Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 364-369, Nisan 2020
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 364-369, April 2020
Copyright © 2020 EJOSAT
Research Article

Suicide Prediction from Hemogram with Machine Learning**

Berna Arı[1**], Ali Arı[2], Abdulkadir Şengür[3]

[1*] Fırat Üniversitesi, Teknoloji Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Elazığ, (ORCID: 0000-0003-1000-2619)
[2] İnönü Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Malatya, Türkiye (ORCID: 0000-0002-5071-6790)
[3] Fırat Üniversitesi, Teknoloji Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Elâzığ, (ORCID: 0000-0003-1614-2639)

(Conference Date: 5-7 March 2020)

(**DOI:** 10.31590/ejosat.araconf47)

ATIF/REFERENCE: Ari, B., Ari, A., & Sengur, A. (2020). Suicide Prediction from Hemogram with Machine Learning. *Avrupa Bilim ve Teknoloji Dergisi*, (Special Issue), 364-369.

Abstract

Suicide; It is a phenomenon that we encounter with different frequencies and methods by hosting social, economic and cultural factors at its base. Adolescence, which is an upper step of childhood, contains complex emotions such as hopelessness, loneliness, and depression in its world, and it is a stage in which the risk of suicide is high. It is of great importance to take necessary measures in neutral and imperceptible ways in terms of adolescence and suicide relationship. Blood, which can be easily taken by experts even in a non-severe illness, appears as numerical data with the parametric values that make up its content in laboratories. The hemogram test showing the measurement of blood parameters is used in the diagnosis of many diseases today. In this study, the relationship between the values obtained as a result of the hemogram test and the possibility of suicide of adolescent individuals were investigated. Leukocyte (WBC), erythrocyte (RBC), basophil (BA), eosinophil (EO), lymphocyte (LY), Monocyte (MO), Neutrophil (NE) and Platelet (PLT) of adolescents who have attempted suicide and whose age and gender are known, blood values of mean platelet volume (MPV) and hemoglobin (HGB) levels were evaluated within the designed system. Complete blood count data of 302 individuals who were healthy and suicidal attempts constituting the dataset were pre-processed and the data that would adversely affect the estimated suicide group were removed from the system by considering the references. While making suicide estimation, the high performance bagging trees and the Support Vector Machines separating the members of the two groups with high accuracy were chosen as a result of the joint study of the classification algorithms. It has been shown that by using 260x13 attribute, the classification results can be obtained with BT and Quadratic SVM and 93.5% accurate predictions can be made with BT. Experts will be able to easily find out how high or at which level this probability is, if the individual has any psychological disorders or if the suicide is suspected.

Keywords: Bagged Tree, suicide prediction, hemogram, support vector machines, complete blood count, machine learning

Makine Öğrenmesi ile Tam Kan Sayımı Değerlerinden İntihar Tahmini

Öz

İntihar; toplumsal, ekomonik ve kültürel etkenleri tabanında barındararak farklı sıklıklar ve yöntemlerle karşımıza çıkan bir olgudur. Çocukluğun bir üst basamağı olan ergenlik ise ümitsizlik, yalnızlık, depresyon gibi karmaşık duyguları dünyasında barındırmakta olup intihar riskinin yüksek oranda görüldüğü bir evredir. Gerekli tedbirlerin tarafsız ve hissedilmez yollarla alınabilmesi ergenlik ve intihar ilişkisi açısından büyük önem arz etmektedir. Ağır seyretmeyen bir rahatsızlıkta dahi uzmanlar tarafından kolayca alınabilen kan, sonrasında laboratuvarlarda içeriğini oluşturan parametrik değerler ile sayısal veriler olarak karşımıza çıkmaktadır. Kan

[*] This paper was presented at the ***International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF 2020)***.

[**] Corresponding Author: Fırat Üniversitesi, Teknoloji Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Elazığ

parametrelerinin ölçümünü gösteren hemogram testi, günümüzde birçok hastalığın tanısında kullanılmaktadır. Bu çalışmada hemogram testi sonucunda elde edilen değerler ile ergen bireylerin intihar etme olasılığı arasındaki ilişki araştırılmıştır. Cinsiyeti ve yaşı bilinen, intihar girişiminde bulunmuş ergenler ile sağlıklı ergenlerin lökosit (WBC), eritrosit (RBC), bazofil (BA), eozinofil (EO), lenfosit (LY), Monosit (MO), Nötrofil (NE) ve Platelet (PLT) sayısı, Ortalama trombosit hacmi (MPV) ve Hemoglobin (HGB) düzeylerine ait kan değerleri tasarlanan sistem içerisinde değerlendirilmiştir. Veri setini oluşturan sağlıklı ve intihar girişiminde bulunmuş 302 kişiye ait tam kan sayımı verileri ön işlemden geçirilerek intihar eden gruba ait tahmini olumsuz etkileyecek veriler referns ralıklar göz önünde bulundurularak sistemden çıkarılmıştır. İntihar tahminini yaparken, sınıflama algoritmalarının ortak çalışması sonucu yüksek başarım gösteren torbalama ağaçları ve iki grubun üyelerini yüksek doğrulukla ayıran Destek Vektör Makineleri (DVM) seçilmiştir. 260x13 öznitelik kullanılarak Topluluk Ağaçları (TA) ve Kuadratik DVM ile sınıflandırma sonuçları alınarak TA ile %93,5 doğru tahminde bulunulabileceği gösterilmiştir. Uzmanlar, bireyin herhangi bir psikolojik rahatsızlığı görülmesi ya da intihar şüphesi olması durumunda, bu ihtimalin ne kadar yüksek ya da ne düzeyde olduğunu tasarlanan sistem sayesinde kolayca öğrenebilecektir.

Anahtar Kelimeler: Torbalama Ağacı, İntihar tahmini, Hemogram, Destek vektör makineleri, Tam kan sayımı, Makine öğrenmesi

1. Introduction

Hemogram (complete blood count) tests form the basis for health status analysis, disease diagnosis and effective treatment in individuals. Hemogram Test (CBC); is a laboratory test that provides specific data for the diagnosis of many diseases, especially anemia, infection, inflammation and bleeding disorders (Aslan,2019). As a result of this test; The hematological picture, created by counting and percentage of blood components, provides experts with information about the physiological and systemic state of the body (Aslan,2019).

In the literature, CBC tests have been used together with artificial intelligence to diagnose many diseases recently. In a study examining the effect of biochemistry values in iron deficiency anemia, Artificial Neural Networks (ANN) and k- Nearest Neighbor (k-EYK) were used. In the designed system, hemoglobin (HGB), hematocrit (Hct), mean corpuscular volume, REM data obtained from 50 patients and 50 healthy individuals were assessed using the RBC data and the highest performance was found by ANN (78.31%). (İlaslaner and Güven, 2019). Another study analyzing CBC samples made evaluations on five disease predictions with leukemia, inflammatory disease, bacterial or viral infection, HIV infection and anemia with the Support Vector Machines (SVM) (Deepa et al., 2019). Another study on the use of hemogram data from childhood psychiatric patients as an inflammatory marker was tested using artificial intelligence architectures. The accuracy of the system, which was evaluated with the ANN model with Multi-Layer Sensor Network Architecture, was tested with three-layer cross-validation and was calculated as 99% (Ucuz et al., 2019). In another study, the classification performance for the field of immunology, where it was difficult to examine the immune cells without label, was obtained using hemogram data, and with ANN 91.3% classification accuracy was obtained (Gupta et al., 2019). In the study that estimated the chronological age of the individual with CBC data, Deep Neural Networks were used and 81.5% accuracy was achieved within the framework of the ten-year age (Putin et al.,2016). Another study describing the diseases of leukemia, inflammatory, bacterial or viral infection, HIV infection and pernicious anemia with CBC values has developed the weighted k-mean algorithm and evaluated the efficiency of the study with the C-mean and K-average clustering algorithms, and the performance of the weighted k-average algorithm from other algorithms reported that it was better. When related studies are analyzed in a more specific dimension, it is seen that the relationship between psychiatric disorders and CBC test is one of the most frequently researched topics that maintains its popularity. In another study investigating the neutrophil / lymphocyte (NLR) ratio of individuals who attempted suicide, the results were compared with the control group and the increase in NLR value was emphasized in people who attempted suicide (Ayhan et al., 2019).

In this study, it is aimed to estimate the possibility of suicide attempt by looking at individuals' CBC data. In patients with psychiatric problems or suspicions, experts will be assisted in assessing the possibility of suicide by entering the CBC data into the system. CBC values obtained from volunteers in the same age group who attempted suicide and who did not have any physical and mental illness were examined. Data belonging to 260 patients were used by eliminating the data of the individuals who attempted suicide outside the reference value range of the relevant test and those that are thought to limit the forecast. The attribute set was given to the SVM and Ensemble Method (EM) classifiers, and it was concluded that 93.5% accurate predictions can be made with EM.

2. Material and Method

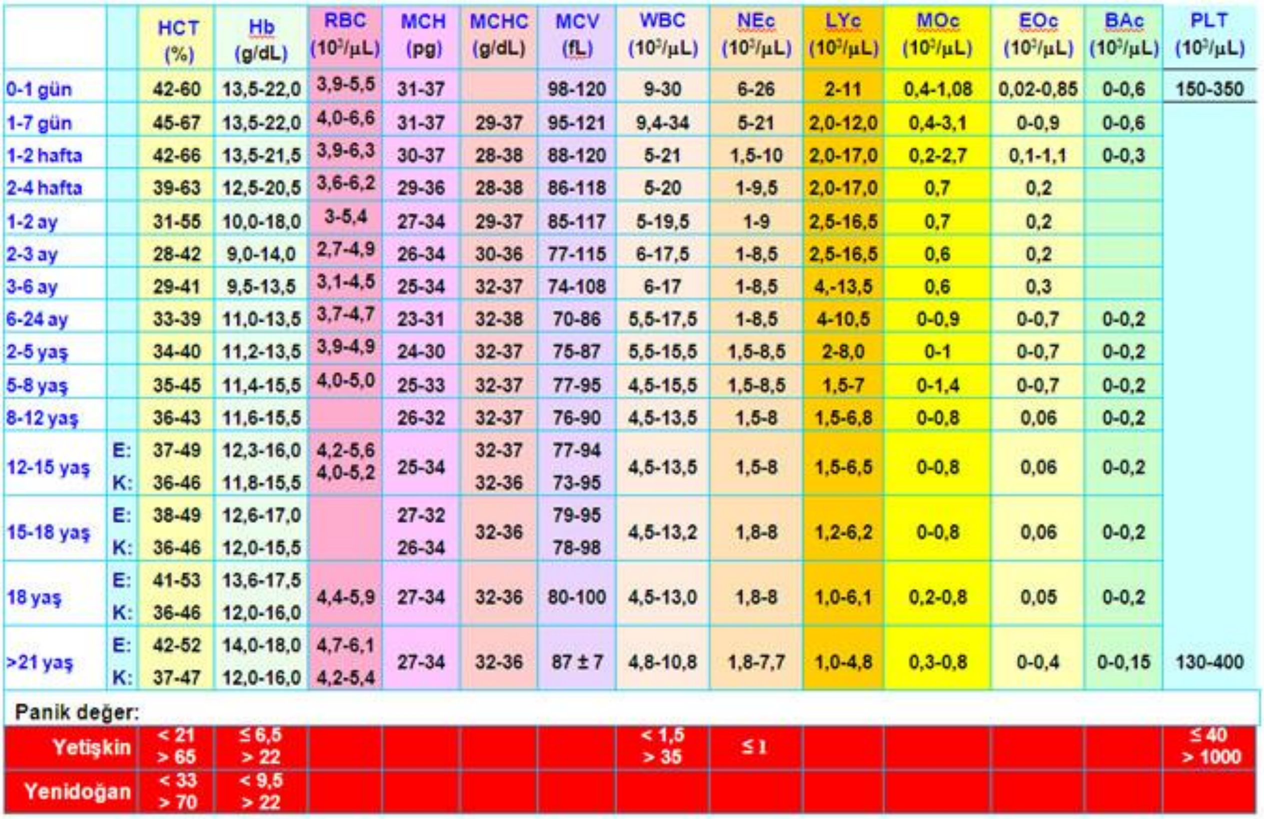
2.1. Data Acquisition and Feature Extraction

In the designed system, CBC test data obtained from 193 patients, 34 males and 159 females aged between 12-18 years, who were followed up in the pediatric psychiatry clinic of Malatya Turgut Özal Medical Center between 2017 and 2020, were taken only at the level of gender and age without sharing personal information. In the control group, CBC data of 41 male and 68 female volunteers from the same age group without physical and mental history were included in the system. CBC values obtained from the blood taken from adolescents who attempted suicide by any method, the majority of which were 'medicines', within 6 hours after drug intake were included in the study. The results of 42 individuals with abnormal (well below or much higher than the relevant value) results in the hemogram reference range and values that would adversely affect the forecast were not included in the study. Considering gender, age and blood values, 260x13 attributes obtained were evaluated from the blood values for the purpose of suicide prediction. The values considered in the study are shown in Table 1 and reference ranges for blood values are shown in Table 2 (Cayci et al.,2015).

Table 1. Attributes of the designed system

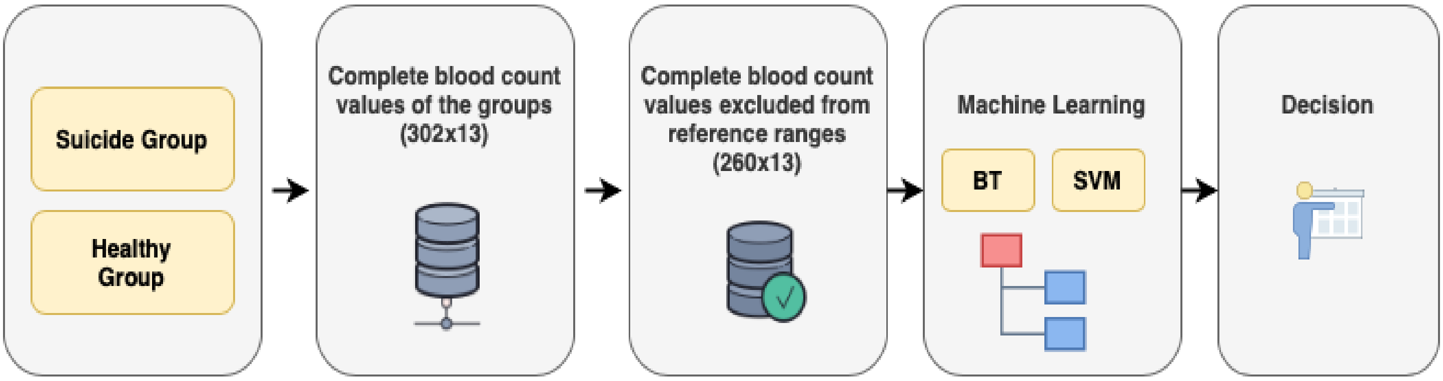
Features
Gender
Age range
WBC (White Blood Cell)
RBC (Red Blood Cell)
HGB (Hemoglobin)
BA (Basophil)
EO (Eosinophill Sayısı)
LY (Lenfosit)
MO (Monosit)
NE (Neutrophil)
MPV (Mean Platelet Volume)
PLT (Platelets)

Table 2. Reference ranges for blood values



		HCT (%)	Hb (g/dL)	RBC ($10^3/\mu L$)	MCH (pg)	MCHC (g/dL)	MCV (fL)	WBC ($10^3/\mu L$)	NEc ($10^3/\mu L$)	LYc ($10^3/\mu L$)	MOc ($10^3/\mu L$)	EOc ($10^3/\mu L$)	BAc ($10^3/\mu L$)	PLT ($10^3/\mu L$)
0-1 gün		42-60	13,5-22,0	3,9-5,5	31-37		98-120	9-30	6-26	2-11	0,4-1,08	0,02-0,85	0-0,6	150-350
1-7 gün		45-67	13,5-22,0	4,0-6,6	31-37	29-37	95-121	9,4-34	5-21	2,0-12,0	0,4-3,1	0-0,9	0-0,6	
1-2 hafta		42-66	13,5-21,5	3,9-6,3	30-37	28-38	88-120	5-21	1,5-10	2,0-17,0	0,2-2,7	0,1-1,1	0-0,3	
2-4 hafta		39-63	12,5-20,5	3,6-6,2	29-36	28-38	86-118	5-20	1-9,5	2,0-17,0	0,7	0,2		
1-2 ay		31-55	10,0-18,0	3-5,4	27-34	29-37	85-117	5-19,5	1-9	2,5-16,5	0,7	0,2		
2-3 ay		28-42	9,0-14,0	2,7-4,9	26-34	30-36	77-115	6-17,5	1-8,5	2,5-16,5	0,6	0,2		
3-6 ay		29-41	9,5-13,5	3,1-4,5	25-34	32-37	74-108	6-17	1-8,5	4,-13,5	0,6	0,3		
6-24 ay		33-39	11,0-13,5	3,7-4,7	23-31	32-38	70-86	5,5-17,5	1-8,5	4-10,5	0-0,9	0-0,7	0-0,2	
2-5 yaş		34-40	11,2-13,5	3,9-4,9	24-30	32-37	75-87	5,5-15,5	1,5-8,5	2-8,0	0-1	0-0,7	0-0,2	
5-8 yaş		35-45	11,4-15,5	4,0-5,0	25-33	32-37	77-95	4,5-15,5	1,5-8,5	1,5-7	0-1,4	0-0,7	0-0,2	
8-12 yaş		36-43	11,6-15,5		26-32	32-37	76-90	4,5-13,5	1,5-8	1,5-6,8	0-0,8	0,06	0-0,2	
12-15 yaş	E:	37-49	12,3-16,0	4,2-5,6	25-34	32-37	77-94	4,5-13,5	1,5-8	1,5-6,5	0-0,8	0,06	0-0,2	
	K:	36-46	11,8-15,5	4,0-5,2		32-36	73-95							
15-18 yaş	E:	38-49	12,6-17,0		27-32	32-36	79-95	4,5-13,2	1,8-8	1,2-6,2	0-0,8	0,06	0-0,2	
	K:	36-46	12,0-15,5		26-34		78-98							
18 yaş	E:	41-53	13,6-17,5	4,4-5,9	27-34	32-36	80-100	4,5-13,0	1,8-8	1,0-6,1	0,2-0,8	0,05	0-0,2	
	K:	36-46	12,0-16,0											
>21 yaş	E:	42-52	14,0-18,0	4,7-6,1	27-34	32-36	87 ± 7	4,8-10,8	1,8-7,7	1,0-4,8	0,3-0,8	0-0,4	0-0,15	130-400
	K:	37-47	12,0-16,0	4,2-5,4										
Panik değer:														
Yetişkin		< 21 > 65	≤ 6,5 > 22					< 1,5 > 35	≤ 1					≤ 40 > 1000
Yenidoğan		< 33 > 70	< 9,5 > 22											

Figure 1. Flow chart of the proposed system


Suicide Group
Healthy Group
Complete blood count values of the groups (302x13)
Complete blood count values excluded from reference ranges (260x13)
Machine Learning
BT
SVM
Decision

2.1. Classifiers

2.1.1. Ensemble Methods

The Bagging Tree (BT) classifier, which is the community learning method, is designed to increase the level of success as a result of using collaborative classification algorithms (Tan et al.,2015). Instead of a single learner, a decision tree is created with more than one copy from the main learner and the classifier output is combined with the voting method (Karasu and Saraç, 2018).

Bagging (bootstrap combination) is a collection of decision trees used for regression (Breiman,1996). Community methods use multiple models to achieve better predictive performance by combining many poor learner results into a high-quality community predictor by bagging. Bagging a community of decision trees is a variance reduction technique that aims to improve predictive performance of decision trees. The bagged community power is described as calculating estimates on out-of-bag observations for each tree and averaging over the whole community for each observation. The estimated out-of-bag response is compared to the actual value of each observation. By comparing predicted out-of-bag responses to actual responses for all observations used in education, the average out-of-bag error is estimated. The average error out of this bag is a neutral estimator for the true union error (Breiman, 1996b), (Lupaşcu et al., 2013).

2.2.1. Support Vector Machine

SVM is an effective method in which the data is classified by drawing the boundary between the members of the two groups in the plane (Suykens and Vandewalle, 1999). It is advantageous in terms of being applicable to linear and nonlinear data, having high accuracy, modeling ability at complex decision boundaries, and working with many independent variables (Suykens and Vandewalle, 1999), (Ari et al., 2019). SVM decision function; The x_i i data point, x_* is a test vector, the x_i Lagrangian multiplier associated with the training example a_i, the class (-1 or +1) of the data point y_i i and the b bias value are defined as follows:

$$f(x_*) = sign\left[\sum_{i=1}^{N} a_i\, y_i \varphi\left(x_*, x_i\right) + b\right] \qquad (1)$$

For the expression of the quadratic (2nd degree) optimization problem, for each $\{(x_i, y_i)\}$ ρ is the width of the separator between the support vector classes, w is the normal of the multi-plane (weight vector), ‖w‖ is the representation of w.

It is maximized by $\rho = \frac{2}{\|w\|}$ (Hwang et al.,2011).

If $y_i = 1; w^T x_i + b \geq 1$

If $y_i = -1; w^T x_i + b \geq -1$.

3. Experimental Results

On the Matlab software to test the proposed method; EM and SVM classification algorithms are used. A total of 3380 features belonging to the healthy group and the suicide group were given to the classifiers with 5 cross validity. The 5-fold cross validation attributes are divided into 5 groups first and the first group is used to train the system, and the remaining four groups are used to test the mentioned steps, and each group is trained as a training group, and the rest is a test group. Estimation accuracy is given in Table 3 by taking the average of the results.

Table 3. Suicide *estimation method and accuracy from blood values*

Method	**Method Type**	**Predict accuracy (%)**
Ensemble	Bagged Trees	93,5
Support Vektor Machine	Quadratik	91,2

While the system's accuracy was found as the performance evaluation criterion, the complexity matrix of the system was used and the True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) metrics were processed (Forbes, 1995). In the study, if the mentioned values are to be explained from the suicide estimate: TP: It is stated that the system has correctly identified the person who committed suicide, FP: The system indicates that a healthy individual has committed suicide, TN: The system has been identified as a healthy individual, and FN has been reported as a healthy diagnosis of the suicidal individual.

Accuracy: It is obtained by dividing the correct classification value by the total value (Forbes, 1995).

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{143+100}{260} = 0{,}93 \qquad (2)$$

Sensivity: It is the effectiveness of the system in knowing the truth. TP is found by dividing FN and TP values by sum, and this is the degree to which the probability of suicide can be determined (Forbes, 1995).

$$\frac{TP}{TP+FN} = \frac{143}{143+9} = 0{,}94 \qquad (3)$$

Specifity: It is found in the system by proportioning the sum of the TN and FP values to the cases with TN, and the real state is to what extent the individuals with other gender are correctly determined (Forbes, 1995).

$$\frac{TN}{TN+FP} = \frac{100}{100+8} = 0{,}92 \qquad (4)$$

In the study, the ROC curve, which enables the diagnosis of suicide status and the performance obtained by controlling the validity status, is given in Figure 2.

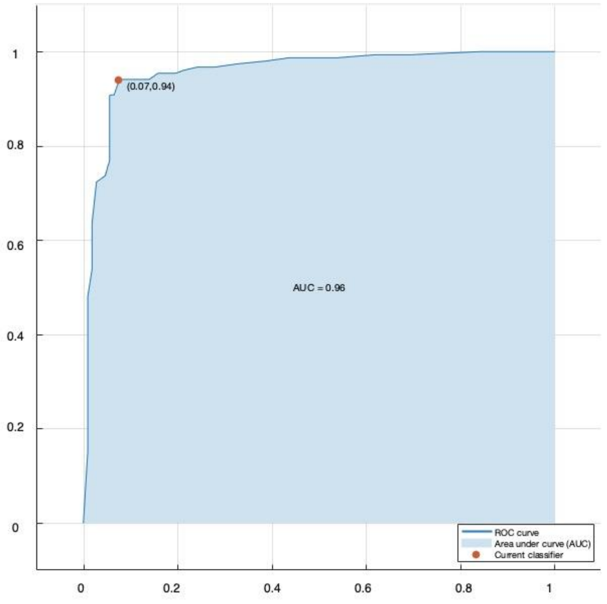


Figure 2. Roc curve (False Positive-Real Positive Ratio)

4. Conclusions

In today's world, where psychological disorders increase and result in undesirable conditions such as suicide, it is important to diagnose diseases early and in expert ways. Many of the patients with psychiatric disorders are thinking of attempting suicide and sometimes doing so. The designed system showed the blood values-suicide relationship with 93.5% success by using BT method which is one of the machine learning approaches. It has been shown that the probability of attempting suicide can be found with an effective method if the individual's blood value measurements are known.

Reference

Aslan, R. (2019). Sağliği Ve Hastaliklari Kan Sayimi Ile Okumak: Hemogram. *Ayrıntı Dergisi*, *7*(76).

İlaslaner, T., & Güven, A. (2019, October). Investigation of the Effects Biochemistry on Iron Deficiency Anemia. In *2019 Medical Technologies Congress (TIPTEKNO)* (pp. 1-4). IEEE.

Deepa, M. N., Gunasekaran, S., Elakiya, R., Haritha, U., Kaleeswari, N., & Purnima, P. Integration of SVM with Artificial Fish Swarm Algorithm for Blood Tumour Prediction.

Ucuz, İ., Özcan, Ö., Mete, B., Arı, A., Tetik, B. K., & Yıldırım, K. (2019). Evaluation of inflammatory markers in childhood-onset psychiatric disorders by using artificial intelligence architectures. *Anatolian Journal of Psychiatry*, 0-0.

Gupta, R. K., Chen, M., Malcolm, G. P., Hempler, N., Dholakia, K., & Powis, S. J. (2019). Label-free optical hemogram of granulocytes enhanced by artificial neural networks. *Optics express*, *27*(10), 13706-13720.

Putin, E. et al. (2016). Deep biomarkers of human aging: application of deep neural networks to biomarker development, Aging, 8, 1021-33.

Ayhan, M. G., Dağistan, A. A., Tanrikulu, C. Ş., Bozdoğan, Ş. Y., & Eren, I. (2019). İntihar girişiminde bulunanlarda artmış nötrofil/lenfosit oranı. *Anadolu Psikiyatri Dergisi*, *20*(3), 305-312.

Çayci, T., Kurt, Y. G., Honca, T., Taş, A., Özgürtaş, T., Ağilli, M., ... & Aydin, I. (2015). Hastane Bilgi Sistemindeki Kayıtlı Hasta Sonuçlarından Tam Kan Referans Aralıklarının Tayini.

Karasu, S., & Saraç, Z. Güç Kalitesi Bozulmalarının 2 Boyutlu Ayrık Dalgacık Dönüşümü ve Torbalama Karar Ağaçları Yöntemi ile Sınıflandırılması. *Politeknik Dergisi*, *21*(4), 849-855.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L., 1996b. Out-of-bag estimation. Technical Report, Department of Statistics, University of California, Berkeley.

Lupaşcu, C. A., Tegolo, D., & Trucco, E. (2013). Accurate estimation of retinal vessel width using bagged decision trees and an extended multiresolution Hermite model. *Medical image analysis*, *17*(8), 1164-1180.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293-300.

Hwang, J. P., Park, S., & Kim, E. (2011). A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. *Expert Systems with Applications*, *38*(7), 8580-8585.

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *Journal of Clinical Monitoring*, *11*(3), 189-206.

Arı, B., Arı, A., Ucuz, İ., Özdemir, F. Ç., & Şengür, A. Grafik Tablet Kullanılarak Makine Öğrenmesi Yardımı ile El Yazısından Cinsiyet Tespiti.