Non-Hodgkin Lymphoma Cell Detection and Patient's Status by Gene Expression Profiling (Using Supervised-Learning)

Roopesh Deepthimahanthi

BTech Computer Science Engineering

Lovely Professional University

April 8, 2020

***Abstract* - This document gives us the strategy behind the detection of Non-Hodgkin Lymphoma by the method gene expression profiling with the help of machine learning especially supervised learning. This is the elaboration of how soft computing techniques can be really useful in the field of medical sciences in discovering diseases and patient's status for a specific disease with the help of previously analysed data.**

INTRODUCTION

Non-Hodgkin lymphoma (NHL, or occasionally simply lymphoma) is a cancer that occurs in white blood cells called lymphocytes, which form part of the body's immune system.NHL is a term that is used by several different types of lymphoma, many of which have some of the same features. Another major type of lymphoma, called Hodgkin lymphoma, is being treated differently.

NHL most commonly attacks adults, but children can develop it too. NHL usually occurs with lymph nodes or other lymph tissue but often it can affect the skin. Lymphoma attacks the lymph system of the body (also known as the lymph gland). The lymph system is an essential part of the immune system that helps combat diseases and certain other disorders. It helps to transfer fluids across the body, too.Lymphomas may begin in any part of the body where lymph tissue is located. The primary lymph tissue areas are:

- Lymph nodes: Lymph nodes are clusters of bean-sized lymphocytes and other cells of the body's immune system, found within the mouth, belly and pelvis. They are linked by lymph vessel network.
- Spleen: The spleen is an organ on the left side of the body, beneath the lower ribs. The spleen contains lymphocytes and other cells within the immune system. It also retains healthy blood cells, and cleans out infected blood cells, viruses, and waste cells.
- Bone marrow: The spongy material between most bones is the bone marrow. That is where new blood cells are created (including a few lymphocytes).
- Thymus: The thymus is a small organ that sits behind the upper part of the breastbone and before the nucleus. It is essential in formation for T Lymphocyte.
- Adenoids and tonsils: There are lymph tissue clusters at the back of the mouth. They tend to produce antibodies against breathing-in or swallowing germs.
- Digestive tract: There's also lymph tissue in the stomach, intestines and many other organs.

Brief description of above types can be showcased in the below diagram

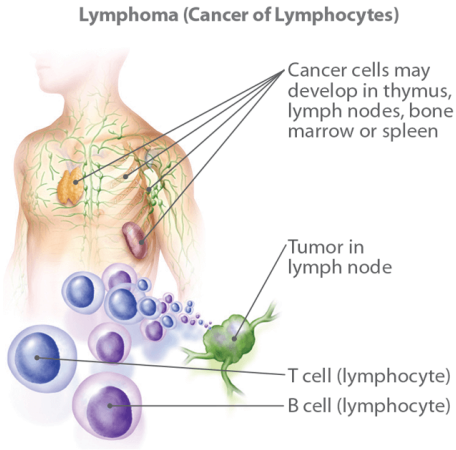

Fig.1. Lymphoma basic information diagram

LITERARTURE REVIEW

Different models have shown different accuracies on

DLBCL patients. According to a research 80% of patients are diagnosed correctly with current detection technologies and are having a prolonged survival after diagnosis. In the study done earlier, artificial neural network has been used to classify patients with DLBCL on the basis of their gene expression profiles. Finally, it has been attempted to extract a number of genes that their differential expression were significant in DLBCL subtypes. Methods: 40 patients and 4026 genes were studied. In this study, Firstly in the whole sample genes were ranked based on their signal to noise (S/N) ratios. Then a suitable threshold was selected, according to which samples having less value than threshold was removed. Then PCA was used for further reduce the number of samples and Perceptron neural network was used for classification of these patients. Some appropriate genes based on their prediction ability were extracted.
Results: Various targets considered for patients classifying. Thus patients were classified based on their 5 years survival with accuracy of 93%, in regard to Alizadeh et al study results with accuracy of 100%, and regarding with their International Prognosis Index (IPI) with accuracy of 89%.
Conclusion: According to research combining PCA and S/N ratio has proven to be an effective method for the reduction of the dimension and neural network is proven to be robust tool for classification of patients according to their gene expression profile.

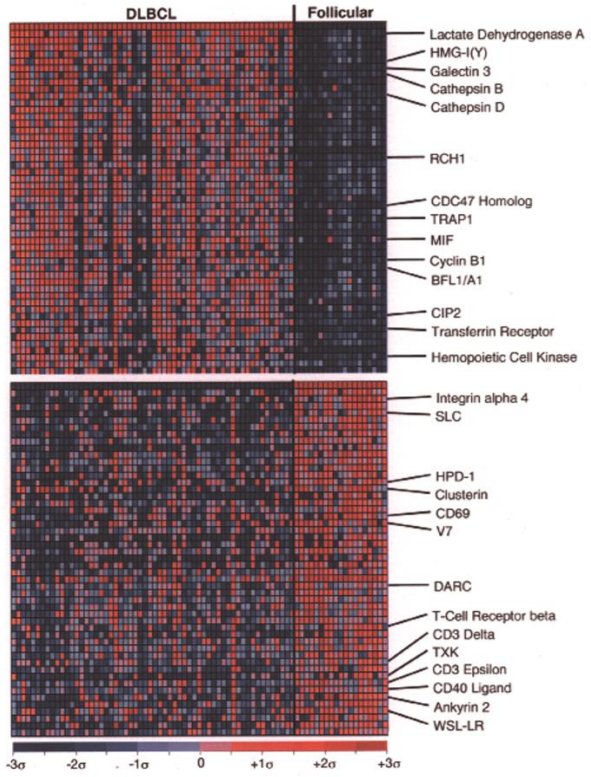

Fig. 2. Features comparison across different samples For DLBCL vs FL

Another model has been proposed for prediction, with artificial neural network (ANN) classifier. Due to the limited amount of training data and the fact that one output was needed model was limited to linear Perceptron (LP) with 10 input nodes representing the PCA components. Using more than 10 components did not improve the classification of the samples.
Classification steps are as bellow:

At first, the 40 labelled samples were randomly split into 3 equal groups. 2 groups were used as trained and one group was used as test. Since there were not enough samples available, a leave one out cross validation on 26 training samples was performed. In which one sample is used for testing, and others are used for training the predictor, the testing sample is classified by this predictor, and the process is repeated iteratively. In this method, 26 networks were trained. In the final step, these 26 models were tested on 14 blinded test samples. Then we used average committee vote to classify these 14 test samples. It means that the outputs of 26 networks were averaged on each of the 14 test samples and this average forced to 0 or 1.

PROPOSED METHODOLOGY

The process starts with getting samples of patients, the 77 patients' data have been used. The process applied is same as the process for distinguishing DLBCL from FL with supervised learning. Long-term clinical data was available for all 58 DLBCL as well as remaining FL patients in the study. These patients were divided into two groups: those with cured disease (n = 32) and those with fatal or refractory disease (n = 45). Then different supervised learning classification approaches have been used to develop a Cell and Curable or Fatal outcome predictors. The algorithms will be compared for their accuracies, to determine the optimal algorithm finally to be adopted for the next stage of work.

First the data goes through Feature Extraction then the extracted Features are put through different standardization techniques. To increase the accuracy, firstly we converted categorical data into numerical values using label encoder from sklearn and standardized the whole featured data of training and testing and then applied Synthetic Minority Oversampling Technique(SMOTE) to upscale the data and finally using Extra trees classifier from sklearn to fetch top 20 features from the data to enable easy learning .After the segmentation of lymphocytes soft computing classifiers are used to classify using the extracted features such as area, perimeter, convex area, solidity, major axis length, orientation filled area, ratio between cell and nucleus area, mean gray level, rectangularity and circularity can be used for the classification of lymphocytes. In the proposed model 7129 features down-scaled to 20 have been used of 77+ upscaled patient samples. The data has been classified for two targets DLBCL/FL and cured/fatal.

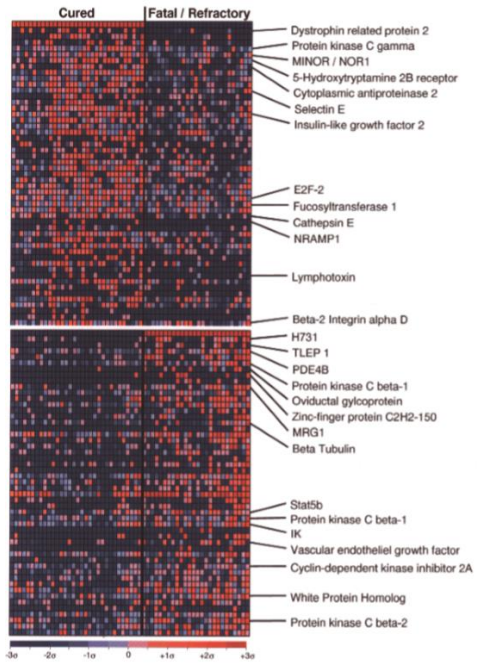

Fig. 3. Features comparison across different samples for Cured vs Fatal

The classifier used must achieve lowest miss-classification error with lowest standard deviation. For these 5 different classifiers have been deployed like K-Nearest Neighbors, Decision Tress, logistic Regression, Random Forest, Support Vector Machine and ultimately a Stacking Classifier using all the above models. Then the classifier giving the best result has been choosen for further work. To avoid the statistical problem of over-estimating prediction accuracy that occurs when a model is trained and evaluated with the same samples, we used a 'leave-one-out' cross-validation testing method.

RESULTS AND DISCUSSION

In this model 77 patients with DLBCL (n = 58) or FL (n = 19) were subjected to transcriptional profiling using oligonucleotide microarrays containing probes for 6,817 genes. The samples were sorted according to their degree of correlation with the DLBCL versus FL distinction. Genes were found more at higher levels in DLBCL patients than in FL patients. . The trained predictor is effective in predicting the outcome of subsets of DLBCL patients. Further the gene-expression–based outcome predictor was investigated if it contained additional information not captured by the IPI. Patients with the 'cured' gene-expression signature had significantly higher OS rates than patients with the 'fatal/refractory' signature.

Total of 5 models have been trained on the dataset that are K-Nearest Neighbors, Decision Tress, logistic Regression, Random Forest, Support Vector Machine and on top, a Stacking Classifier. But K Nearest Neighbour algorithm has a drawback that it needs user to pre assign the k value. More work can be done to overcome this drawback. Larger dataset can be used and applied to the system in order to test the results.

Results of DLBCL/FL

K Nearest neighbour classified the data with accuracy of 80% where Logistic Regression is providing accuracy of 95%. Support Vector Machine is providing accuracy of 91%. Decision tree is providing accuracy of 83%. Random Forest Classifier is providing accuracy of 91%.

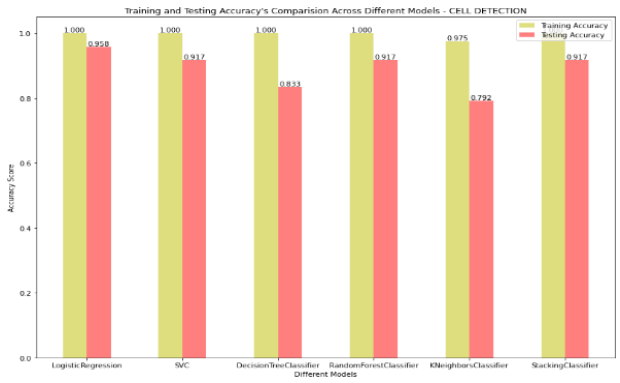

Fig. 2. Graph Depicting Accuracy Comparison of Cell Detection

Results of Cured/Fatal

K Nearest neighbour classified the data with accuracy of 66% where Logistic Regression is providing accuracy of 54%. Support Vector Machine is providing accuracy of 50%. Decision tree is providing accuracy of 66%. Random Forest Classifier is providing accuracy of50%.

Best Classifier for DLBCL/FL data segmentation is Support Vector Machine. And for Cured/Fatal data segmentation best classifier is KNN or Decision Tree.

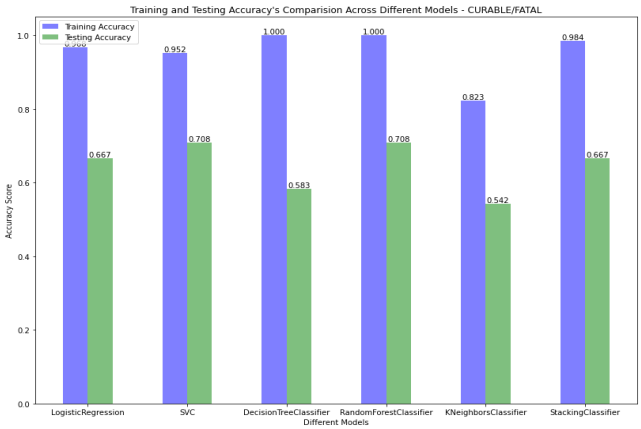

Fig. 3. Graph Depicting Accuracy Comparison of Cured/Fatal

The Specifications of proposed modal for DLBCL/FL are as follow

- Model is preprocessed with Standardization, Minority oversampling and then choosing top 20 features from 7129 features.
- K has been taken as 5 in K-Nearest-Neighbors Classification
- In Logistic Regression max iterations are 200
- In Support Vector Machine degree has been taken as 3 and kernel as radial basis function
- Decision tree has been trained with max_depth as 10, min_sample_leaf as 1 and criteria as Gini
- In Random forest n_estimators have been taken as 150.

The Specifications of proposed modal for Cured/Fatal are as follow

- Model is preprocessed with Standardization, Minority oversampling and then choosing top 20 features from 7129 features.
- K has been taken as 5 in K-Nearest-Neighbors Classification
- In Logistic Regression max iterations are 200
- In Support Vector Machine degree has been taken as 3 and kernel as radial basis function
- Decision tree has been trained with max_depth as 10, min_sample_leaf as 1 and criteria as Gini
- In Random forest n_estimators have been taken as 150.

CONCLUSION

In this paper we have proposed a technique for detection and classification of Non-Hodgkin Lymphoma into its types and patients status as cured or fatal. Dataset of 77 patients with 7129 features has been used. Genes expressed at higher levels in DLBCL patients than in FL patients K Nearest Neighbor is used for classification. Various models have been deployed and as well as various feature selection and extraction process took place to avoid overfitting of models. Various statistical features are extracted for classification purpose. Different segmentation and classification algorithms are used. Best Classifier for DLBCL/FL data segmentation is Logistic regression, and for Cured/Fatal data segmentation best classifier is SVM or Random forest.

REFRENCES

[1] Wikipedia - https://en.wikipedia.org/wiki/NonHodgkin_lymphoma.
[2] Cancer.org - https://www.cancer.org/cancer/non-hodgkin-lymphoma/about/b-cell-lymphoma.html.
[3] Feature Details - www.genome.wi.mit.edu/MPR/lymphoma
[4] National Cancer Institute – Dataset - https://llmpp.nih.gov/lymphoma