is consistent with the clinical findings. Note that the OR ranking only considers bivariate association between each symptom and lymphedema status. It does not consider the redundancy between two features. This may be one reason that the rankings based on the three examined feature selection methods are not consistent with the OR ranking, except for the first ranked feature.

V. CONCLUSION

Worrying about developing lymphedema or progression to chronic and severe lymphedema has been a focal phenomenon for breast cancer survivors. An e-platform designed for patients self-assessment is feasible and reliable. Providing real-time assessment of lymphedema using self-report symptom features is pragmatic and cost-effective for this much-needed patient care. Our study compared 5 well-known supervised learning algorithms and showed that a well-trained Artificial Neural Network classifier can provide highly accurate assessment of the patients lymphedema status from the reported symptom severity scores, with a cross-validation accuracy of 93.75%, sensitivity of 95.65%, specificity of 91.03%. Furthermore, we have identified symptoms that are most important for lymphedema detection, which are consistent with the clinical observations.

REFERENCES

TABLE IV: Feature Ranking Order Summary

<table>
<thead>
<tr>
<th>nRMR</th>
<th>GBM</th>
<th>ANN</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>23</td>
<td>12</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>16</td>
<td>13</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>22</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>22</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

Department of Electrical and Computer Engineering, Polytechnic School of Engineering, New York University, New York, New York, 2015.


