#10

Improved Biomarker-Based Diagnostics of Leukemia Subtypes using Machine Learning Methods

Katherine Schmidt[1], Purnima M. Kodate[2], and Kirti M. Yenkie[3*]

[1]Department of Mathematics, Rowan University, Glassboro, NJ, USA

[2]Department of Pathology, Government Medical College (GMC), Nagpur, India

[3]Department of Chemical Engineering, Rowan University, Glassboro, NJ, USA

Introduction & Motivation

- With cancer being a leading cause of death worldwide, it is of utmost importance to detect it at an early stage
- One possible way to detect cancers early and recognize its correct sub-type is through the use of biomarkers
- Clinical diagnosis and classification of Acute Leukemias is primarily performed via morphological identification and cytochemistry analysis, while the final confirmation is obtained via immunophenotyping (IPT)
- The former two methods are available at GMC Nagpur but they have to send samples to Tata Memorial Mumbai for the IPT test
- Thus, ***we have used the clinical data collected from morphological tests and cytochemistry analysis and applied machine learning methods to identify the most critical variable (biomarker) for the classification of leukemia subtypes***

Myeloperoxidase (MPO):

- MPO is an enzyme found in the white blood cells that is released when the arterial wall is inflamed[1]
- Significant in classifying 5 out of the 8 subtypes of AML[2]
- A 2012 study[3] found that MPO expression in AML may help to identify patients that will benefit from stem cell transplantation

Clinical Data

The data consists of 174 observations of 15 variables. Of these variables, ten are binary, four are numerical, and the dependent variable (immunophenotyping diagnosis) is categorical with five levels

Table 1: Description of variables in this study (Kodate et. al., 2018)

Variable	Abbreviation	Type	Levels
Age	Age	Numerical	
Gender	Gender	Binary	Male or female
Hepatomegaly	H	Binary	Present or absent
Splenomegaly	S	Binary	Present or absent
Mediastinal mass	Me	Binary	Present or absent
Bleeding	B	Binary	Present or absent
Total leucocyte count (mcl)	TLC	Numerical	
Hemoglobin (gm/dl)	Hb	Numerical	
Platelet count (mcl)	Plt	Numerical	
Periodic acid Schiff	PAS	Binary	Present or absent
Myeloperoxidase	MPO	Binary	Present or absent
Immunophenotyping diagnosis	IPT	Categorical	AML, ALL, MPAL, Normal, Inconclusive

- Hepatomegaly: enlargement of the liver
- Splenomegaly: enlargement of the spleen
- Mediastinal mass: benign or cancerous growth of the chest area, separating the lungs
- Total leucocyte count: can signify the presence of leukemia when rapidly produced (another name for white blood cells)
- Hemoglobin: protein in red blood cells that carry oxygen throughout the body - can signify leukemia in low counts
- Platelet count: low counts can signify leukemia/chemotherapy
- Periodic Acid Schiff: displays glycogen and mucopolysaccarides, patterns can be characteristic of lymphoblastic leukemia
- Immunophenotyping diagnosis: a blood sample that may detect leukemias through the presence or absence of white blood cell antigens

  Levels: Acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), mixed-phenotype acute leukemia (MPAL), normal, inconclusive

[1]"MPO-Practitioner-One-Pager-CHL-D006b.pdf." .

[2]WHO Classification of acute myeloid leukemias (AML). Retrieved April 23, 2019, from http://hemepathreview.com/WHO-Review/Outline/Chapter4.htm

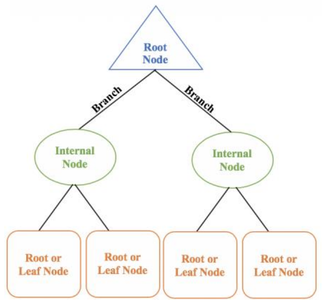
[3]Y. Kim, S. Yoon, S. J. Kim, J. S. Kim, J.-W. Cheong, and Y. H. Min, "Myeloperoxidase Expression in Acute Myeloid Leukemia Helps Identifying Patients to Benefit from Transplant," Yonsei Med J, vol. 53, no. 3, pp. 530-536, May 2012.

Machine Learning Methods

- 16 models were fitted: 8 decision trees & 8 artificial neural networks (ANN)
- Each method was fitted for four different training and testing levels as well as each of those were fitted with and without MPO as a predictor variable
- The overall goal is to analyze the effect of MPO in classifying AML in comparison to ALL, MPAL, normal and inconclusive results

Method#1: Decision Trees

- Eight decision trees were fitted
- The original tree was then pruned using a k-fold cross validation method to reduce overfitting
- The lowest cross-validation error corresponds to a complexity parameter, which is then used to build a tree with a given number of terminal nodes



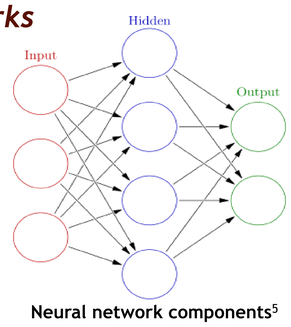
Decision tree components[4]

Method#2: Artificial Neural Networks

- Eight artificial neural networks were fitted
- The activation used is the sigmoid function that has the following form:

$$\varphi(x) = \frac{1}{1 - e^{-x}}$$

- One, two, and three hidden layers were all tested. One hidden layer performed the best in all the eight cases



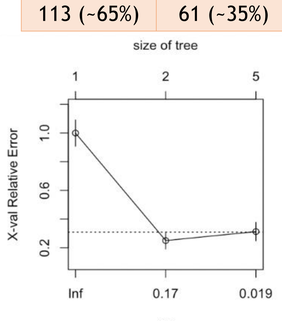
Neural network components[5]

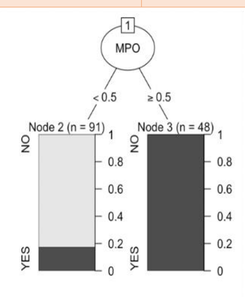
Results

Method#1: Decision Trees

Table 2: Accuracy of decision tree models

Dataset sizes		Model with MPO		Model without MPO	
Training	Testing	Terminal Nodes	Accuracy	Terminal Nodes	Accuracy
139 (~80%)	35 (~20%)	2	97.14%	5	91.43%
130 (~75%)	44 (~25%)	2	97.14%	7	75%
122 (~70%)	52 (~30%)	2	97.14%	3	82.69%
113 (~65%)	61 (~35%)	2	97.14%	5	73.77%



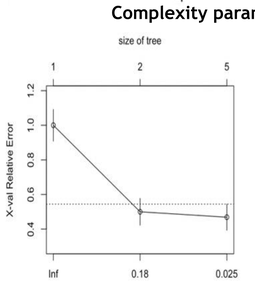


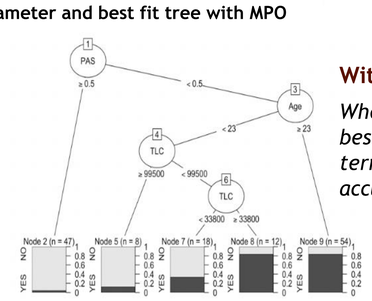
Complexity parameter and best fit tree with MPO

With MPO

In decision tree with MPO, all training and testing sets perform the same, with the optimal tree containing two terminal nodes

*The overall accuracy of the model is **97.14%***





Complexity Parameter and best fit tree without MPO

Without MPO

*When MPO is removed, the best fit tree contains 5 terminal nodes and has an accuracy of **91.43%***

[4]Sa, J.A.S.; Almeida, A.C.; Rocha, B.R.P.; Mota, M.A.S.; Souza, J.R.S.; Dentel, L.M. Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy

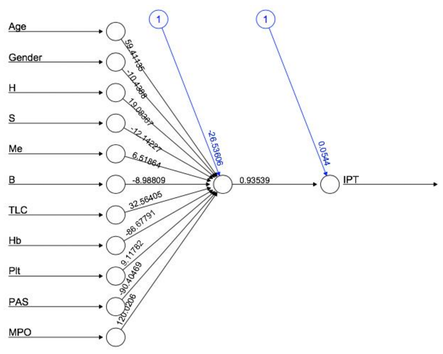
[5]https://en.wikipedia.org/wiki/Artificial_neural_network#Components_of_an_artificial_neural_network

Results

Method#2: Artificial Neural Networks

Table 3: Accuracy of artificial neural networks

Dataset sizes		Model with MPO	Model without MPO
Training	Testing	Accuracy	Accuracy
139 (~80%)	35 (~20%)	1.00	0.885714
130 (~75%)	44 (~25%)	0.977272	0.909091
122 (~70%)	52 (~30%)	0.961538	0.788462
113 (~65%)	61 (~35%)	0.950820	0.836066

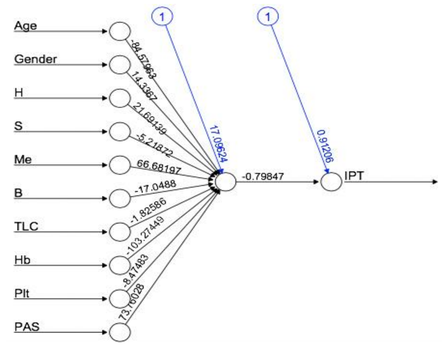


Best fit ANN with MPO

With MPO

*All ANN fit contain one hidden layer. The neural network with MPO has an accuracy of **100%** at the 80/20 training-testing split*

In the model, MPO has the highest weight at 120.02. This is significant in this case, as there is only one hidden layer



Best fit ANN without MPO

Without MPO

The neural network without MPO performs significantly worse than the previous model

*The accuracy drops from 100% to **90.91%**, again signifying the importance of MPO as an indicator of AML*

Conclusions

- A total of 16 models were fitted with the highest accuracy being 100% in classifying AML against ALL, MPAL, normal, and inconclusive patients
- The highest accuracy model was the artificial neural network with 80% training and 20% testing data split
- MPO was clearly a significant indicator of the presence of AML, as it had decreased the accuracy greatly in both cases when removed from the model

Acknowledgments

- Rowan University IRB (Institutional Review Board) for carefully evaluating the study protocol
- The Rowan University Seed Funding Program
- Chemical Engineering Department at Rowan University, NJ, USA
- Hematopathology Department at the Government Medical College in Nagpur, Maharashtra, India for providing resources for collection of the diagnostic data

Contact Information:

Katherine Schmidt - schmidt6@students.rowan.edu

Dr. Purnima Kodate - mkpurnima@gmail.com

Dr. Kirti M. Yenkie - yenkie@rowan.edu

Sustainable Design & Systems Medicine Lab: https://yenkiekm.com/