IJIRR

International Journal of Information Research and Review
Vol. 09, Issue, 08, pp.7433-7439, August, 2022

RESEARCH ARTICLE

EXPERIMENTAL RESULTS ON THE AUC OF THE BI-GENERALIZED EXPONENTIAL ROC MODEL USING SPREAD SHEET FUNCTIONS

Prasuna.Ch[1,*] and R.V.S. Nagabhushana Rao[2] and Dr.K.Sujatha[3]

Academic consultant, Dept of Statistics, VS Univerisity

ARTICLE INFO

Article History:
Received 19th May, 2022
Received in revised form
07th June, 2022
Accepted 19th July, 2022
Published online 29th August, 2022

Keywords:
ROC, AUC, Bi-Generalized Exponential Distribution, Sensitivity, Specificity.

ABSTRACT

In medical diagnosis, testing of the performance of a particular diagnostic test is a common and good practice, after classifying the subjects in to different groups by various classification techniques, in particularly binary classification. Assessment of the performance of a diagnostic test can be achieved by Area under the Receiver Operating Characteristic Curve, simply denoted by AUC. For Diseased (D) and Healthy (H) normal populations, Bi-exponential model gives a closed form expression to the Area under the Curve (AUC). In this paper we report the results of simulated experiments on the properties of the AUC of the bi-generalized exponential ROC model. We first study the sensitivity of the AUC to changes in Scale parameter (λ) and Location parameter (μ). We will show that changes in λ do not alter the AUC of the model but with fixed scale parameter in D and H groups the AUC changes quickly when the distance between the location parameters is changed. Numerical illustrations for the proposed method are given with simulated data.

INTRODUCTION

The Receiving Operating Characteristic (ROC) curve is a graphical tool for evaluating the performance of a binary classifier. Even though it originated during Second World War, (Green and Swets, 1966, Lloyd (1998)), many researchers highlighted its significance in medicine, experimental psychology, finance, banking, data mining etc., in later years. In biomedical applications ROC curve is used for assessing the effectiveness of a measured marker (like tumor volume) in distinguishing between diseased and healthy individuals. A person is classified as positive or negative as $X > c$ or $X \leq c$. Let X denote a continuous random variable indicating the measured value for a marker on an individual and c be a cutoff value. The result of classification is compared with a result diagnosis (often with a gold standard). After classification, the outcome will have four possible states viz., True Positive (TP), False Positive (FP),True Negative (TN) and False Negative (FN). The objective is to find out such a classification rule that minimizes the chance of occurrence of FP and FN cases. The Probability of false positive fraction {TN/(TN+FP)} is known as the Specificity (S_p) of the test while the probability of true positive fraction {TP/(TP+FN)} is called the Sensitivity (S_n). The ROC curve is a tradeoff between S_n and (1-S_p) at all possible cutoff values. At each cutoff, we get pair (S_n, 1-S_p) and the plot of S_n against (1-S_p) gives the ROC curve.

ROC Curve is used for the following purposes

- To estimate the probability of correct classification of an unknown individual into one of the two known groups
- To identify the optimal cut-off value for the marker under consideration.

The performance of a classifier can be understood from empirical data in terms of percentage of misclassification. A good classifier shall have nearly zero percent misclassification. In the case of a theoretical ROC model, the performance is measured in terms of the Area under the ROC curve (AUC). Larger the AUC value better will be the classification power of the classifier. A perfect test will have zero misclassification and hence AUC = 1. When $S_n = (1-S_p)$ it means the test is unable to distinguish between TP and FP cases. The ROC curve for such test will be a 45^o line connecting the points (0,0) with (1,1) producing 50% chance of correct classification. For a good marker, it is expected that $S_n > (1- S_p)$ at any cutoff and hence markers having AUC > 0.5 have interest. When the underlying distributions are not known, we use non-parametric method to estimate the ROC curve and AUC in this method reduces to the well known Mann-Whitney U test (Hanley and Neil (1982)).

The Bi-normal and Bi-exponential models: We denote the healthy and diseased groups by H and D respectively. Let μ_H and μ_D denote the means and σ_D^2 and σ_H^2 denote the variances of the normal distribution in the H and D groups respectively.

Corresponding author:** ***Prasuna.Ch,
Academic consultant, Dept of Statistics, VS Univerisity.

The bi-normal ROC model produces a closed form expression to plot the curve as well as to calculate the AUC. Properties of this model can be found in Kronozowski and Hand (2009). Then the bi-normal ROC model is of the form

$$ROC(t) = \Phi\,(a + b\,\Phi^{-1}(x(t)) \qquad (1)$$

where $a = \{\mu_D-\mu_H\}/\sigma_D$ and $b = \sigma_D/\sigma_H$, $\Phi(\ )$ is denotes the cumulative N(0,1) and x(t) is a parameter indicating the False Positive Rate ($0 \le x(t) \le 1$) which is obtained at possible cutoff value . When different cutoff values are close to each the ROC curve becomes smooth. Faraggi and Reiser (2002) have shown that with bi-normal model, the AUC is given by

$$AUC=\phi\left(\frac{\mu_D-\mu_H}{\sqrt{\sigma_D^2+\sigma_H^2}}\right) \qquad (2)$$

As such the AUC is simply the percentile on the normal distribution. Not all test results display a normal distribution and some of them can be modeled by using a skewed distribution. For instance when the distribution is exponential with parameter λ in both the groups, Betinec (2008) has shown that the bi-exponential ROC is of the form

$$ROC(t)=t^{ç} \text{ where } ç = \frac{\lambda_H}{\lambda_D} \qquad (3)$$

Vishnu vardhan , Pundir and Sameera (2012) have shown that the AUC of bi-exponential model is

$$AUC = \frac{\lambda_D}{\lambda_H+\lambda_D} \qquad (4)$$

As in the case of binormal model here also the AUC has a closed form expression. The Generalized Exponential distribution (GED) is a distribution widely used in reliability theory and survival analysis. A theoretical ROC model with this distribution has potential applications in developing new markers for predicting an event in survival analysis. In the following section a brief review about this distribution is given.

The Generalized Exponential Distribution: The GED also called Exponentiated Weibull distribution is a three parameter Weibull model proposed by Mudholkar and Srivastava, Freimer (1995) as a suitable model for life time distribution. Gupta & Kundu (1997) studied a special case of the GE model assuming the location parameter to be zero and compared its performance with the two-parameter Gamma family and the two-parameter Weibull family. The three-parameter GE is characterized by three parameters (α, λ, μ) where and α, λ, μ denotes the shape, scale and location parameters respectively. Then the density function of GED is given by

$$f(x;\alpha,\lambda,\mu) = \frac{\alpha}{\lambda}\left(1-e^{\frac{-(x-\mu)}{\lambda}}\right)^{\alpha-1} e^{\frac{-(x-\mu)}{\lambda}} \qquad (5)$$

Without loss of generality we take α = 1

$$f(x;1,\lambda,\mu) = (1/\lambda)\, e^{-(x-\mu)/\lambda}\ ; \quad (x > \mu;\ \lambda > 0;\ ) \qquad (6)$$

and the distribution function is

$$F(x;1,\lambda,\mu) = (1-e^{-(x-\mu)/\lambda})\ ; \quad (x > \mu;\ \lambda > 0;). \qquad (7)$$

We use the notation GE(1, λ,μ) to indicate this distribution. The shape of the density function of GED is shown in Figure-1.

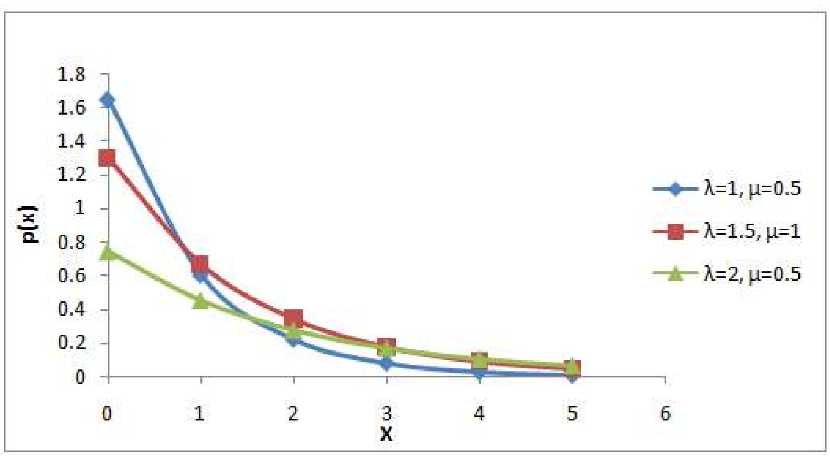

Figure 1. Probability density function of GED(1, λ,μ)

In the following section we derive the Bi-GE ROC model using the first principles.

Bi-GED ROC Model: Let X and Y be the random variables indicating the test value in the H & D groups with underlying GE density functions.

$$g(x;\alpha_{H,}\lambda_H,\mu_H) = \frac{1}{\lambda_H} e^{-\left(\frac{x-\mu_H}{\lambda_H}\right)} \quad \alpha_H=1\;\&\; \lambda_H,\mu_H>0,\ x>0 \qquad (8)$$

$$g(y;\alpha_{D,}\lambda_D,\mu_D) = \frac{1}{\lambda_H} e^{-\left(\frac{x-\mu_D}{\lambda_D}\right)} \quad \alpha_D=1\&\lambda_D,\mu_D>0,x>0 \qquad (9)$$

Since the data on both X and Y is continuous, each data point serves as a possible cutoff denoted by t. At each cutoff t we find the false positive rate x (t) either by counting number of cases of interest or by using a formula (in case of theoretical model). When the location parameters is set zero then $\mu_H = \mu_D = 0$. Again with $\alpha_H = \alpha_D = 1$ we get the case of one parameter bi-GED model proposed by Ehtesham Hussain (2011) as

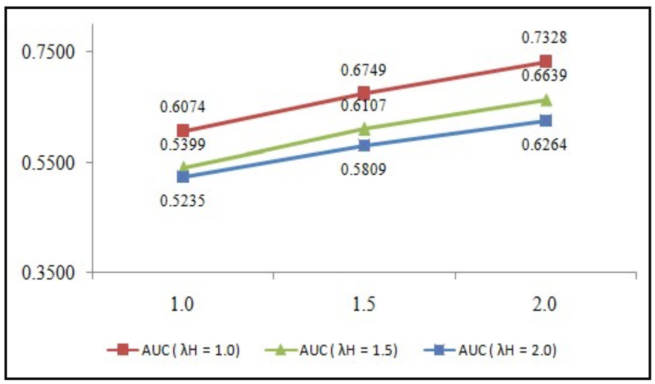

Figure 2. Estimated AUC for different values of λ and fixed values of μ

$$\text{ROC}(t) = 1-(1-t)^{\beta} \qquad 0 \le t \le 1 \qquad (10)$$

where $\beta = \frac{\lambda_D}{\lambda_H}$

Under this model the false positive rate with threshold t is given by

$$x(t) = P(X > t|H) \qquad = 1 - \left[1 - e^{\frac{-(t-\mu_H)}{\lambda_H}}\right] \qquad (11)$$

This gives

$$t = \mu_H - \lambda_H{*}\ \ln(x(t)) \qquad (12)$$

Now the ROC curve is given by

$Y(t) = P(X > t \mid D)$

$$= 1 - \left[1 - e^{\frac{-(t-\mu_D)}{\lambda_D}}\right] \qquad (13)$$

$$= e^{-\left(\mu_H-\lambda_H \ln(x(t))-\mu_D\right)/\lambda_D}$$

$$\Rightarrow \quad Y(t) = x(t)^{\beta} e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \qquad (14)$$

Then the Bi GE ROC model is $Y(t) = x(t)^{\beta} e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)}$ (14) where $\beta = \frac{\lambda_H}{\lambda_D}$.

Substituting for x(t) from (11) we get the ultimate from of ROC curve as

$$Y(t)=e^{-\beta\left(\frac{t-\mu_H}{\lambda_H}\right)}e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \quad (15)$$

In the following section we find a relationship between μ_H and μ_D when a predetermined AUC is required.

Sensitivity of ROC curve and AUC due to changes in λ and μ: From (14) it follows that the bi-generalized exponential ROC model is given by

$$Y(t) = x(t)^{\beta}e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)}, \text{ where } \beta = \frac{\lambda_H}{\lambda_D} \quad (16)$$

Substituting for x(t) from $e^{\frac{-(t-\mu_H)}{\lambda_H}}$ we get the ultimate from of ROC curve as

$$Y(t) = e^{-\beta\left(\frac{t-\mu_H}{\lambda_H}\right)}e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \quad (17)$$

where μ_H , μ_D , λ_H and λ_D are the location and scale of the test result in the two groups.

The AUC is obtained as $\int_0^1 y(t)\,dFP$ and this reduces to

$$AUC = \frac{\lambda_D}{\lambda_D+\lambda_H}e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \quad (18)$$

When the location parameters is set zero then $\mu_H = \mu_D = 0$, we get the case of one parameter bi-exponential ROC model. Let $\hat{\mu}_H$ and $\hat{\mu}_D$ represent the estimated means and $\hat{\lambda}_H$, $\hat{\lambda}_D$ represent the estimated standard deviations in the H and D groups respectively based on samples of size n_1 and n_2.

This gives the estimated AUC as

$$\widehat{AUC} = \frac{\hat{\lambda}_D}{\hat{\lambda}_D+\hat{\lambda}_H}e^{\left(\frac{\hat{\mu}_D-\hat{\mu}_H}{\hat{\lambda}_D}\right)} \quad (19)$$

For numerical illustration the parameters are shown as λ_H and λ_D without ^.

Illustration-1: Let us consider fixed location parameters $\mu_H = 0.30$, $\mu_D = 0.60$. We have selected three levels for λ_D and λ_H as 1.0, 1.5 and 2.0. Three are 9 combinations of λ_H and λ_D at which the AUC is found from (19) and shown in Table- 1. All the functions evaluated using Excel template.

Table 1. List of possible combinations for scale and location parameters

λ_H	λ_D	AUC
1.0	0.5	0.6074
	1.0	0.6749
	1.5	0.7328
1.5	1.0	0.5399
	1.5	0.6107
	2.0	0.6639
2.0	1.5	0.5234
	2.0	0.58091
	2.5	0.6263

The above pattern is shown in Figure-2

ROC model with specified AUC : Suppose we wish to have the model with AUC = A^* (predetermined). From (17) it follows that the AUC of bi-generalized exponential model is given by .

From the above results the following observations can be made

- At each value of λ_H, the AUC increases linearly with increase in the value of λ_D
- As λ_H increases the AUC value decreases

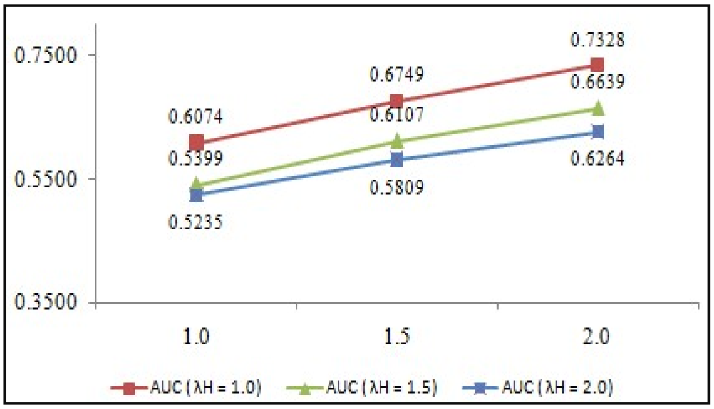

Figure 2. Estimated AUC for different values of λ and fixed values of μ

At each combination of λ_H and λ_D the factor $R = \frac{\lambda_D}{\lambda_D+\lambda_H}$ and the corresponding AUC separately evaluated and shown in Table-2.

Table 2 Estimated AUC and variance ratio (Figures in bracket indicate AUC)

λ_D	λ_H		
	1.0	1.5	2.0
1.0	0.500 (0.6074)	0.400 (0.6749)	0.333 (0.7328)
1.5	0.600 (0.5399)	0.500 (0.6107)	0.429 (0.6639)
2.0	0.667 (0.5235)	0.571 (0.5809)	0.500 (0.6264)

It is found that there is a linear relationship between the factor R and the corresponding the AUC as shown table-3.

Variance ratio (R)	AUC
0.500	0.6074
0.600	0.6749
0.667	0.7328
0.400	0.5399
0.500	0.6107
0.571	0.6639
0.333	0.5235
0.429	0.5809
0.500	0.6264

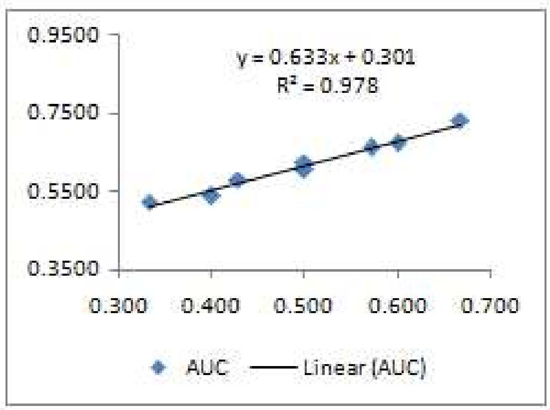

Table 3. Relationship between R and AUC

$$A^* = \frac{\lambda_D}{\lambda_D+\lambda_H} e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \tag{20}$$

Hence a one unit change in the factor R results in 0.633 units of change in the AUC.

ROC model with specified AUC: Suppose we wish to have the model with AUC = A^* (predetermined). From (17) it follows that the AUC of bi-generalized exponential model is given by

Assuming that λ_D and λ_H are known constants and putting $k = \frac{\lambda_D}{\lambda_D+\lambda_H}$ it follows that

$$A^* = k e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)}$$

$$\Rightarrow \ln(A^*/k) = \frac{\mu_D-\mu_H}{\lambda_D}$$

$\Rightarrow \ln(A^*/k)\,\lambda_D = \mu_D - \mu_H$

This gives

$$\mu_D = \mu_H + (\lambda_D * \ln(A^*/k)) \tag{21}$$

Thus the mean score for D group can be calculated from μ_H such that a prefixed AUC is attained.

Suppose $\lambda_D = \lambda_H$. Then we get k = 1/2 and

$$\mu_D = \mu_H + (\lambda_D * \ln(2A^*)) \tag{22}$$

Suppose, we wish to develop an ROC curve such that it has a predetermined A*, like 0.9, 0.8 etc. Using (21) μ_D and the AUC corresponding AUC can be obtained from (20). In the following section the results of a simulated experiment are presented.

Numerical results: Fixing the required A^* as 0.90 we have estimated μ_D by using (5.3.2). By changing the values of n_1, n_2 λ_D, λ_H, and μ_H we have calculated the AUC by using (19) with the help of the Excel spared sheet as shown in Figure 5.1. The experimental results for the target A^* as 0.90 are obtained by changing the values of n_1, n_2,λ_D, and λ_H at fixed μ_H. For each set of input parameters, the corresponding estimate of μ_D is obtained using (21). In the following section, the calculations are made of two new estimates for different values of n_1, n_2 λ_D, λ_H.

Illustration-1: This experiment is performed with 3 different sets of parameters and different sample sizes. In each trail λ_H, λ_D and μ_H are fixed. The correcting μ_D is estimated at the target AUC and shown in square brackets. The AUC is obtained by both simple average and fixed weight methods. The bias from the target is also calculated and the results are shown in Table-4.

Table 4. Estimated AUC and Bias as a function sample size (Figures in bracket indicate bias)

Target A^* = 0.90						
Sample size	Trail-1 λ_H=0.4, λ_D=3.5, μ_H = 0.06 [μ_D =0.07]		Trail-2 λ_H=1, λ_D=4, μ_H = 0.5, [μ_D =0.97]		Trail-3 λ_H=2, λ_D=5, μ_H = 2, [μ_D =2.57]	
	AUC $_{AVG}$	**AUC $_{FW}$**	**AUC $_{AVG}$**	**AUC $_{FW}$**	**AUC $_{AVG}$**	**AUC $_{FW}$**
10	0.7923 (0.1077)	0.8287 (0.0713)	0.7103 (0.1897)	0.7332 (0.1667)	0.6506 (0.2494)	0.6548 (0.2452)
20	0.8548 (0.0452)	0.8664 (0.0331)	0.8120 (0.0880)	0.8050 (0.0949)	0.7788 (0.1212)	0.7515 (0.1484)
50	0.8837 (0.0163)	0.8864 (0.0136)	0.8664 (0.0336)	0.8528 (0.0472)	0.8528 (0.0472)	0.8222 (0.0777)
80	0.8900 ((0.0100)	0.8913 (0.0087)	0.8792 (0.0208)	0.8665 (0.0334)	0.8707 (0.0293)	0.8437 (0.0563)
100	0.8921 (0.0079)	0.8930 (0.0070)	0.8834 (0.0166)	0.8715 (0.0284)	0.8766 (0.0234)	0.8516 (0.0484)
120	0.8935 (0.0065)	0.8940 (0.0059)	0.8862 (0.0138)	0.8750 (0.0249)	0.8805 (0.0195)	0.8572 (0.0428)

Illustration-2: Similar experiment A^* = 0.8 gives the results shown in Table-5

Table 5. Estimated AUC and Bias as a function sample size (Figures in bracket indicate bias)

Target A^* = 0.80						
Sample size	Trail-1 λ_H=0.4, λ_D=3.5, μ_H = 0.06 , [μ_D =0.07]		Trail-2 λ_H=1, λ_D=4, μ_H = 0.5, [μ_D =0.50]		Trail-3 λ_H=2, λ_D=5, μ_H = 2, [μ_D =2.57]	
	AUC $_{AVG}$	**AUC $_{FW}$**	**AUC $_{AVG}$**	**AUC $_{FW}$**	**AUC $_{AVG}$**	**AUC $_{FW}$**
10	0.7279 (0.0721)	0.7852 (0.0148)	0.6514 (0.1486)	0.6940 (0.1060)	0.5950 (0.2045)	0.6186 (0.1814)
20	0.7727 (0.0273)	0.8043 (0.0043)	0.7322 (0.0678)	0.7460 (0.0539)	0.7006 (0.0994)	0.6950 (0.1050)
50	0.7909 (0.0091)	0.8077 (0.0077)	0.7744 (0.0256)	0.7773 (0.0226)	0.7613 (0.0387)	0.7486 (0.0513)
80	0.7946 (0.0054)	0.8080 (0.0080)	0.7843 (0.0157)	0.7855 (0.0145)	0.7760 (0.0240)	0.7641 (0.0358)
100	0.7957 (0.0043)	0.8084 (0.0084)	0.7875 (0.0125)	0.7882 (0.0117)	0.7809 (0.0191)	0.7697 (0.0302)
120	0.7965 (0.0035)	0.8087 (0.0087)	0.7896 (0.0104)	0.7901 (0.0098)	0.7841 (0.0159)	0.7736 (0.0264)

From tables 4 & 5 we observe the following As samples increases, the estimated AUC converges to the target AUC.

The bias of each estimate decreases when sample size increases: Hence with large sample we can estimate the AUC of the bi-generalized exponential ROC model with very small bias. We can compare the significance of the difference between the two estimates as given below.

Hypothesis testing of new estimates with Target AUC: The two estimates can be compared for their significance of difference by t-test. Since the 81 AUC estimates can be treated as a sample of bi-generalized exponential AUC's and each of the two estimates is an average it is possible to compare them with A^* by using Student's t- test. The null hypothesis is that there is no significant difference between the target AUC and the estimated AUC.

The test statistic for comparing AUC_{AVG} with A* is given by

$$t = \left\{ \frac{AUC_{AVG} - A^*}{S/\sqrt{n-1}} \right\} \quad (23)$$

where $S = \sqrt{V(AUC_{AVG})}$ and n is number of available AUC values. In the following section, the calculations are made of two new estimates for different values of n_1, n_2, λ_D and λ_H at fixed μ_H. The statistical significance of the difference between the target AUC and the AUC obtained by the new estimators is tested and the results are shown in Table-6.

Table -6: t-test for comparing the summarized AUCs with the Target (Figures in bracket indicate p-value)

Trial	t-test with A^*= 0.9		t-test with A^*= 0.8	
	AUC_{AVG}	AUC_{FW}	AUC_{AVG}	AUC_{FW}
1	0.8961 (0.1106)	0.8963 (0.1312)	0.7979 (0.4516)	0.7866 (0.0981)
2	0.8866 (0.1361)	0.8616 (0.1000)	0.7887 (0.1216)	0.7748 (0.1008)
3	0.8884 (0.1227)	0.8694 (0.1001)	0.7905 (0.1200)	0.7819 (0.1036)

REFERENCES

Green, D.M and Swets, J.A 1966. "Signal Detection theory and Psychophysics". Wiley, Newyork.

Hanley, J.A, Mc Neil J.B 1982. "The meaning and use of the Area under the Receiver Operating Characteristic curves", Radiology; April; 143; 29-36. 2.

Mudholkar, G.S. Srinivastava, D.K. & Freimer. M. 1995. The exponentiated Weibull family: a reanalysis of the bus motor failure data. Techno metrics 37, 436 – 445.

Gupta, R.D. & Kundu. D. 1997. Exponentiated exponential family: an alternative to gamma and Weibull distribution. Technical report. Dept of Math., Stat. & Comp. Sci., University of New

Brusnwick, Saint – John, NB, Canada.

Lloyd, C. J. 1998. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association, 93, 1356-1364.

Faraggi. D and Reiser, B. 2002. Estimation of the Area under the ROC curve ,Statistics in Medicine; 21 :3093-3106

Betinec. M. 2008. "Testing the difference of the ROC Curves in Bi-exponential Model" Tatra Mountains Mathematical Publications,39,215-223.

Krzanowski W. D and J. Hand 2009. ROC Curves for Continuous data, Monographs on statistics and Applied Probability, CRS Press, Taylor and Francis Group, LLC.

Ehtesham Hussain. 2011. " the ROC Curve Model from Generalized –Exponential distribution", Pak, j. stat. oper.res.Vol.VII No.2, pp323-330.

Vishnu Vardhan, R. SudeshPundir and G.Sameera 2012. " Estimating of Area Under the ROC Curve Using Exponential and Weibull distributions", Bonfring International Journal of Data Mining Vol.2,No.2,June.

****