Development of predictive models for lymphedema prediction by using blood test and chemotherapy data

Xuan-Tung Trinh  
Seoul National University

Pham Ngoc Chien  
Seoul National University

Nguyen-Van Long  
Seoul National University

Le Thi Anh  
Seoul National University

Nguyen Ngan Giang  
Seoul National University

Sun Young Nam (✉ 99261@snubh.org )  
Seoul National University Bundang Hospital

Yujin Myung  
Seoul National University Bundang Hospital

Research Article

Keywords: Lymphedema, Machine learning, Predictive model, Complete blood count, Therapy

Posted Date: September 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2090471/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

**Background:** Lymphedema is a disease that refers to tissue swelling caused by an accumulation of protein-rich fluid that is usually drained through the lymphatic system. Detection of lymphedema is often based on expensive diagnoses such as bioimpedance spectroscopy, shear wave elastography, computed tomography, etc. Applications of data science and machine learning in predicting medical conditions offered support for medical doctors and patients in the early detection of diseases. Although current studies proposed machine learning models to predict lymphedema by using symptoms reported by patients, there might be uncertainty in patient-input data. In this study, we proposed to use more reliable input data such as complete blood count, serum, and therapy data to develop predictive models for lymphedema.

**Methods:** We collected data from 2137 patients, including 356 patients having lymphedema and 1781 patients not having lymphedema. The lymphedema status of each patient was confirmed by clinicians. Data of each patient includes: 1) complete blood count (CBC) test, 2) serum test, and 3) therapy information. We used machine learning algorithms (i.e., random forest, gradient boosting, support vector machine, decision tree and artificial neural network) to develop predictive models on training dataset (i.e., 80% of the data) and tested the models on the test dataset (i.e., 20% of the data). After choosing the best predictive models, we developed web application for medical doctors and clinicians to use our models for quick screening lymphedema patients.

**Results:** A dataset of 2137 patients was collected from Seoul National University Bundang Hospital. Predictive models based on random forest algorithm showed satisfactory performance (balanced accuracy = 86.7 ± 0.9%, sensitivity = 84.3 ± 0.6%, specificity = 89.1 ± 1.5%, precision = 97.4 ± 0.4%, F1 score = 90.4 ± 0.4%, and AUC = 0.931 ± 0.007). A web application was made to assist medical doctors in quick screening lymphedema: [https://snubhtxt.shinyapps.io/SNUBH_Lymphedema](https://snubhtxt.shinyapps.io/SNUBH_Lymphedema).

**Conclusions:** Our study would provide a tool for the early detection of lymphedema and be the basement for future studies predicting lymphedema stages.

Background

Lymphedema refers to a group of pathologic disorders with the excessive accumulation of protein-rich fluid drained through the lymphatic system of the body [1, 2]. The disorders develop from an imbalance between the capacity of the lymphatic circulation and the demand for lymphatic flow. There are two types of lymphedema: primary (i.e., lymphedema due to congenital or inherited conditions) and secondary (i.e., lymphedema triggered by acquired damage that occurs after surgical lymph node dissection). Lymphedema negatively affects patients' quality of life because it leads to adverse outcomes such as pain, arm/leg swelling, and arm/leg heaviness [3–5]. Early detection of lymphedema is essential for disease treatment and minimizing patients’ physical impairment and depression. Traditional methods for detecting lymphedema are a limb circumference measurement, bioimpedance spectroscopy [6, 7], shear...
wave elastography [8], and infrared perometry [9]. These methods would require time and high cost to diagnose hundreds of patients. Machine learning and data science currently gave great support to medical science that many predictive models were able to predict disease outcomes with high accuracy [10–26]. Machine learning-based detection of lymphedema currently assists doctors and patients in real-time monitoring lymphedema [14, 15]. Fu and colleagues proposed an artificial neural network model to predict lymphedema. The model used 26 lymphedema symptom features to predict lymphedema in 355 American patients with an accuracy of 93.75%, sensitivity of 95.65%, and specificity of 91.03%. This model had detection accuracy that was significantly higher than bioimpedance spectroscopy [15]. However, in the study of Fu and colleagues, confirmation of lymphedema status was provided by patient self-report but not by double-check of clinicians, and there was no graphical user interface for assessing the model. Wei and colleagues developed another predictive model for lymphedema based on a logistic regression algorithm [14]. This model used 24 lymphedema-associated symptoms to predict lymphedema in 533 Chinese patients with a sensitivity of 77.1%, specificity of 88.3%, and accuracy of 82.5%. In this study, Wei and colleagues also provided an open-access web application for patients to real-time monitor lymphedema status. Both models of Fu and Wei provided helpful tools for patients and medical doctors to detect lymphedema early. Although those models show good performance in accuracy, sensitivity, and specificity, they have limitations regarding input variables (i.e., symptoms of lymphedema). Lymphedema-related symptoms were previously developed and recognized by researchers [27, 28]. The symptom features include swelling in the arm/hand/breast, heaviness, firmness, tightness, stiffness, pain/aching/soreness, numbness, tenderness, stiffness, redness, blistering, burning, stabbing, tingling, skin toughness or thickness, impaired mobility in shoulder/arm/elbow/wrist/fingers [15]. The model of Wei and colleagues suggested that skin toughness or thickness, tingling (pins and needles), arm limitation, arm swelling, and blistering are the five most important symptoms for the early detection of lymphedema [14]. Patients who have symptoms such as skin toughness or thickness, tingling, and swelling in arm/hand often have high risk to lymphedema. Although it is possible to detect lymphedema by using those human-eye visible symptoms early, the detection might be earlier if we could have models that use human-eye invisible parameters (e.g., blood-related parameters). Complete blood count (CBC) is a common blood test during a routine checkup. The test will provide parameters related to red blood cells, hemoglobin, white blood cells, hematocrit, and platelets. Compared to self-reported symptoms, the CBC parameters are more reliable because they minimize errors caused by patient self-reporting.

Therefore, the objectives of this study were: 1) developing predictive models for lymphedema by using blood test data and machine learning algorithms (e.g., random forest, gradient boosting, logistic regression, decision tree, artificial neural network), 2) providing medical doctors and patients an open-access web application for quick screening of lymphedema.

**Methods**

**Study population**
The approval of this study was obtained from the Institutional Review Board Statement of Bundang Seoul National University Hospital (approval number: B2007-624-101). We collected data from 2137 patients, including 356 patients having lymphedema and 1781 patients not having lymphedema.

**Data collection and cleaning**

The lymphedema status of each patient was confirmed by clinicians and physicians using the medical records of patients. Data of each patient includes: 1) complete blood count (CBC) test, 2) serum test, and 3) therapy information. After cleaning the missing data, we obtained a data table of 28 parameters, including 16 CBC parameters (Table 1), three serum test parameters (Table 2), nine therapy parameters (Table 3), and one lymphedema status parameter. The data table was included in the Supplementary Information.

**Models development and validation**

Models were developed to predict the lymphedema status of patients (i.e., yes or no). For that purpose, popular classification algorithms such as random forest [29], gradient boosting tree [30], C5.0 decision tree, logistic regression [31, 32], and neural network were used for developing the predictive models.

In this study, we used R version 4.2.0 [33] and Rstudio [34] programs to analyze and develop predictive models of lymphedema. Installed and used R packages were: openxlsx [35], svDialogs [36], caret [37], randomForest [38], xgboost [39], C50 [40], nnet [41], shiny [42].

We randomly split the clean dataset into training (80% data) and test sets (20% data). The splitting was repeated to obtain three random splits for mean and standard deviation. The algorithms were applied to the training set via three-fold cross-validation [43], in which the training data was randomly partitioned into three mutually exclusive subsets, with two subsets for training and one for testing. After obtaining trained models, we applied them on the test sets to validate the application of the trained models. Metrics for validating the performance of trained models were: balanced accuracy, sensitivity, specificity, precision, F1 score, and area under the curve (AUC) measured through the receiver operating characteristic (ROC) curve [44].

**Web application for screening lymphedema**

Based on the performance of developed models, we chose the best models for developing a web application. We developed the web application to assist medical doctors in screening lymphedema by using shiny package in R [42]. The address of the web application is: https://snubhtxt.shinyapps.io/SNUBH_Lymphedema. Source code of models and web application is available at: https://github.com/trinhxt/SNUBH_Lymphedema. Detail description of the web application is described in the Result section.

**Results**

**Clinical and histopathological Characteristics**
A dataset of 2706 rows and 29 columns was obtained after data collection and data cleaning. Number of data rows (i.e., 2706) was higher than the number of patients (i.e., 2137) because some patients checked blood/serum/therapy diagnosis several times at the Seoul National University Bundang Hospital. Among 29 columns, 16 columns are CBC test variables (Table 1), three columns are serum test variables (Table 2), nine columns are therapy variables (Table 3), and one column was lymphedema status confirmed by medical records and clinicians. Most of the patients were female (99.6%). The average age of patients was 55.9 ± 11.4; the youngest patient was 27, and the oldest patient was 95 years old. Student t-test was conducted to compare the mean difference between control and lymphedema groups. The p-values of the t-test are shown in Tables 1–3. Box plots of those comparisons were included in the Supplementary Information. Nine numerical variables showing significant difference between control and lymphedema groups are: number of lymph nodes harvested (p-value = 0.000), amount of radiation (p-value = 0.000), radiation fraction (p-value = 0.000), mean corpuscular hemoglobin (p-value = 0.001), mean corpuscular volume (p-value = 0.002), mean corpuscular hemoglobin concentration (p-value = 0.041), hemoglobin (p-value = 0.021), segmented neutrophil (p-value = 0.031), lymphocyte (p-value = 0.018).
Table 1
Summary of blood test data

<table>
<thead>
<tr>
<th>No</th>
<th>Variable</th>
<th>Unit</th>
<th>Full name</th>
<th>Control (n = 2246)</th>
<th>Lymphedema (n = 460)</th>
<th>p-value (t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MCH</td>
<td>pg</td>
<td>Mean corpuscular hemoglobin</td>
<td>29.86</td>
<td>29.44</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>MCV</td>
<td>fL</td>
<td>Mean corpuscular volume</td>
<td>90.43</td>
<td>89.51</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>Lymphocyte</td>
<td>cells/mL</td>
<td>Lymphocyte</td>
<td>32.44</td>
<td>31.39</td>
<td>0.018</td>
</tr>
<tr>
<td>4</td>
<td>Hb</td>
<td>g/dL</td>
<td>Hemoglobin</td>
<td>13.05</td>
<td>12.89</td>
<td>0.021</td>
</tr>
<tr>
<td>5</td>
<td>MCHC</td>
<td>g/dL</td>
<td>Mean corpuscular hemoglobin concentration</td>
<td>32.99</td>
<td>32.86</td>
<td>0.039</td>
</tr>
<tr>
<td>6</td>
<td>Seg.neu</td>
<td>cells/mL</td>
<td>Segmented neutrophil</td>
<td>58.79</td>
<td>59.83</td>
<td>0.040</td>
</tr>
<tr>
<td>7</td>
<td>Hct</td>
<td>%</td>
<td>Hematocrit</td>
<td>39.50</td>
<td>39.20</td>
<td>0.089</td>
</tr>
<tr>
<td>8</td>
<td>Monocyte</td>
<td>cells/mL</td>
<td>Monocyte</td>
<td>6.42</td>
<td>6.58</td>
<td>0.103</td>
</tr>
<tr>
<td>9</td>
<td>Basophil</td>
<td>cells/mL</td>
<td>Basophil</td>
<td>0.46</td>
<td>0.45</td>
<td>0.363</td>
</tr>
<tr>
<td>10</td>
<td>Eosinophil</td>
<td>cells/mL</td>
<td>Eosinophil</td>
<td>1.80</td>
<td>1.75</td>
<td>0.516</td>
</tr>
<tr>
<td>11</td>
<td>MPV</td>
<td>fL</td>
<td>Mean platelet volume</td>
<td>10.16</td>
<td>10.19</td>
<td>0.532</td>
</tr>
<tr>
<td>12</td>
<td>RBC</td>
<td>cells/mL</td>
<td>Red blood cell</td>
<td>4.38</td>
<td>4.39</td>
<td>0.564</td>
</tr>
<tr>
<td>13</td>
<td>ANC</td>
<td>cells/mL</td>
<td>Absolute neutrophil count</td>
<td>3770.44</td>
<td>3795.98</td>
<td>0.744</td>
</tr>
<tr>
<td>14</td>
<td>WBC</td>
<td>cells/mL</td>
<td>White blood cells</td>
<td>6.28</td>
<td>6.25</td>
<td>0.779</td>
</tr>
<tr>
<td>15</td>
<td>PCT</td>
<td>ng/mL</td>
<td>Procalcitonin</td>
<td>0.28</td>
<td>0.28</td>
<td>0.804</td>
</tr>
<tr>
<td>16</td>
<td>PLT</td>
<td>cells/mL</td>
<td>Platelets</td>
<td>272.39</td>
<td>272.98</td>
<td>0.869</td>
</tr>
</tbody>
</table>

Table 2
Summary of serum data

<table>
<thead>
<tr>
<th>No</th>
<th>Variable</th>
<th>Unit</th>
<th>Full name</th>
<th>Control (n = 2246)</th>
<th>Lymphedema (n = 460)</th>
<th>p-value (t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sodium</td>
<td>g/dL</td>
<td>Sodium serum</td>
<td>140.64</td>
<td>140.83</td>
<td>0.096</td>
</tr>
<tr>
<td>2</td>
<td>Chloride</td>
<td>g/dL</td>
<td>Chloride serum</td>
<td>104.24</td>
<td>104.11</td>
<td>0.255</td>
</tr>
<tr>
<td>3</td>
<td>Potassium</td>
<td>g/dL</td>
<td>Potassium serum</td>
<td>4.23</td>
<td>4.22</td>
<td>0.368</td>
</tr>
</tbody>
</table>
Table 3
Summary of therapy data

<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>Unit</th>
<th>Full name</th>
<th>Control (n = 2246)</th>
<th>Lymphedema (n = 460)</th>
<th>p-value (t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>lnn</td>
<td></td>
<td>Number of lymph nodes harvested</td>
<td>8.27</td>
<td>18.58</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>age</td>
<td></td>
<td>Age</td>
<td>55.90</td>
<td>55.83</td>
<td>0.899</td>
</tr>
<tr>
<td>3</td>
<td>fx</td>
<td></td>
<td>Radiation fraction</td>
<td>8.88</td>
<td>13.97</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>Gy</td>
<td></td>
<td>Amount of radiation (Gray)</td>
<td>21.16</td>
<td>32.24</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>sex</td>
<td></td>
<td>Gender</td>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n = 2241</td>
<td>n = 457</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Male</td>
<td>n = 5</td>
<td>n = 3</td>
</tr>
<tr>
<td>6</td>
<td>recon</td>
<td></td>
<td>Breast reconstruction</td>
<td>No reconstruction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n = 1846</td>
<td>n = 416</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TRAM flap</td>
<td>n = 171</td>
<td>n = 24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Implant</td>
<td>n = 229</td>
<td>n = 20</td>
</tr>
<tr>
<td>7</td>
<td>tax</td>
<td></td>
<td>Taxane-based chemotherapy</td>
<td>No taxane</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n = 1124</td>
<td>n = 55</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Type 1</td>
<td>n = 639</td>
<td>n = 179</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Type 2</td>
<td>n = 483</td>
<td>n = 226</td>
</tr>
<tr>
<td>8</td>
<td>che</td>
<td></td>
<td>Chemotherapy</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n = 1311</td>
<td>n = 159</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>n = 935</td>
<td>n = 301</td>
</tr>
<tr>
<td>9</td>
<td>axi</td>
<td></td>
<td>Axilla radiation therapy</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n = 1820</td>
<td>n = 218</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>n = 426</td>
<td>n = 242</td>
</tr>
</tbody>
</table>
Table 4
Comparison of models in this study and previous models

<table>
<thead>
<tr>
<th></th>
<th>This study</th>
<th>Wei 2021[14]</th>
<th>Fu 2018[15]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>RF</td>
<td>LR</td>
<td>ANN</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>86.7 ± 0.9</td>
<td>82.5 ± NA</td>
<td>93.8 ± 0.1</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>84.3 ± 0.6</td>
<td>77.1 ± NA</td>
<td>95.7 ± 0.1</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>89.1 ± 1.5</td>
<td>88.3 ± NA</td>
<td>91.0 ± 0.1</td>
</tr>
<tr>
<td>AUC</td>
<td>0.931 ± 0.007</td>
<td>0.889 ± 0.049</td>
<td>NA</td>
</tr>
<tr>
<td>(N_{\text{test}})</td>
<td>541</td>
<td>160</td>
<td>71</td>
</tr>
<tr>
<td>(N_{\text{train}})</td>
<td>2165</td>
<td>373</td>
<td>284</td>
</tr>
<tr>
<td>No. variables</td>
<td>28</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Type of variables</td>
<td>Blood, serum, therapy</td>
<td>Lymphedema symptoms</td>
<td>Lymphedema symptoms</td>
</tr>
<tr>
<td>Data source</td>
<td>Clinical tests</td>
<td>Patient self-report</td>
<td>Patient self-report</td>
</tr>
<tr>
<td>Lymphedema confirmation</td>
<td>Clinicians</td>
<td>Limb circumference</td>
<td>Patient self-report</td>
</tr>
<tr>
<td>Web application</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Predictive models for lymphedema

By using five algorithms (RF, XGB, C5.0, LR, and ANN), we obtained five predictive models. The performance of those models on the training and test datasets is shown in Fig. 1–2. Among those models, RF model shows the best predictive performance on both training and test datasets, followed by XGB, C5.0, LR, and ANN models. For performance on the training dataset, the RF model shows that balanced accuracy = 99.9 ± 0.1%, sensitivity = 99.9 ± 0.1%, specificity = 99.9 ± 0.1%, precision = 99.9 ± 0.1%, F1 score = 99.9 ± 0.1%, and AUC = 0.999 ± 0.001. For performance on the test dataset, the RF model shows that balanced accuracy = 87.0 ± 0.7%, sensitivity = 82.8 ± 3.2%, specificity = 91.3 ± 3.2%, precision = 97.9 ± 0.7%, F1 score = 90.0 ± 1.6%, and AUC = 0.931 ± 0.007. ROC plots of the five models are shown in Fig. 1.

Important variables to predict lymphedema

Based on model performance, we chose RF model for further analysis of variable importance. The relative importance of variables in deciding the risk of lymphedema for a patient was based on their weights in the RF model and shown in Fig. 3. Among 28 variables of the RF models, the number of lymph nodes harvested (\(Lnn\)) is the most important, followed by taxane-based chemotherapy (\(tax\)), \(age\), and
other variables. The weight of the \( \text{Lnn} \) variable is almost as twice as the weight of \( \text{tax} \) and \( \text{age} \), indicating that the \( \text{Lnn} \) variable is significantly important compared to other variables.

**Applicability domain of predictive models**

The applicability domain of predictive models is the region in the space of model variables (i.e., descriptors) representing the limitation of models toward new data \([45, 46]\). In this study, by following other studies, we used the Euclidean distance method to define the applicability domain of our models \([45–47]\). Visualization of training and test datasets is shown in a t-SNE plot (Fig. 4) by using \textit{sniffer} package \([48–50]\). Data in the test set within the applicability domain defined by the training dataset. If new data has a high Euclidean distance to the training set (over \( 2 \times 10^4 \)), then the prediction on this data would have high uncertainty.

**Web application for screening lymphedema**

A web application available at https://snubhtx.shinyapps.io/SNUBH_Lymphedema, was made by using \textit{shiny} package in R. Users need to upload a dataset to the web so that the model can predict the lymphedema risk of patients (Fig. 5). A template for the dataset was provided in the bottom left of the web (Fig. 5). After uploading the dataset, the model will predict the score for patients based on the probability of lymphedema. A score over 0.25 suggests that a patient might have a high risk of lymphedema and vice versa. Users can select each patient to see his/her predicted score and suggestion related to lymphedema risk.

**Discussions**

There are two current studies that used machine learning algorithms to develop models predicting lymphedema \([14, 15]\) (Table 4). They applied machine learning algorithms such as logistic regression and artificial neural networks and used lymphedema symptoms to predict lymphedema disease. Our random forest model shows better prediction performance than the model of Wei and colleagues regarding the accuracy, sensitivity, specificity, and AUC. However, the model made by Fu and colleagues outperforms our model in terms of accuracy, sensitivity, and specificity. The model in our study uses a dataset that is about five times larger than the dataset of Wei and about seven times larger than the dataset of Fu. A bigger dataset would provide models that could predict a wider range of patients. Therefore, our model would provide a higher potential for future prediction tasks. Our model used blood and serum test data for lymphedema prediction along with therapy data (e.g., number of lymph nodes harvested, taxane-based chemotherapy). Although we could not obtain those parameters as quickly as patient-self-report parameters as in the studies of Wei and Fu, the blood and serum parameters are more reliable than self-report parameters. The blood and serum parameters that were collected by using complete blood count and serum tests of hospitals would have minimal human errors caused by patient-self-reports. Due to the difference in data collection, the model in our study might benefit medical doctors and clinicians in the quick screening of lymphedema, while the models of Fu and Wei might benefit...
patients for real-time self-monitoring of lymphedema. Because patients tend to test blood annually, our models might also be suitable for early detection of lymphedema. Future studies would be conducted to predict stages of lymphedema (i.e., early and late). Additionally, lymphedema confirmation in our study was provided by clinicians and medical doctors based on the medical records of patients. This confirmation is more accurate than the patient self-report and limb circumference as in the studies of Fu and Wei.

Determining lymphedema status is a time-consuming task that clinicians and medical doctors have to make decision based on medical records of patients and expensive diagnoses such as bioimpedance spectroscopy, shear wave elastography, computed tomography, etc. Instead of using those expensive techniques and checking the medical records of every patient, clinicians and medical doctors might use our models and web application for quick screening of patients who might have a high risk of lymphedema. After that, the doctors might further analyze medical records for those patients with a high risk of lymphedema. Because our models used blood and serum data, which could only be obtained from hospitals or clinics, the models might not be suitable for real-time monitoring of lymphedema at home. Patients might use our models but need to test their blood at clinics and collect their own therapy data to predict their lymphedema status.

The dataset in this study includes 2137 patients (356 of them have lymphedema, and 1781 of them do not have lymphedema). We could further extend this dataset and update predictive models in the future so that the models could be applicable to a wider range of patients. This study is the first step in applying machine learning to the detection of lymphedema stages. According to the classification system of the International Society of Lymphology (ISL), there are four stages of lymphedema [5]. Stage 0 refers to a latent or subclinical condition where swelling is not evident despite impaired lymph transport. Stage I represents an early accumulation of fluid that is relatively high in protein content, and that subsides with limb elevation. Stage II signifies that limb elevation alone rarely reduces tissue swelling; pitting is manifest. Stage III encompasses lymphostatic elephantiasis in which pitting is absent and trophic skin changes such as acanthosis, fat deposits, and warty overgrowths develop. Stage classification of hundreds of lymphedema patients using ISL system would be a time-consuming task. Future studies would be conducted to develop predictive models for lymphedema stages.

**Conclusion**

This study successfully developed machine learning models for predicting lymphedema using blood, serum and therapy data. The models based on the random forest algorithm showed good performance on lymphedema prediction; thus, the models could be used to detect lymphedema. Our models used more reliable variables than previous machine learning models of lymphedema prediction (i.e., patient-self report symptoms). An open-access web application was developed to assist medical doctors in quickly screening lymphedema. This study might be the first step in predicting the stages of lymphedema (i.e., stages I, II, III).
# Abbreviations

MCH  Mean corpuscular hemoglobin

MCV  Mean corpuscular volume

Hb   Hemoglobin

MCHC Mean corpuscular hemoglobin concentration

Hct  Hematocriti

MPV  Mean platelet volume

RBC  Red blood cell

ANC  Absolute neutrophil count

WBC  White blood cells

PCT  Procalcitonin

PLT  Platelets

CBC  Complete Blood count

# Declarations

## Ethics approval and consent to participate

This study was approved by institutional review board of Seoul National University Bundang Hospital (IRB number B2007-624-101). All methods were carried out in accordance with the tenets set by the declaration of Helsinki. Informed consent was obtained from all patients and patient data was anonymized to protect confidentiality.

## Consent for publication

Not applicable

## Availability of data and materials

Data are available as Supplementary Files. Source codes of predictive models and web application are available at https://github.com/trinhxt/SNUBH_Lymphedema.

## Competing interests
The authors declare that they have no competing interests.

Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR22C1363)

Authors’ contributions

Xuan-Tung Trinh: Conceptualization, Methodology, Software, Validation, Data curation, Formal analysis, Investigation, Writing – original draft, Visualization, Writing – review & editing. Ngoc-Chien Pham: Validation, Writing – review & editing. Van-Long Nguyen: Writing – review & editing. Van-Anh Thi Le: Writing – review & editing. Ngan-Giang Nguyen: Writing – review & editing. Sun-Young Nam: Validation, Writing – review & editing. Yujin Myung: Writing – review & editing.

Acknowledgements

Not applicable

References


Figures
Figure 1

Receiver operating characteristic of predictive models on training dataset (A) and test dataset (B).

Figure 2

Performance metrics (sensitivity, specificity, accuracy, precision, F1 score, and AUC) of different models on the training dataset (A) and test dataset (B).
Figure 3

Relative importance of variables in deciding the risk of lymphedema for a patient. Results are based on the random forest model. Green variables are therapy, blue variables are serum, and orange variables are CBC parameters. Error bars are standard deviations of triplicates.
Figure 4

t-SNE visualization of training and test data. The size of the scatters is directly proportional to the Euclidean distance of data to the training dataset.
Figure 5

Web application for predicting risk to lymphedema of patients

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20220927Lymphedema.ver.03.Supplementary1.docx