

Original Article

Running Title: Prediction Model for Successful Allogenic Hematopoietic Stem Cell Transplantation
Received: January 01, 2022; Accepted: April 26, 2022

Presenting a Prediction Model for Successful Allogenic Hematopoietic Stem Cell Transplantation in Adults with Acute Myeloid Leukemia

Mostafa Langarizadeh^{*}, PhD, Boshra Farajollahi^{*♦}, MSc, Abbas Hajifathali^{**}, MD

^{*}*Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran*

^{**}*Hematopoietic Stem Cell Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

♦Corresponding Author

Boshra Farajollahi, MSc

Department of Health Information Management,
School of Health Management and Information Sciences,
Iran University of Medical Sciences, Tehran, Iran

Tel: 02188794301

Email: boshraafrij@gmail.com

Abstract

Background: Allogenic hematopoietic stem cell transplantation is considered as an effective treatment for patients with acute myeloid leukemia. However, complications of transplantation, like aGVHD, affect the efficiency of allogenic hematopoietic stem cell transplantation.

The present study aimed to implement different models of data mining (DT) (single and ensemble) for prediction of allogenic hematopoietic stem cell transplantation in patients with acute myeloid leukemia (transplantation against host disease).

Method: We conducted this developmental study on 94 patients with 34 attributes in Taleghani Hospital, Tehran, Iran, during 2009–2017. In this practical study, the data were analyzed via DT algorithms, including decision tree, random forest and gradient boosting (ensemble learning), artificial neural network (Single Learning), and support vector machine. Some criteria, like specificity, accuracy, F-measure, AUC (area under curve), and sensitivity, were reported in order to evaluate DT algorithms.

Results: There were 34 transplantation-related variables; some predictors, such as liver, hemoglobin, and donor blood group, were found to be the most important ones. To predict aGVHD, the two selected algorithms included the most appropriate DM models, ANN and SVM classifiers, with ROC of 100.

Conclusion: This study indicated that DT algorithms could be successfully used for approving the efficiency of the models predicting allogenic hematopoietic stem cell transplantation.

Keywords: Data mining, Allogenic hematopoietic stem cell transplantation, Leukemia, Myeloid, Acute

Introduction

Blood cancers are those which attack bone marrow, blood, and lymphatic system.^{1, 2} Among different kinds of blood cancers, acute myeloid leukemia (AML) cancers, which are malignant disorder in bone marrow, grow in myeloma or lymphocyte cells. They are the most dangerous blood malignant with a high and quick progression.^{3, 4} Despite a considerable progress in the diagnosis of AML and improvement in options treatment, AML is still a deadly disease with high mortality rate.^{5, 6} Chemotherapy and Allo-HSCT could be mentioned as effective treatment strategies for this disease.^{7, 8} Although numerous patients obtain complete recovery after chemotherapy, their disease may return stronger and more offensive; therefore, improvement in treatment strategies is needed.⁹ However, Allo-HSCT is a good way to treat AML patients, even those with a high risk of disease return, but there are some challenges associated with its application, such as acute/chronic GVHD, life-threatening infections, and relapse of the disease.^{5, 8, 10-13}

Complicated and different factors prior to transplant surgery affect patients' condition after the transplantation. However, due to the complications related to the transplantation of Allogenic Hematopoietic Stem Cells, prediction of patients' condition after the surgery based on experimental combination of special factors is not always purposive.¹³ In fact, since there are side effects and high range of death due to transplantation, it making decision about the time, the way, and the patient who should be transplanted is of great importance.¹⁴ In medicine, understanding and treating blood diseases could be challenging. It is necessary that oncologists find out the most important variables and factors for predicting the treatment since a wrong decision in

treatment, in addition to wasting time and money, could lead to death.^{15, 16}

Thus, choosing an appropriate way to treat the patient is of particular necessity; updated information and related causes to death are prerequisite for improvement of outcomes.¹⁷ Since deciding whether to perform Allo-HSCT is challenging both for physicians and patients, powerful tools are needed to help physicians in terms of recognizing patients for Allo-HSCT.¹⁸

Health and treatment are important sections in most countries. But some challenges, such as increased costs and complications, would be coming up. Better decision making based on the existing data could minimize those challenges. Examining different effective factors for overcoming these challenges, several problems have been faced, for example, differentiation in the final result of research projects due to absence of standardization and limitation of determining of many patients' outcomes.

Different analyses could provide tools and techniques, thereby further information from mass and complicated data and introduce some useful information to help decision making in health care.¹⁹ Solutions based on analyses, such as data mining (DT), make it possible to overcome the above-mentioned challenges.²⁰ Therefore, predicted models for Allo-HSCT have been provided in order to improve decision making and help physicians optimize care strategies and categorize risks in long and provident analyses. Moreover, new revolutions in DT algorithms have provided better understanding of patients' biology, and improvement in prognosis and anticipated tools could be conducive to making clinical decisions; yet, they are not completely applied in clinical situations.^{15, 16} aGVHD is a complicated disease, including 10 multi-systematic diseases, and is one of the important reasons behind nonrecurring death and low-quality life in patients with long lifetime. Accordingly, in the first 100

days following the transplantation, due to the donor T cells attack to the host tissues, certain disorders occur in skin, digestion system, and liver functionality in 35 to 50% of patients under Allo-HSCT transplant.²¹ Hence, the main objective of this study was predicting the absence of aGVHD using different DT algorithms.

Materials and Methods

In this developmental study, to predict aGVHD, the data were obtained from 94 AML patients who received Allo-HSCT in Taleghani hospital (Tehran-Iran) between 2009 to 2017. A total of 34 features (including base-line and biomarker data) were used, which are believed to be effective in transplantation results. Prediction was done using 5 DT algorithms applied in MatLab software. At the first phase, as shown in Figure 1, preprocessing was performed. The records and features with missing values (more than 50% of cases) were removed from the dataset. In addition, the missing values of continuous and discrete variables were replaced separately, with mean and mode in each class, respectively.

For feature selection, genetic algorithm with wrapper method was utilized. The last process was the evaluation and comparison of the algorithms efficiency. The performance of the algorithms was assessed using accuracy, sensitivity, specificity, F-measure, and AUC (area under curve) criteria. Finally, the algorithm (or algorithms) with the greatest area under the curve of ROC were suggested.

This essay was extracted from a thesis with the morality certificate from the ethics national committee of biological studies of Iran University of Medical Sciences (IR.IUMS.REC.1399.1031).

Results

Features

Table 1 represents the dataset variables and their descriptions.

Pre-processing

After discarding incomplete patient records, the patient dataset was reduced to 93 patients (65 case-patients diagnosed with aGVHD vs. 28 controls who did not experience aGVHD). The implementation process was once done with the 34 obtained features from the preprocessing step without and with feature selection. The results of feature selection showed that out of the 34 features included in transplanting, the most important ones are the donor blood group, liver, and hemoglobin.

Tables 2 and 3 as well as figure 2 depict the evaluation of DT algorithms with and without feature selection in this step.

Based on the evaluation criteria, including accuracy, sensitivity, specificity, F-measure, and AUC, the ANN and SVM had the best performance.

Discussion

Comparing the performance of models with and without feature selection

In this study, five algorithms were implemented to predict aGVHD with and without feature selection. Performance of the algorithms with feature selection in GB, SVM, and ANN algorithms increased considerably; in random forest implementation, algorithms with and without feature selection had the same performance and DT algorithm performance with feature selection in comparison with the implementation without feature selection saw a slight decrease.

Selected features

In the present work, three features were identified as the most important factors associated with aGVHD on transplantation day. The features included donor blood group, liver, and hemoglobin as based line

factors showed that biomarkers are important for oncologists.

Background

In 2018, Aria et al. conducted a study entitled “predicting aGVHD following Allo-HSCT using an ML algorithm”, which they published about patients with malignant and benign blood diseases. The mentioned study aimed to develop ML models to predict grades 2 to 4 of aGVHD accurately. They draw a comparison among the 5 algorithms of machine learning (Naïve Bayes, Alternative decision tree, Multilayer Neural Network, random forest, and Adaboost). They suggested ADTree algorithm and reported 62/3% AUC for grades 3-4.²²

In 2018, Lee et al. conducted a study on patients with malignant hematology (Acute Myeloid Leukemia, Acute Lymphocyte Leukemia, Mito-dysplastic Syndrome, and Chronic Myeloid Leukemia). To make the predicting tools, Super Learning (a group of learning devices developed recently. It is a combination of the results of a number for algorithms) was used for a single optimum prediction. The reported AUC was in the range of 61.3 to 64% for these ensemble models.²³

In 2019, Fuse et al. published a paper using the data of lymphocyte and AML patients. Through the use of ADTree algorithm, prediction of Acute Leukemia blood cancer turn was analyzed after one year of transplantation. With the model of training data (n = 148), prediction accuracy, the AUC of ROC, and k-static value were 78/4%, 0/746, and 0/508, respectively.¹²

In 2019, Okamara et al. published a study entitled “Interactive web application for plotting personalize prognosis prediction curves in allogenic hematopoietic cell transplantation using machine learning” employing the data of patients with malignant disorder blood disease. To develop the application of random survival forest model, eight features were used. The AUC

for 1-year overall survival, progression-free survival, relapse progression, and non-relapse mortality in test cohort were 0.70, 0.72, 0.73, and 0.77, respectively.²⁴

In 2020, Tang et al., using the data of patients with malignant hematology disease, reported 65/9% AUC for grades 2-4 aGVHD.²⁵

In 2020, SalehNasab et al. published an essay entitled “An intelligent clinical decision support system for predicting acute graft-versus host disease (aGVHD) following allogenic hematopoietic stem cell transplantation” using the data of patients with malignant hematology disease. In their work, after implementation and evaluation of different algorithms based on the average of different evaluation ways (Accuracy, F-measure), two algorithms, namely HistGradient Boosting and XGB, with the mean of 90.53 and 90.82, respectively, were selected as the most functional algorithms; therefore, CDSS was implemented.²⁶

Using the AML patient under Allo-HSCT in this study, more professional factors associated with this disease and the type of treatment is identified. This research was performed on AML patients under Allo-HSCT in Taleghani Hospital; using the data of the two bone marrow transplantation centers affiliated to Tehran and Iran universities was not possible.

Conclusion

Comparing the performance of algorithms with and without feature selection suggested better effect of using feature selection. The selected features included donor blood group, liver, and hemoglobin. Furthermore, two algorithms, namely SVM and ANN, had the best performance in predicting aGVHD in patients with AML who underwent the transplantation of Allo-HSCT as one of the most important factors in transplantation. The high performance of the suggested algorithms, unlike previous studies, revealed that independent analysis of Allo-HSCT in

patients with AML was effective. It could be also suggested that predicting recurrence in AML patients under Allo-HSct transplantation should be considered as one of the factors in Allo-HSct success in future studies.

Acknowledgement

This study was part of the MSc. project conducted at Iran University of Medical Sciences, Tehran, Iran (Grant Number: 99-3-37-19600).

Conflict of Interest

None declared.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(1):7-34. doi: 10.3322/caac.21551.
2. Dhillon A, Singh A. Machine learning in healthcare data analysis: A survey. *J Biol Today's World*. 2018;8(2):1-10. doi: 10.15412/J.JBTW.01070206.
3. Chu Y, Chen F, Sheng Z, Zhang D, Zhang S, Wang W, et al. Blood cancer diagnosis using ensemble learning based on a random subspace method in laser-induced breakdown spectroscopy. *Biomed Opt Express*. 2020;11(8):4191-202. doi: 10.1364/BOE.395332.
4. Mousavinasab SN, Yazdani Cherati J, Karami H, Khaksar S. Risk Factors influencing the survival of pediatric acute leukemia using competing risk model. [In Persian] *J Mazandaran Univ Med Sci*. 2015;24(121):31-8.
5. Bittencourt MC, Ciurea SO. Recent advances in allogeneic hematopoietic stem cell transplantation for acute myeloid leukemia. *Biol Blood Marrow Transplant*. 2020;26(9):e215-e21. doi: 10.1016/j.bbmt.2020.06.007.
6. England JT, Saini L, Hogge D, Forrest D, Narayanan S, Power M, et al. Day 14 Bone marrow evaluation during acute myeloid leukemia induction in a real-world Canadian cohort. *Clin Lymphoma Myeloma Leuk*. 2020;20(7):e427-e36. doi: 10.1016/j.cmlm.2020.02.012.
7. Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood Rev*. 2019;36:70-87. doi: 10.1016/j.blre.2019.04.005.
8. Doppelhammer M, Fraccaroli A, Prevalsek D, Bücklein V, Häbe S, Schulz C, et al. Comparable outcome after haploidentical and HLA-matched allogeneic stem cell transplantation for high-risk acute myeloid leukemia following sequential conditioning—a matched pair analysis. *Ann Hematol*. 2019;98(3):753-62. doi: 10.1007/s00277-019-03593-2.
9. Wang W, Stiehl T, Raffel S, Hoang VT, Hoffmann I, Poisa-Beiro L, et al. Reduced hematopoietic stem cell frequency predicts outcome in acute myeloid leukemia. *Haematologica*. 2017;102(9):1567. doi: 10.3324/haematol.2016.163584.
10. Shouval R, Bonifazi F, Fein J, Boschini C, Oldani E, Labopin M, et al. Validation of the acute leukemia-EBMT score for prediction of mortality following allogeneic stem cell transplantation in a multi-center GITMO cohort. *Am J Hematol*. 2017;92(5):429-34. doi: 10.1002/ajh.24677.
11. Muhsen IN, Jagasia M, Toor AA, Hashmi SK. Registries and artificial intelligence: Investing in the future of hematopoietic cell transplantation. *Bone Marrow Transplant*. 2019;54(3):477-80. doi: 10.1038/s41409-018-0327-x.
12. Fuse K, Uemura S, Tamura S, Suwabe T, Katagiri T, Tanaka T, et al. Patient-based prediction algorithm of relapse after allo-HSCT for acute Leukemia and its usefulness in the decision-making process using a machine learning approach. *Cancer*

- Med.* 2019;8(11):5058-67. doi: 10.1002/cam4.2401.
13. Aytan P, Yeral M, Korur A, Gereklioglu C, Kasar M, Buyukkurt NH, et al. Factors associated with overall survival in acute myeloid leukemia patients before and after hematopoietic stem cell transplant. *Exp Clin Transplant.* 2021;19(8):856-64. doi: 10.6002/ect.2018.0352.
14. Shouval R, Labopin M, Unger R, Giebel S, Ciceri F, Schmid C, et al. Prediction of hematopoietic stem cell transplantation related mortality-lessons learned from the in-silico approach: A European society for blood and marrow transplantation Acute Leukemia working party data mining study. *PLoS One.* 2016;11(3):e0150637. doi: 10.1371/journal.pone.0150637.
15. Radakovich N, Cortese M, Nazha A. Acute myeloid leukemia and artificial intelligence, algorithms and new scores. *Best Pract Res Clin Haematol.* 2020;33(3):101192. doi: 10.1016/j.beha.2020.101192.
16. Kashef A, Khatibi T, Mehrvar A. Treatment outcome classification of pediatric Acute Lymphoblastic Leukemia patients with clinical and medical data using machine learning: A case study at MAHAK hospital. *Inform Med Unlocked.* 2020;20:100399. doi: 10.1016/j.imu.2020.100399.
17. Ghorbani R, Ghousi R. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science.* 2019;3(2):47-70. doi: 10.5267/j.ijdns.2019.1.003.
18. Elsayy M, Sorrow M. Up-to-date tools for risk assessment before allogeneic hematopoietic cell transplantation. *Bone Marrow Transplant.* 2016;51(10):1283-300. doi: 10.1038/bmt.2016.141.
19. Islam MS, Hasan MM, Wang X, Germack HD, Noor-E-Alam M. A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare (Basel).* 2018;6(2):54. doi: 10.3390/healthcare6020054.
20. Potdar R, Varadi G, Fein J, Labopin M, Nagler A, Shouval R. Prognostic scoring systems in allogeneic hematopoietic stem cell transplantation: Where do we stand? *Biol Blood Marrow Transplant.* 2017;23(11):1839-46. doi: 10.1016/j.bbmt.2017.07.028.
21. Salehnasab C, Hajifathali A, Asadi F, Parkhideh S, Kazemi A, Roshanpoor A, et al. An Intelligent Clinical Decision Support System for Predicting Acute Graft-versus-host Disease (aGvHD) following Allogeneic Hematopoietic Stem Cell Transplantation. *J Biomed Phys Eng.* 2021;11(3):345-56. doi: 10.31661/jbpe.v0i0.2012-1244.
22. Arai Y, Kondo T, Fuse K, Shibasaki Y, Masuko M, Sugita J, et al. Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Adv.* 2019;3(22):3626-34. doi:10.1182/bloodadvances.2019000934.
23. Lee C, Haneuse S, Wang H-L, Rose S, Spellman SR, Verneris M, et al. Prediction of absolute risk of acute graft-versus-host disease following hematopoietic cell transplantation. *PLoS One.* 2018;13(1):e0190610. doi: 10.1371/journal.pone.0190610.
24. Okamura H, Nakamae M, Koh S, Nanno S, Nakashima Y, Koh H, et al. Interactive web application for plotting personalized prognosis prediction curves in allogeneic hematopoietic cell transplantation using machine learning. *Transplantation.* 2021;105(5):1090-6. doi: 10.1097/TP.0000000000003357.
25. Tang S, Chappell GT, Mazzoli A, Tewari M, Choi SW, Wiens J. Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO Clin Cancer Inform.* 2020;4:128-35. doi: 10.1200/CCI.19.00105.

26. Salehnasab C, Hajifathali A, Asadi F, Roshandel E, Kazemi A, Roshanpoor A. Machine learning classification algorithms to predict aGvHD following Allo-HSCT: A

systematic review. *Methods Inf Med.* 2019;58(6):205-212. doi: 10.1055/s-0040-1709150.

Table 1. The dataset features and their descriptions

Type	Row	Feature	Description	Role
Base-line	1	Patient's Gender		Input
	2	Donor's Gender		Input
	3	Donor's and Patient's Gender		Input
	4	Patient's Blood group		Input
	5	Donor's Blood group		Input
	6	Blood group compatibility	Donor and the recipient have the same blood group antigens and plasma antibodies	Input
	7	Marital Status		Input
	8	Patient's Age		Input
	9	Donor's Age		Input
	10	Patient's BMI		Input
	11	Donor's BMI		Input
	12	Smoking		Input
	13	Donor-Recipient Relationship	The relationship between the donor and patient, including related and siblings	Input
	14	Delivery	The process of giving birth for the donor	Input
	15	Profilaxy Regimen	Regimen use for the prevention of a specific disease	Input
	16	Chemotherapy Regimen	Regimen 1-3: Myeloablative is an intensive conditioning regimen to destroy the bone marrow cells	Input
	17	Complete Remission	Including tests, physical exams, and scans showing that all signs of cancer are gone	Input
	18	Radiothrapy Pre-BMT	The treatment of disease with ionizing radiation	Input
	19	MNC		Input

Biomarker	20	Time from diagnosis to transplantation	The time between the disease diagnosis and hematopoietic stem cell transplantation	Input
	21	Creatinine		Input
	22	LDH		Input
	23	Uric Acid		Input
	24	CRP		Input
	25	Platelet		Input
	26	WBC		Input
	27	Hemoglobin		Input
	28	Liver		Input
	29	Intestine		Input
	30	Pulmonary		Input
	31	Skin		Input
	32	CMV		Input
	33	Platelet single donor transfusion number 0 to 10 days		Input
	34	Platelet single donor transfusion number 0 clearance days		Input
35	aGVHD		Target	

BMI: Body mass index; BMT: Bone marrow transplantation; MNC: Mononuclear cell; LDH: Lactat dehydrogenase; CRP: C-reactive protein; WBC: White blood cells; aGVHD: Acute graft-versus-host disease

Table 2. Results of evaluation of the algorithms process without feature selection

Row	Classifier	Accuracy	Sensitivity	Specificity	F-measure	AUC
1	ANN	96.29	100	94.44	94.73	97.22
2	SVM	62.96	NaN	62.96	0	62.96
3	DT	92.59	90	94.11	90	92.05
4	GB	62.96	NaN	62.96	94.73	78.84
5	RF	62.29	100	94.44	94.73	97.22

ANN: Artificial neural network; SVM: Support vector machine; DT: Decision tree; GB: Gradient boosting; RF: Random forest; AUC: Area under the curve

Table 3. Results of evaluation of the algorithms process with feature selection

Row	Classifier	Accuracy	Sensitivity	Specificity	F-measure	AUC
1	ANN	100	100	100	100	100
2	SVM	100	100	100	100	100
3	DT	92.59	83.33	100	90.90	91.66
4	GB	96.29	90.90	100	95.23	95.45
5	RF	96.29	100	94.44	94.73	97.22

ANN: Artificial neural network; SVM: Support vector machine; DT: Decision tree; GB: Gradient boosting; RF: Random forest; AUC: Area under the curve

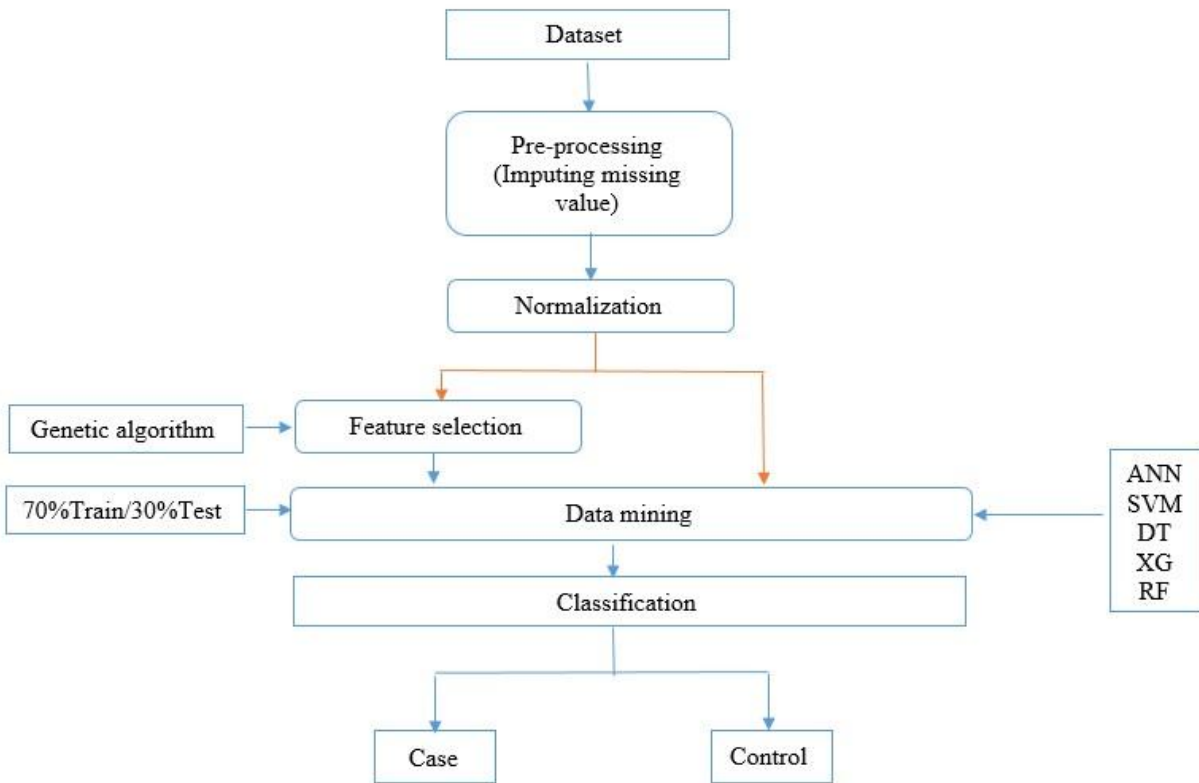


Figure 1. This figure shows an overview of the proposed data mining algorithms.

ANN: Artificial neural network; SVM: Support vector machine; DT: Decision tree; GB: Gradient boosting; RF: Random forest

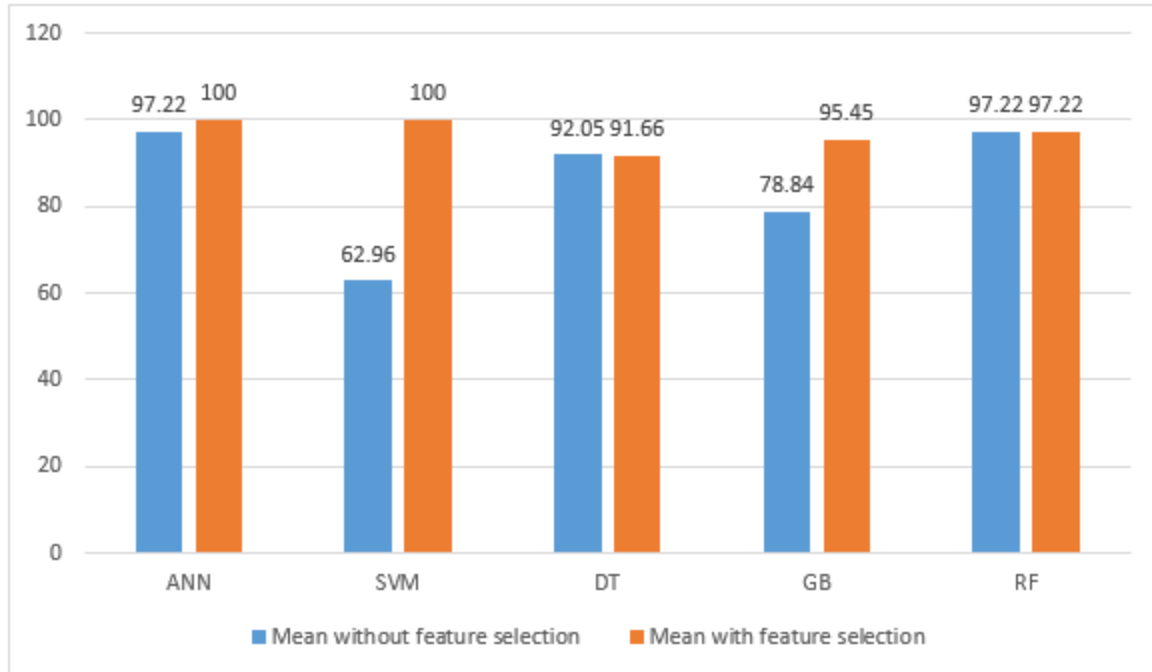


Figure 2. This figure shows the comparison surface under the ROC of models with and without feature selection.

ANN: Artificial neural network; SVM: Support vector machine; DT: Decision tree; GB: Gradient boosting; RF: Random forest