ProgMod: An Analytical Model for Prognosis Prediction of AML Patients Using Survival Regression and Gene Expression Levels

Ahmad Al Sayyid, Rafiqul Haque, Yehia Taher, Sara Makki, Ali Jaber

ProgMod: An Analytical Model for Prognosis Prediction of AML Patients Using Survival Regression and Gene Expression Levels

Ahmad Al Sayyid
Department of Computer Science
Lebanses Univesity – Faculty of Sciences
Beirut, Lebanon
aelsayyid94@gmail.com

Rafiqul Haque
Department of Technological Development
Intelligencia R & D
Paris, France
rafiqul.haque@intelligencia.fr

Yehia Taher
Department of Computer Science
Universite de Versailles St-Quentin-en-Yvelines
Paris, France
yehia.taher@uvsq.fr

Sara Makki
Department of Technological Development
Cognitus
Paris, France
sara.makki@cognitus.fr

Ali Jaber
Department of Computer Science
Lebanses Univesity – Faculty of Sciences
Beirut, Lebanon
alijaber30@hotmail.com

Abstract – An accurate prediction of prognosis to the patient diagnosed with Acute Myeloid Leukemia (AML) is an enormously difficult task. Several solutions have been proposed for prognosis prediction however there is a scope to improve current solutions. In this paper we aim at developing a solution that estimates the survival time that is the Prognosis of patients diagnosed with AML. To that end, we used a machine learning model that is built on an algorithm called Survival Regression. The model consumes as input the Expression Levels of a small number of the genes of the patient.

Keywords - AML, Gene Expression Levels, Machine Learning, Prognosis, Survival Regression

I. INTRODUCTION

Acute Myeloid Leukemia (AML) is a form of cancer that is characterized by infiltration of the bone marrow, blood, and other tissues by proliferative, clonal, abnormally differentiated, and occasionally poorly differentiated cells of the hematopoietic system. Although it was incurable 50 years ago, AML is now cured in 35 to 40% of adult patients who are 60 years of age or younger and in 5 to 15% of patients who are older than 60 years of age. The outcome in older patients who are unable to receive intensive chemotherapy without unacceptable side effects remains dismal, with a median survival of only 5 to 10 months [1]. Producing an accurate prognosis to the patient diagnosed with AML difficult due to its enormous molecular heterogeneity, which means that the disease manifests itself with wide diversity on the molecular level. However, now with the medical insight provided by the genetic profile of the patient, the possibility of acquiring a prognosis by studying his genes is promising, especially with powerful algorithms that mainly developed within ML field. Numerous machine learning algorithms are available. However, there is no *one size fits all* algorithm that can be used in developing a model for complex analysis. Typically, the algorithms serve specific purposes or specific domain of interest. Therefore, to build an analytics model, a modeler follows trial-and error principle to discover the best fitting algorithm. For instance, a wide number of machine learning algorithms and statistical methods have been proposed in a large body of literature such as [2] [3] [4] [5] [6] [7]. However, the algorithms proposed in these literature produce different results and different levels of accuracy. Furthermore, it is commonly seen that the existing machine learning algorithms very often cannot be used as is, but need to be extended. That being said, building a complex analytics model is not a straightforward operation; it needs testing various methods and algorithms to find best fit in terms of accuracy and correctness.

We studied a number of existing analytics models for prognosis. According to our investigation, these models are built on the top of different statistical methods and machine learning algorithms. The study shows that the accuracy and correctness of the results produced by these models vary greatly. Also, we strongly believe that the accuracy level achieved by the existing solutions is *low* and can be improved.

Our objective is to develop and train an efficient machine learning model that can estimate the survival time, or the prognosis, of a patient diagnosed with AML, with high accuracy, using solely the Gene Expression Levels of a small group genes. The small number of genes means that the testing time and costs will be significantly reduced, as the gene expression levels of such number of genes can be measured using low-multiplicity technologies. The challenge will be to identify from thousands of genes the ones that their expression level can be used effectively to give an accurate prognosis. And the algorithm that the model will depend on is Survival Regression.
The rest of this paper is organized as follows: the second part is a background on some of the concepts needed for the understanding of the contribution. The third part is a literary view on the subject of machine learning in healthcare. The fourth part is our contribution and the fifth is results. The last part will be a conclusion.

II. BACKGROUND

A. Gene Expression Levels

Gene Expression is the process by which the information from a gene is used in this synthesis of a functional gene product such as proteins. Gene expression is context-

dependent and is regulated in several basic ways: by region, dynamic response to environmental elements, by gene activity, and in *disease states*. So it can be said that gene expression level is one way of mapping his biological functions, and how well his body is being regulated by his genes. That is why the amount of gene expression, or the Gene Expression Level of the genes of a certain individual can give insight into his biological and thus medical profile.

B. Survival Regression

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical system. This is done by studying the relationship between the Explanatory Variables that lead to that event in a certain group and the Outcome Variable. For example, in cancer survival analysis, the event to be observed is the death of the patient. The time of survival of the patient is "regressed" against the variables that are under study, and the prognosis of the patient, or the duration of time before his death, is the outcome variable.

Survival models can be viewed as consisting of two parts: the underlying baseline function, often denoted $\lambda_0(t)$, describing how the risk of event per time unit changes over time at *baseline* levels of covariates; and the effect parameters, describing how the hazard varies in response to explanatory covariates [8].

One of the branches of survival analysis is the survival analysis under the *Proportional Hazards Conditions*. The proportional hazards condition states that covariates are multiplicatively related to the hazard. Sir Cox, whom the Cox Proportional Hazard Model is attributed to, observed that if the proportional hazards assumption holds, then it is possible to estimate the effect parameter(s) without any consideration of the hazard function. This approach to survival data is called application of the Cox proportional hazards model. Cox also noted that biological interpretation of the proportional hazards assumption can be complicated [8].

III. Related Work

Several works related to prognosis analysis have been found in a large bodies of literature. We reported some of the notable works related to our research.

Sara Haddou Bouazza et al. [9] discussed a comparison between five feature selection algorithms to extract the most significant features from the Gene Microarray Data: F test, T test, Signal to noise Ratio (S/R), ReliefF and Pearson Product-Moment Correlation Coefficient (CC). They used datasets of five cancers: Leukemia, Lung, Lymphoma, Central Nervous System and Ovarian, and five supervised learning classifiers: K Nearest Neighbors (KNN), Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Decision Tree for Classification (DTC), and Naïve Bayes classifier (NB). They concluded that the selection methods with highest accuracies across classifiers where: S/R, ReliefF, and CC respectively.

Feiyu Xion et al. [10] constructed a machine learning framework called KITML (Kernelized Information-Theoretic Metric Learning), which depended on the KNN algorithm but with improved distance metrics to diagnose cancer by finding similar patient biological profiles while dealing efficiently with the high dimensionality of microarray gene data. The performance of the algorithm was compared to others including KNN with Euclidian Metric, SVM, Random Forest (RF), and DT. The Macro Average F1 score was used as an evaluation metric and it showed higher accuracy of the algorithm, with considerable less execution time.

Sara Tarak et al. [11] classified cancer types by studying the Gene Expression Levels in 3 datasets: leukemia, breast, and colon using the Matlab Bioinformatics and Statistics Toolboxes. They used the KNN algorithm, with K=3 and built five classifiers each employing a feature selection method. The feature selection methods were chosen to avoid redundancy and noise and are: Backward Elimination Hilbert-Schmidt Independence Criterion (BEHSIC), Extreme Value Distribution (EVD) gene selection, and Singular Value Decomposition Entropy (SVDEntropy) gene selection. Tuned their results using Error Estimation, and evaluated them by using Receiver Operating Characteristic (ROC) and Bayesian Credible Interval (BRI) methods.

Kenneth R. Foster et al. [12] published a paper in which they discussed the great challenges facing the construction of classifiers in the field of bioinformatics, arguing they are mostly prone to over-fitting, and generally lack accuracy, and face a great challenge in the vast number of parameters that can interfere in the diagnosis process that are not taken into account. They stated that building classifiers should not be views simply as an add-on statistical analysis, but a parcel of the experimentation process, and that the validation of the classifiers for diagnostic applications should be considered as part of a much larger process of establishing the clinical validity of the techniques. The article mainly focuses on methods to improve the accuracy of SVM classifiers in bioinformatics unavoidable.

Discussion

The technologies studied in the above tackle the problem of prognosis much lesser than diagnosis, as there are many solutions proposed on cancer diagnosis using different approach like use of medical imaging and Gene Expression Levels, however we found little about the prognosis problem. To the best of our knowledge, there is no solution proposed to tackle the problem of prognosis through the use of Gene Expression Levels alone, which have the potential to overcome the difficulty of producing accurate models, by eliminating much of the diversity in patients and their data, by focusing on data from a single, highly reliable source; their genetic profiles.

Furthermore, the implementation of the solutions proposed in literature the is done using classical solutions like Matlab and other statistical analysis tools rather than using advanced tools and technologies. This has limitations from a technical standpoint. Firstly, modification of any algorithm extremely difficult, very often not possible; the modeler relies on as is library provided by the development framework. Hence, improvement of a model is a non-trivial task. Optimization of performance in terms of computation is entirely impossible. Sometimes, technologies of these sorts provide some parameters to optimize performance.

However, very often it is not an effective approach. Specifically, for large-scale datasets, the existing solutions cannot be used meaning that the analysis cannot be performed these existing solutions. To the best of our knowledge, the scale of the dataset is critical to produce a comprehensive result of analysis – which is not possible with existing solutions.

IV. ProgMod – A Model for Prognosis of AML Patients

In this section, we present *ProgMod* for finding the prognosis of AML patients using solely their gene expression levels. The process of model construction was sequential and iterative (shown in Figure 1), in order to find the model with the highest score. The number of features of different models varied as the aim was to find an optimal number that is preferably below 48.

A. ProgMod Development Cycle

The ProgMod development lifecycle consists of four phases:

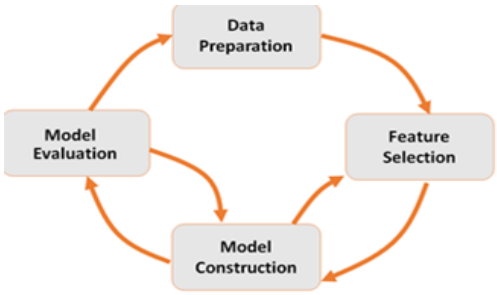

Figure 1 - Sequential and Iterative Development Cycle of ProgMod

Data Preparation, Feature Selection, Model Construction, Model Evaluation. These tasks briefly described in the following.

1) Data Peparation: The dataset we worked with was of 240 AML patients, and is the combination of 2 datasets. 1) GSE 12417 (78 Patients) 2) TCGA (162 Patients). The information available about the patinets were three types: 1) Social Data (Age and Gender) 2) Clinical Informartion Data (Blast Count, Overall survival time osTime, and overall survival status osStatus) 3) Genetic Profile Data (Gene Expression Levels). The first step was to clean the data and remove the redundant information and extract the ones needed i.e. osTime, osStatus, and Gene Expression Levels.

2) Feature Selection: This was done in two different phases:

a) Preliminary Feature Dimensinonality Reduction: The process of applying variance thresholding to remove features that have low variance regardless of the value of the outcome variable i.e. the overall survival time. This reduced the number of featurs from *20000* to *287*.

b) Outcome Reliant Feature Selection: Select from the reduced variables the ones that best correlate to the outcome variable. Three features selection methods were used and tested in order to find the best one, they are: 1) F test 2) Recursive Feature Elimination 3) Mutual Info Selection.

Sklearn's feature selection module can be used for feature selection/dimensionality reduction on sample sets, either to improve estimators' accuracy scores or to boost their performance on very high-dimensional dataset

All the features that had a very low variance in 80% of the patients were dropped. The result of this process was to reduce the number of features from about 20,000 to 287, these are the ones that will be used in the next step.

The second step is outcome reliant feature selection i.e. selecting the features that best correlate to the outcome variable, the survival time. First, and since the two dataset have different methods of measurements of Gene Expression Levels, they could not be used merged together to perform feature selection. Instead, the larger dataset, TCGA, containing 178 patients, was used as a features selection set, and the second dataset was used as a testing set.

Three feature selection methods were tried, evaluated, and the one that had the best results was eventually used in the final model. The three feature selection methods are all modules in sklearn's feature selection module**.** They are the modules:

- *f_regresion*: the f_test for regression tests whether any of the independent variables in a multiple linear regression model is significant using the ratio of variances obtained from the means squared value. [1]
- *rfe_regression*: recursive feature elimination recursively removes features, builds a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable.[2]
- *Mutual_info_regression*: Mutual information is a measure between two (possibly multi-dimensional) random variables XX and YY that quantifies the amount of information obtained about one random variable, through the other random variable. [3]

The pseudo-code for the algorithm used to select the best model was the following:

```
for num_features in range (10, 48):
      features = select_features
(num_features, Feature_Selection_Set)
  model_score = evaluate_model
   (features, Test_Set_1, Test_Set_2)
```

3) Model Constrction: Cox Proportional Hazard Regression Model is used for developing analytics model. It is probably the most popular regression technique for regression analysis of survival data *[8]*. The power of the Cox Model, and what distinguishes it from linear or logistic regression, is its ability to account for censored data points, i.e. data points that the time of event remains unknown. In our example, if the patient that did not die at the end of the standard five year follow up period, his overall survival status is labeled as "censored event". The Cox Model does not assume the hazard value to be constant, but rather a function of time. This is due to the fact the hazard in Cox Model is the same as the incidence rate i.e. the ratio of the subjects who died over the overall number of subjects ((*number of death events)/(total number of patinets*)), thus as time passes and subjects die, the incidence rate, or the hazard, varies. This keeps the Model in continuous variation, with the incidence rate and model predictions variating as data points are

[1] http://facweb.cs.depaul.edu/sjost/csc423/documents/f-test-reg.htm
[2] https://medium.com/@aneesha/recursive-feature-elimination-with-scikit-learn-3a2cbdf23fb7
[3] https://thuijskens.github.io/2017/10/07/feature-selection/

modified or added to it. If *Xi = {X1, X2 ... Xn}* are the values of the covariates of subject i, the equation for the Cox Model:

$$\ln incidence = \beta_0 + \sum \beta_i (X_i)$$

This equation produces a survival function for each patient, showing the probability of survival over a period of time (Figure 2).

4) Model Evaluation: As mentioned the Cox Model predicts a survival function of the probability of the event of

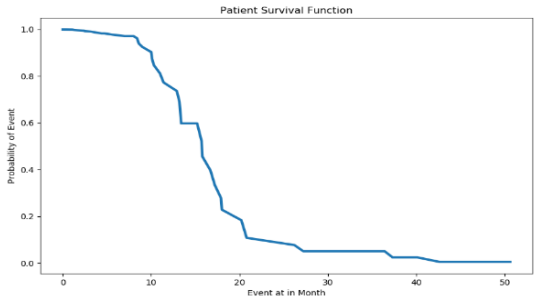

Figure 2 - Patient Survival Function

death over time. While this result is ultimately desired, it is difficult to evaluate the accuracy of the model when the real value is a number and the predicted value is a function. So an algorithm was improvised.

Each model was evaluated by fitting the data with the features resulting from the features selection process in to the testing set and scoring the model. The score of each model was acquired using the module *r2_squard_accuracy* in sklearn's module**:** *metrics*. R2 Squared Accuracy is an algorithm used to predict how well the model will perform on future unseen data. It takes two matrices, one with the real values and one with the predicted values.

The real values used were the actual survival months of the patients. The predicted values were the values in the survival function corresponding to constant probability value. The probability value is irrelevant in this case because the purpose is comparison between different models. The predicted survival value for each patient was the value corresponding to 0.8 probabilty in the survival function.

B. *Implementation of ProgMod*

ProgMod was implemented using different Python-based, open-source libraries available to perform data acquisition and transformation (Pandas and Numpy), analysis (Scikit Learn), and visualization (PyPlot).

For the implementation of the Cox Proportional Hazard Model, we used the open-source library lifelines, downloaded from the lifelines website[4], and performed some tweaking and a wrapper around its functions to suit the needs of my implementation. Lifelines is a library that implements survival analysis models, including the Cox Proportional Hazard Model.

V. Experiments and Results

In this section, we reported results of experiments that we conducted to evaluate *ProgMod*. I discussed how my model produced results, more specifically prognosis prediction.

A. *Model Fitting with Test Data*

Using the test set, the gene expression levels of the 47 genes selected in the previous step were fitted in the Cox Model alongside the overall survival time and the survival status. The model was the used to predict the survival function the patients.

For the purpose of scoring the model the value from the survival function of each patient corresponding to the fixed probabilty 0f 0.8 was taken, and was considered the predicted survival time for that patient. Then the array of real values and predicted values were used to score and visualize the result of the model.

B. *Model Scoring*

Two metrics were used to score the model in each test set, both in sklearn's module: *metrics***:**

- *R2 Squared Accuracy*: used to predict how well the model will perform on future unseen data. Its formula is: $1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Yi - Y'i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Yi - Y''i)^2}$ where ***Yi*** is the real value, ***Y'i*** is the predicted value, and ***Y''i*** is the mean value.
- *Root Mean Squared Error*: used to calculate the average error per data point. Its formula is: $\sqrt{\frac{1}{n}\sum_{j=1}^{n}(yj - y'j)^2}$ where ***yj*** and ***y'j*** are the real and predicted values.

Finally, the lifeline implementation of the Cox Model has its own score. When data is fitted into the model, it provides a summery containing a score called the *concordance* score, and it show how well the data fit the model.

C. *Model Results & Evaluation*

For the test set GSE12417, we plotted the Real and predicted values to visualize which is shown in Figure 3.

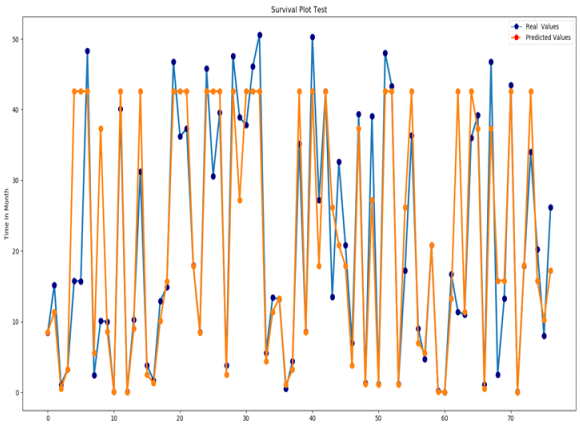

Figure 3 - Plot of Real and Predicted Values

Table I shows the model score by different metrics.

[4] https://lifelines.com

TABLE I Model Score

Metric	R2 Squared	RMSE
Score	0.75	9

We tested the Cox model. Table II shows the results of the Cox model.

TABLE II Summary of experiments with Cox Model

Gene	Coef	exp(coef)	se(coef)	Z	P	Lower_0.95	Upper_0.95
14107	-2.3919	0.0915	1.5373	-1.5559	0.1197	-5.405	0.6211
19804	0.1528	1.1651	0.533	0.2867	0.7743	-0.8918	1.1974
19804	0.1528	1.1651	0.533	0.2867	0.7743	-0.8918	1.1974
2663	0.9846	2.6768	0.6921	1.4226	0.1549	-0.372	2.3412
5250	-1.4543	0.2336	0.9365	-1.553	0.1204	-3.2898	0.3812
11694	0.3508	1.4202	0.6431	0.5454	0.5855	-0.9097	1.6113
9597	0.1556	1.1684	0.3848	0.4043	0.686	-0.5987	0.9099
18177	0.5393	1.7148	0.2627	2.0526	0.0401	0.0243	1.0543
18585	0.1553	1.168	0.4928	0.3151	0.7527	-0.8106	1.1211
6655	-0.6685	0.5125	0.7003	-0.9546	0.3398	-2.0411	0.7041
8213	-0.6822	0.5055	0.6484	-1.0523	0.2927	-1.953	0.5885
3048	-0.6301	0.5326	0.3583	-1.7583	0.0787	-1.3324	0.0722
1419	0.0886	1.0926	0.3315	0.2671	0.7894	-0.5612	0.7383
12264	0.733	2.0813	0.8757	0.837	0.4026	-0.9834	2.4494
19282	-3.3336	0.0357	1.058	-3.151	0.0016	-5.4072	-1.2601
11120	-0.3026	0.7389	0.6639	-0.4559	0.6485	-1.6038	0.9985
15456	0.4814	1.6183	0.2691	1.7889	0.0736	-0.046	1.0088
9945	0.559	1.749	0.5105	1.095	0.2735	-0.4416	1.5596
11893	1.2217	3.3931	0.4649	2.6281	0.0086	0.3106	2.1329
5604	2.0518	7.7817	0.5245	3.9115	0.0001	1.0237	3.0799
9279	-0.5672	0.5671	0.5421	-1.0462	0.2955	-1.6297	0.4954
19432	-1.3956	0.2477	0.3146	-4.4364	0	-2.0122	-0.7799
3910	1.9807	7.2479	0.798	2.482	0.0131	0.4166	3.5448
8240	-1.7741	0.1696	0.5566	-3.1871	0.0014	-2.8651	-0.6831
15024	1.7129	5.545	0.4103	4.1751	0	0.9088	2.517
8704	-1.6538	0.1913	0.4858	-3.4044	0.0007	-2.6059	-0.7017
2449	0.072	1.0747	1.022	0.0705	0.9438	-1.9311	2.0751
6773	-0.1585	0.8534	0.4584	-0.3457	0.7296	-1.057	0.74
11050	-1.1174	0.3271	0.4018	-2.781	0.0054	-1.9048	-0.3299
3675	0.7704	2.1606	0.3591	2.1451	0.0319	0.0665	1.4743
4025	1.475	4.3712	0.7743	1.9049	0.0568	-0.0426	2.9927
5510	-1.01	0.3642	0.6528	-1.547	0.1219	-2.2895	0.2696
4780	0.3181	1.3745	0.3647	0.872	0.3832	-0.3968	1.0329
1420	1.055	2.8721	0.3775	2.7946	0.0052	0.3151	1.795
59	0.1503	1.1621	0.3629	0.4141	0.6788	-0.561	0.8615
5521	1.9321	6.9042	0.5174	3.7345	0.0002	0.9181	2.9462
6772	0.9128	2.4913	0.366	2.494	0.0126	0.1955	1.6302
4181	1.5159	4.5537	0.5488	2.7624	0.0057	0.4404	2.5915
18134	-1.0503	0.3498	0.4779	-2.1977	0.028	-1.987	-0.1136
6552	0.315	1.3702	0.6254	0.5037	0.6145	-0.9107	1.5407
7596	1.0771	2.9362	0.308	3.4974	0.0005	0.4735	1.6807
18504	0.4695	1.5992	0.475	0.9885	0.3229	-0.4614	1.4005
3039	-0.2358	0.79	0.2144	-1.0997	0.2715	-0.6559	0.1844
445	-3.9446	0.0194	1.0507	-3.7541	0.0002	-6.004	-1.8852
6764	0.1693	1.1845	0.8073	0.2097	0.8339	-1.413	1.7516
4550	-1.0298	0.3571	0.3189	-3.2294	0.0012	-1.6549	-0.4048
17847	0.5091	1.6638	0.1573	3.2372	0.0012	0.2009	0.8173

3677	-0.5596	0.5714	0.3721	-1.504	0.1326	-1.2889	0.1697

The interpretation of the summary is the following:

- Statistical Significance: The column marked "z" gives the Wald statistic value. It corresponds to the ratio of each regression coefficient to its standard error (z = coef/se (coef)). The Wald statistic evaluates, whether the beta (β) coefficient of a given variable is statistically significantly different from 0.
- The Regression Coefficients: The second feature to note in the Cox model results is the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable.
- Hazard Ratios: The exponentiated coefficients (exp (coef) = exp (-0.53) = 0.59), also known as hazard ratios, give the exact size of covariates.
- Confidence Intervals of the Hazard Ratios. The summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp (coef)).

For Example, the genes 5604, 19432, 5521, 7596, and 445 have a p value < 0.001 and are the ones with the most statistical significance. These means that although we need all 47 genes to predict an accurate prognosis, these 5 genes are pretty good indicators of the prognosis of the patient.

VI. Conclusion and Future Works

In recent years, Bioinformatics is adopting mainstream technologies of computer science in performing various complex tasks. Machine learning – a branch of computer science- is one of the most popular fields of development of analytical models for performing analysis on critical, lifesaving subjects such as acute myeloid leukemia (AML). However, since there is an exhaustive number of algorithms and methods offered by ML, it is non-trivial task to find the best one without experimenting them. Furthermore, different models developed using different approaches produce different results with different accuracy levels.

We investigated different models proposed in literature for prognosis to the patients diagnosed with AML. According to our study, the results of these models vary which essentially imply that there is a possibility that the existing models could be improved in terms of accuracy.

In this paper, we presented a model to estimate the prognosis of AML patients using the Cox Proportional Hazard Model for survival analysis, using their Gene Expression Levels solely as the explanatory variables, thus overcoming the issue of high molecular heterogeneity of AML when it comes to prognosis prediction.

We have shown that it can estimate the survival time with r2_squared accuracy of 0.75 with the used dataset. And the model is valid to perform prediction on future patients. The number of genes that the model was constructed to work with is 47, a small number that with modern technologies can be tested with relatively small costs and time.

An extension of this work is lined up. That is, this approach can be extended to all sorts of cancer. However, the model has to be adjusted and generalized to select the significant genes, and perform the prognosis prediction of any cancer, based on the Gene Expression Levels of these selected genes.

References

[1] H. Döhner, J. D. Weisdorf and C. D. Bloom, "Acute Myeloid Leukemia," *The New England Journal of Medicine,* 2015.

[2] Golub, T. R. et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science, 286(5439), 531-537.,* 1999.

[3] M. L. Gulley, T. C. Shea and Y. Fedoriw, "Genetic tests to evaluate prognosis and predict therapeutic response in acute myeloid leukemia," *The Journal of Molecular Diagnostics,* 2010.

[4] Den Boer et al., "A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study," *The Lancet Oncology,* 2009.

[5] S. Mi, J. Lu, M. Sun, Z. Li, H. Zhang, M. B. Neilly and S. K. Bohlander, "MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia," *Proceedings of the National Academy of Sciences,* 2007.

[6] Ali Nehme et al., "Atlas of tissue renin-angiotensin-aldosterone system in human: A transcriptomic meta-analysis," *Scientific reports,* 2015.

[7] K. Wheatly et al., "Prognostic factor analysis of the survival of elderly patients with AML in the MRC AML11 and LRF AML14 trials," *British journal of haematology,* 2009.

[8] P. C. van Dijk , K. J. Jager, A. H. Zwinderman, C. Zoccali and F. W. Dekker, "The analysis of survival data in nephrology: basic concepts and methods of Cox regression," *abc of epidemiology,* 2008.

[9] S. H. Bouazza, K. Auhmani, A. Zeroual and N. Hamdi, "Selecting Significant Marker Genes From Microarray Data by Filter Approach for Cancer Diagnosis," *Procedia Computer Science 127,* 2018.

[10] Feiyo Xiong et al, "Kernelized Information-Theoretic Metric Learning for Cancer Diagnosis Using High-Dimensional Molecular Profiling Data," 2015.

[11] S. Tarek, M. AbdelWahab and M. Shouman, "Gene Expression Based Cancer Classification," *Egyption Informatics Journal,* 2017.

[12] K. R. Forester, R. Koprowski and J. D. Scufca, "Machine Learning, Medical Diagnosis, and Biomedical Engineering Research," *BioMedical Engineering OnLine,* 2014.