

# Supplementary material

## “BiTNet: Deep Hybrid Model for Ultrasonography Image Analysis of Human Biliary Tract and Its Applications”

Thanapong Intharah, Kannika Wiratchawa, Yupaporn Wanna, Prem Junsawang, Attapol Titapun, Anchalee Techasen, Arunnit Boonrod, Vallop Laopaiboon, Nittaya Chamadol, Narong Khuntikeo

**I. THE VIEWING ANGLE WAS THE POSITION OF THE ULTRASOUND PROBE WHEN THE IMAGE WAS CAPTURED (PAGE 8).**

Table 1: Viewing Angles Explanation.

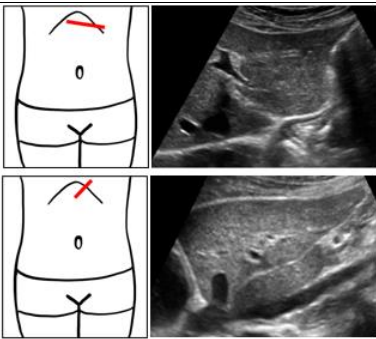
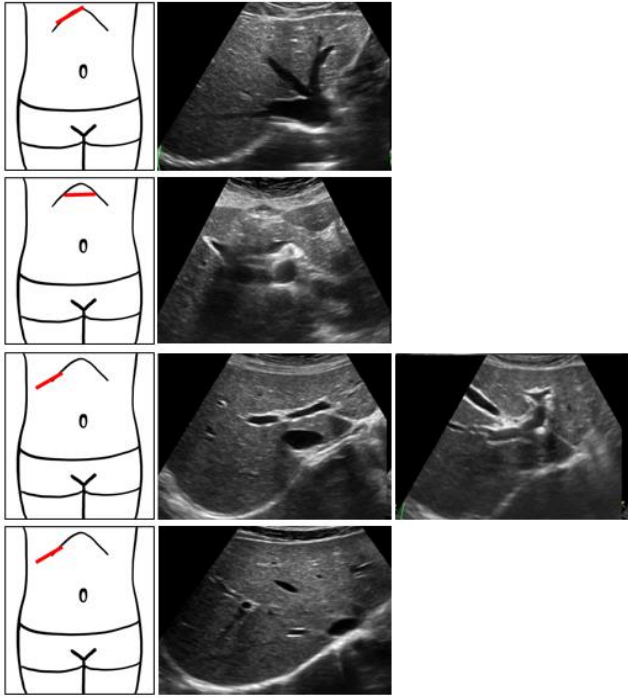
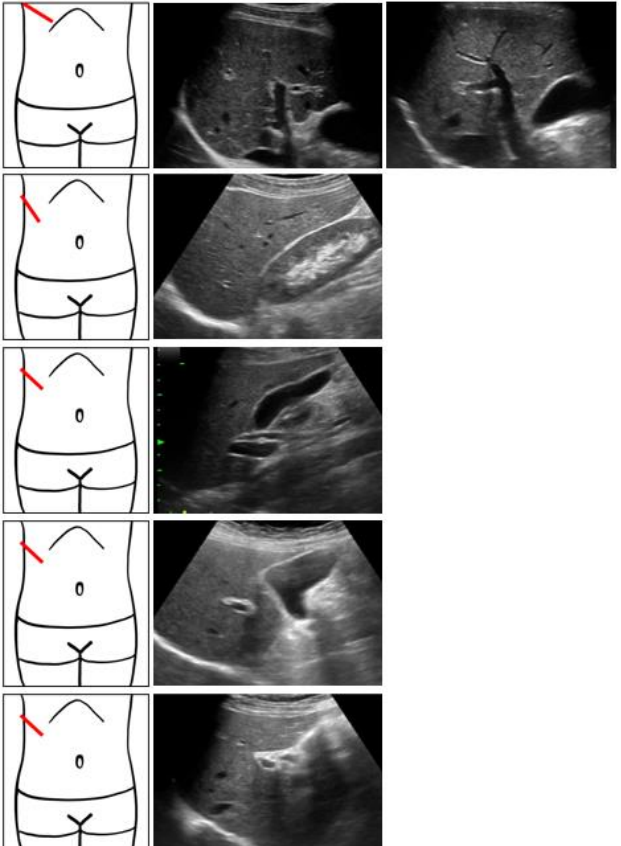
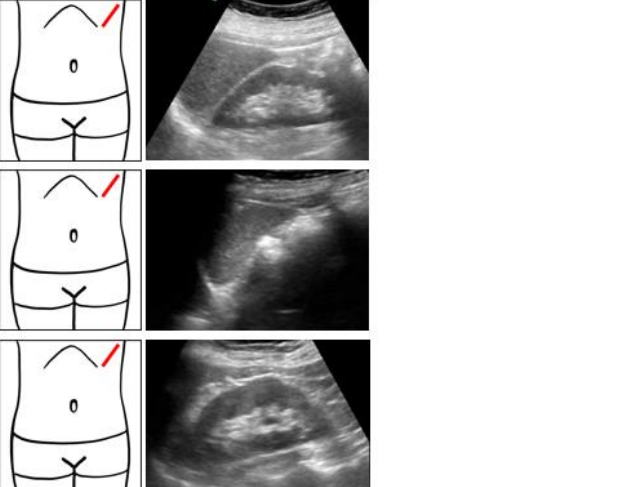
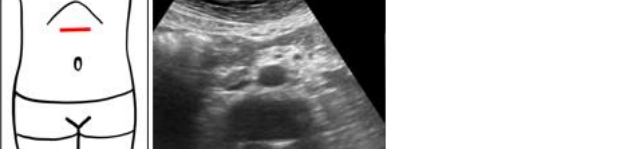
Viewing angle	Main position	Probe position	Key structure	Example images
FP-A	Left lobe liver	Left subcostal scanning and Sagittal plane left upper Q	<ul style="list-style-type: none"> <li>▪ Left lobe liver (S1, 2, 3 and 4)</li> <li>▪ Left portal vein</li> <li>▪ Fissure for ligamentum venosum</li> </ul>	
FP-B	Right lobe liver	Subcostal scanning at epigastric region, Transverse scanning at epigastric region and Right subcostal scanning (Transverse plane at right subcostal region-right upper Q) superior angulation	<ul style="list-style-type: none"> <li>▪ Intrahepatic IVC</li> <li>▪ Hepatic vein (left and middle hepatic vein)</li> <li>▪ Liver S1, S4, S7, 4, 8 and 7</li> <li>▪ Pancreas</li> <li>▪ Right and left portal vein</li> <li>▪ Hepatic vein ,IVC</li> </ul>	

Table 1: Viewing Angles Explanation (continued).

Viewing angle	Main position	Probe position	Key structure	Example images
FP-C	Gallbladder, Common Bile Duct and Right kidney	Sagittal scan right intercostal plane (Just postero-inferior to 5-1) and Sagittal right subcostal scanning	<ul style="list-style-type: none"> <li>▪ Right portal vein</li> <li>▪ Main portal vein</li> <li>▪ Right kidney</li> <li>▪ Liver S5, 6, 7 and 8</li> <li>▪ Liver segment 5 and 6</li> <li>▪ Gallbladder and CBD</li> </ul>	
FP-D	Spleen and Left kidney	Sagittal Left intercostal scanning	<ul style="list-style-type: none"> <li>▪ Spleen</li> <li>▪ Left kidney</li> </ul>	
FP-E	Abdominal aorta	Transverse scanning at epigastric region (midline)	<ul style="list-style-type: none"> <li>▪ Abdominal aorta</li> </ul>	

## II. LIST THE DISTRIBUTION OF THE LABELED IMAGES (PAGE 10).

Table 2: The distribution of the abnormalities versus the viewing angles.

Class number	Label	FP-A	FP-B	FP-C	FP-D	FP-E	Total
1	AB01	105	164	100			369
2	AB02	128	123	77			328
3	AB03	53	31	24			108
4	AB04	105	46	46	3		200
5	AB05	44	78	5			127
6	AB06	76	9				85
7	AB07	3	67	25			95
8	AB081	27	72	57			156
9	AB082	32	56	49			137
10	AB083	11	27	16			54
11	AB09		2	122			124
12	AB10			53			53
13	AB11			73	203		276
14	AB12			1	165		166
Abnormal (Class number 1-14)		584	675	648	371	0	2,278
Normal (Class number 1-14)		748	1,329	1,261	605	348	4,291
Total		1,332	2,004	1,909	976	348	6,569

## III. PERFORMANCE COMPARISON OF DIFFERENT CNN'S WITH DIFFERENT NUMBERS OF PARAMETERS AND INPUT IMAGE SIZES (PAGE 15).

Table 3: Performance comparison of different CNN's with different numbers of parameters and input image sizes.

Networks	Input image size	Parameter (m)	Accuracy	Precision	Recall	F1-score
Eff-B0	224×224	5.3	0.86	0.78	0.60	0.68
Eff-B1	240×240	7.8	0.86	0.74	0.64	0.69
Eff-B2	260×260	9.2	0.86	0.86	0.86	0.86
Eff-B3	300×300	12	0.87	0.87	0.87	0.87
Eff-B4	380×380	19	0.87	0.87	0.87	0.87
Eff-B5	456×456	30	0.88	0.88	0.88	0.88
Eff-B6	528×528	43	0.87	0.87	0.87	0.87
Eff-B7	600×600	66	0.84	0.85	0.84	0.84
ResNet-50	224×224	23.59	0.64	0.47	0.64	0.54
ResNetv2-50	224×224	23.56	0.83	0.83	0.83	0.83
ResNet-101	224×224	42.66	0.65	0.43	0.65	0.52
ResNetv2-101	224×224	42.63	0.80	0.82	0.80	0.81
InceptionResNetV2	299×299	54	0.69	0.62	0.69	0.65
InceptionV3	299×299	22	0.68	0.55	0.68	0.61
NASNetLarge	331×331	84.9	0.71	0.78	0.71	0.74
NASNetMobile	224×224	4.2	0.76	0.77	0.76	0.76

## IV. COMPARISON OF THE PERFORMANCE BETWEEN THE EFFICIENTNET MODEL AND THE BITNET MODEL (PAGE 16).

### ON VALIDATION SET

A. Compare the median of accuracy between the EfficientNet model and the BiTNet model

#### 1) Null and Alternative Hypotheses

$$H_0 : \theta_1 = \theta_2$$

$$H_1 : \theta_1 \neq \theta_2$$

Where

$\theta_1$  = Median of accuracy of the EfficientNet model.

$\theta_2$  = Median of accuracy of the BiTNet model.

#### 2) The Assumption tests

- There is no relationship of accuracy between the EfficientNet model and the BiTNet model.

- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of accuracy score for each model.

The EfficientNet model:

Hypothesis:

$H_0$  : Accuracy scores of the EfficientNet model follow a normal distribution.

$H_1$  : Accuracy scores of the EfficientNet do not follow a normal distribution.

Table 4: Result of Test of Normality of accuracy scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.86	0.12
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.860$ ,  $p = 0.120$ , which indicates that the accuracy scores of the EfficientNet model are normally distributed.

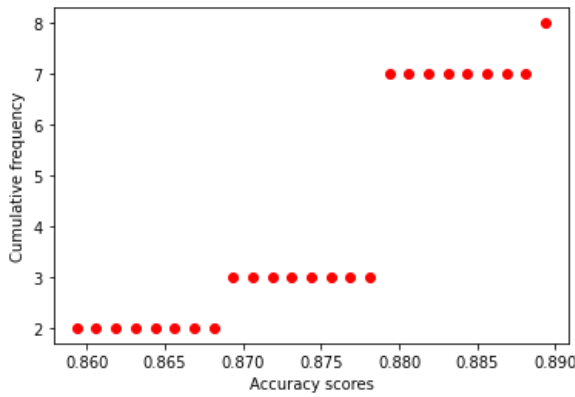


Figure 1: Probability Plots (PP Plots) of accuracy scores of the EfficientNet model.

The BiTNet model:

Hypothesis:

$H_0$  : Accuracy scores of the BiTNet model follow a normal distribution.

$H_1$  : Accuracy scores of the BiTNet model do not follow a normal distribution.

Table 5: Result of Test of Normality of accuracy scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.66	0.00
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test had a significant,  $W = 0.665$ ,  $p = 0.000$ , which indicates that the accuracy scores of the BiTNet model do not follow a normal distribution.

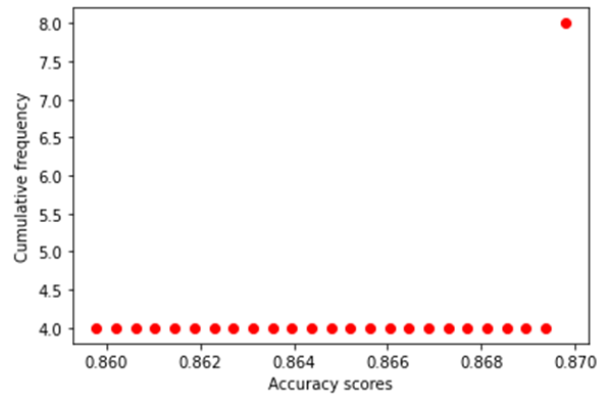


Figure 2: Probability Plots (PP Plots) of accuracy scores of the BiTNet model.

**3) Test Statistics**

To compare group rank differences, we use **Mann Whitney U-Test**, denoted as U.

Table 6: Result of Mann Whitney U-Test between the EfficientNet model and the BiTNet model: accuracy scores.

Mann-Whitney Test	
	EfficientNet $\times$ BiTNet
U	50.00
P-value	$5.32 \times 10^{-2}$
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two - tailed $p \leq 0.01$ was considered statistically significant).	

B. Compare the mean of precision between the EfficientNet model and the BiTNet model

**1) Null and Alternative Hypotheses**

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Where

$\mu_1$  = Mean of precision of the EfficientNet model.

$\mu_2$  = Mean of precision of the BiTNet model.

**2) The Assumption tests**

- There is no relationship of precision between the EfficientNet model and the BiTNet model.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of precision scores for each model.

The EfficientNet model:

Hypothesis:

$H_0$  : Precision scores of the EfficientNet model follow a normal distribution.

$H_1$  : Precision scores of the EfficientNet model do not follow a normal distribution.

Table 7: Result of Test of Normality of precision scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.89	0.23
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.89$ ,  $p = 0.23$ , which indicates that the precision scores of the EfficientNet model follow normally distributed.

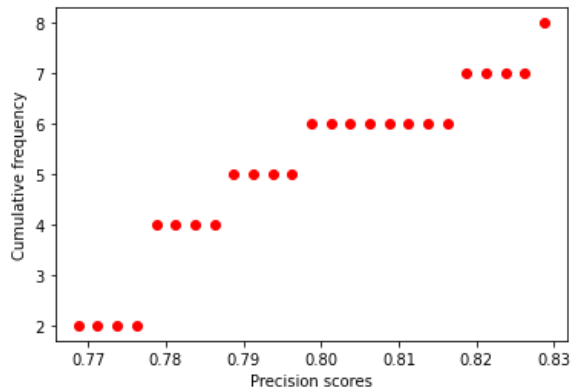


Figure 3: Probability Plots (PP Plots) of precision scores of the EfficientNet model.

The BiTNet model:

Hypothesis:

$H_0$  : Precision scores of the BiTNet model follow a normal distribution.

$H_1$  : Precision scores of the BiTNet model do not follow a normal distribution.

Table 8: Result of Test of Normality of precision scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.88	0.21
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.88$ ,  $p = 0.21$ , which indicates that the precision scores of the BiTNet model follow normally distributed.

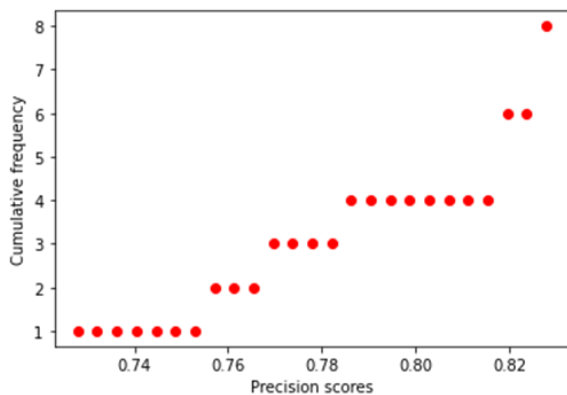


Figure 4: Probability Plots (PP Plots) of precision scores of the BiTNet model.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the precision between the EfficientNet model and the BiTNet model.

Hypothesis

$H_0 : \sigma_1^2 - \sigma_2^2 = 0$

$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$

Where

$\sigma_1^2$  = Variances of the precision of the EfficientNet model.

$\sigma_2^2$  = Variances of the precision of the BiTNet model.

Table 9: Result of Test for Equality of Variances of precision between the EfficientNet model and the BiTNet model.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	3.33	0.08
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 3.33$ ,  $p = 0.08$ , which indicates that the population variances of precision between the EfficientNet model and the BiTNet model are equal. When equal variances are assumed, the calculation uses pooled variances to use the Independent Samples T-Test.

**3) Test Statistics**

We use the **Independent Samples T-Test**, denoted as t. Equal variances are assumed.

Table 10: Result of the Independent Samples T-Test between the EfficientNet model and the BiTNet model: precision scores.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
0.94	-0.08	$-1.25 \times 10^{-3}$	-0.04	0.04
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two-tailed $p \leq 0.01$ was considered statistically significant).				

**4) Interval estimates Using T-score with 99.00% CI**

Table 11: Result of Interval estimates of precision scores using T-score.

Interval estimates using T-score			
Model	Mean of precision scores	99.00% Confident Interval	
		Lower	Upper
EfficientNet	79.25	76.04	82.46
BiTNet	79.37	74.05	84.70

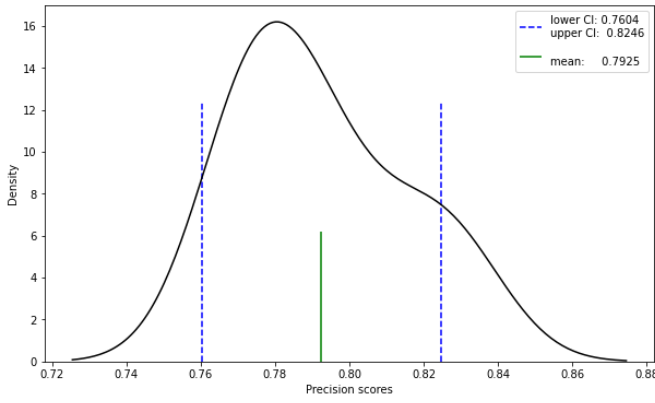


Figure 5: Plot of precision scores of the EfficientNet model, t-statistics - Confidence Level = 99.00%.

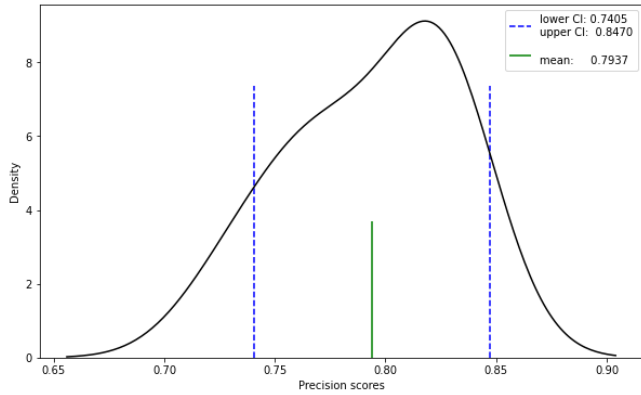


Figure 6: Plot of precision scores of the BiTNet model, t-statistics - Confidence Level = 99.00%.

C. Compare the mean of recall between the EfficientNet model and the BiTNet model

1) Null and Alternative Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where

$\mu_1$  = Mean of recall of the EfficientNet model.

$\mu_2$  = Mean of recall of the BiTNet model.

2) The Assumption tests

- There is no relationship of recall between the EfficientNet model and the BiTNet model.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of recall scores for each model.

The EfficientNet model:

Hypothesis:

$H_0$  : Recall scores of the EfficientNet model follow a normal distribution.

$H_1$  : Recall scores of the EfficientNet model do not follow a normal distribution.

Table 12: Result of Test of Normality of recall scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.96	0.85
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.96$ ,  $p = 0.85$ , which indicates that the recall scores of the EfficientNet model follow normally distributed.

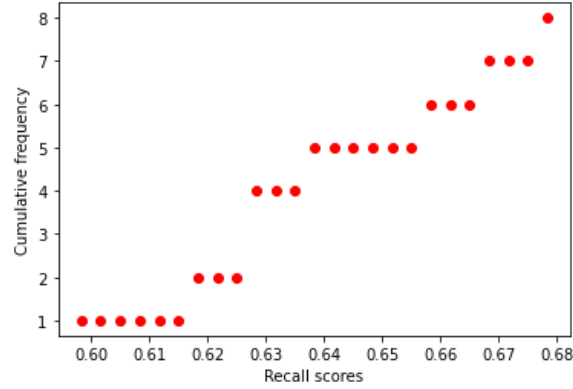


Figure 7: Probability Plots (PP Plot) of recall scores of the EfficientNet model.

The BiTNet model:

Hypothesis:

$H_0$  : Recall scores of the BiTNet model follow a normal distribution.

$H_1$  : Recall scores of the BiTNet model do not follow a normal distribution.

Table 13: Result of Test of Normality of recall scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.97	0.93
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.97$ ,  $p = 0.93$ , which indicates that the recall scores of the BiTNet model follow normally distributed.

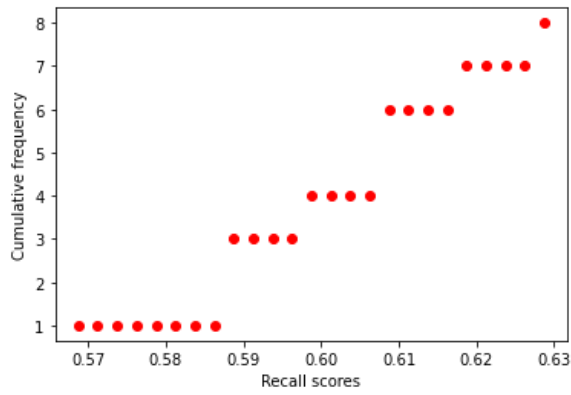


Figure 8: Probability Plots (PP Plot) of recall scores of the BiTNet model.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the recall between the EfficientNet model and the BiTNet model.

*Hypothesis*

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$$

Where

$\sigma_1^2$  = Variances of the recall of the EfficientNet model.

$\sigma_2^2$  = Variances of the recall of the BiTNet model.

Table 14: Result of Test for Equality of Variances of recall between the EfficientNet model and the BiTNet model.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	1.14	0.30
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 1.14$ ,  $p = 0.30$ , which indicates that the population variances of recall between the EfficientNet model and the BiTNet model are equal. When equal variances are assumed, the calculation uses pooled variances to use the Independent Samples T-Test

### 3) Test Statistics

We use the **Independent Samples T-Test**, denoted as  $t$ . Equal variances are assumed.

Table 15: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: recall scores.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$5.07 \times 10^{-3}$	3.32	0.04	$3.98 \times 10^{-3}$	0.07
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two-tailed $p \leq 0.01$ was considered statistically significant).				

### 4) Interval estimates Using T-score with 99.00% CI

Table 16: Result of Interval estimates of recall scores using T-score.

Interval estimates using T-score			
Model	Mean of recall scores	99.00% Confident Interval	
		Lower	Upper
EfficientNet	64.12	60.28	67.97
BiTNet	60.25	57.53	62.97

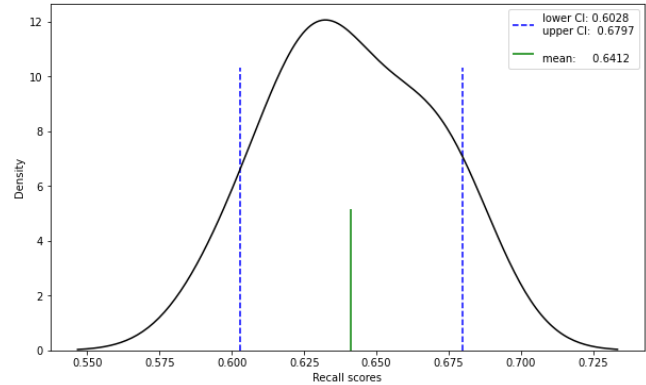


Figure 9: Plot of recall scores of the EfficientNet model, t-statistics - Confidence Level = 99.00%.

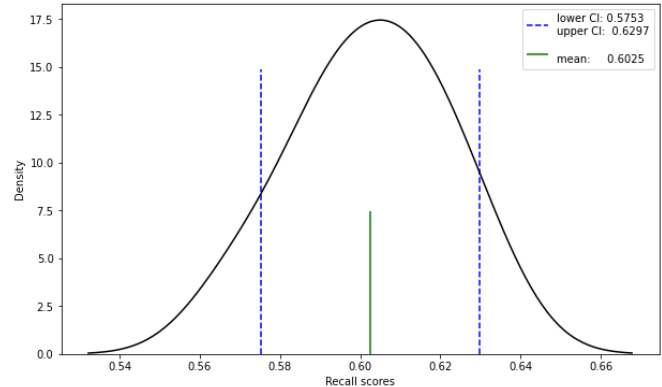


Figure 10: Plot of recall scores of the BiTNet model, t-statistics - Confidence Level = 99.00%.

### ON TEST SET

A. Compare the mean of accuracy between the EfficientNet model and the BiTNet model

#### 1) Null and Alternative Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where

$\mu_1$  = Mean of the accuracy of the EfficientNet model.

$\mu_2$  = Mean of the accuracy of the BiTNet model.

#### 2) The Assumption tests

- There is no relationship of accuracy between the

EfficientNet model and the BiTNet model.

- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of accuracy scores for each model.

The EfficientNet model:

Hypothesis:

$H_0$  : Accuracy scores of the EfficientNet model follow a normal distribution.

$H_1$  : Accuracy scores of the EfficientNet do not follow a normal distribution.

Table 17: Result of Test of Normality of accuracy scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.83	0.05
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.83$ ,  $p = 0.05$ , which indicates that the accuracy scores of the EfficientNet model follow normally distributed.

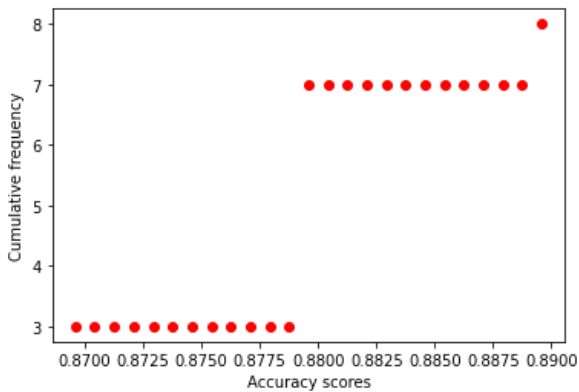


Figure 11: Probability Plots (PP Plots) of accuracy scores of the EfficientNet model.

The BiTNet model:

Hypothesis:

$H_0$  : Accuracy scores of the BiTNet model follow a normal distribution.

$H_1$  : Accuracy scores of the BiTNet do not follow a normal distribution.

Table 18: Result of Test of Normality of accuracy scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.80	0.03
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.80$ ,  $p = 0.03$ , which indicates that the accuracy scores of the BiTNet model follow

normally distributed.

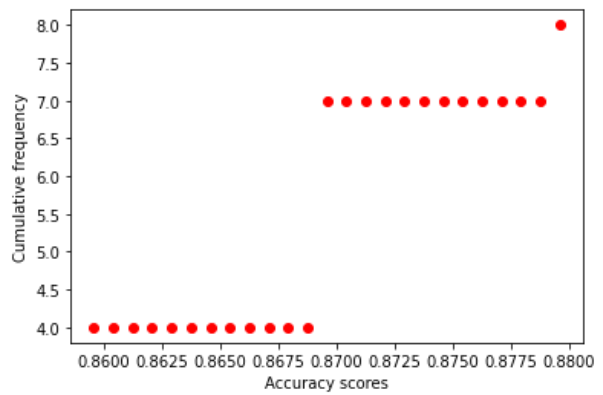


Figure 12: Probability Plots (PP Plots) of accuracy scores of the BiTNet model.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the accuracy between the EfficientNet model and the BiTNet model.

Hypothesis

$H_0 : \sigma_1^2 - \sigma_2^2 = 0$

$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$

Where

$\sigma_1^2$  = Variances of the accuracy of the EfficientNet model.

$\sigma_2^2$  = Variances of the accuracy of the BiTNet model.

Table 19: Result of Test for Equality of Variances of accuracy between the EfficientNet model and the BiTNet model.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	0.13	0.73
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 0.13$ ,  $p = 0.73$ , which indicates that the population variances of accuracy between the EfficientNet and the BiTNet model are equal. When equal variances are assumed, the calculation uses pooled variances to use the Independent Samples T-Test

**3) Test Statistics**

We use the **Independent Samples T-Test**, denoted as t. Equal variances are assumed.

Table 20: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: accuracy scores.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$7.83 \times 10^{-3}$	3.10	0.01	$4.47 \times 10^{-4}$	0.02
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two-tailed $p \leq 0.01$ was considered statistically significant).				



4) Interval estimates Using T-score with 99.00% CI

Table 21: Result of Interval estimates of accuracy scores using T-score.

Interval estimates using T-score			
Model	Mean of accuracy scores	99.00% Confident Interval	
		Lower	Upper
EfficientNet	87.75	86.74	88.76
BiTNet	86.62	85.57	87.68

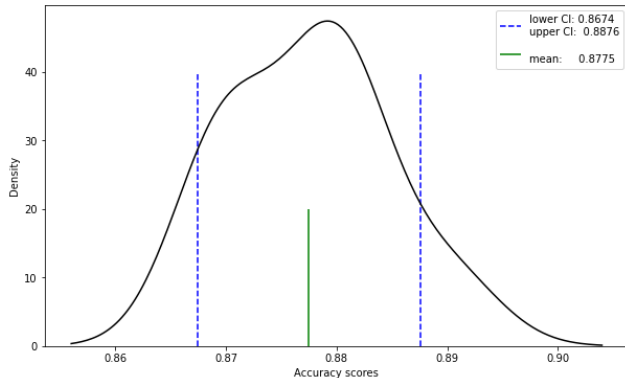


Figure 13: Plot of accuracy scores of the EfficientNet model, t-statistics - Confidence Level = 99.00%.

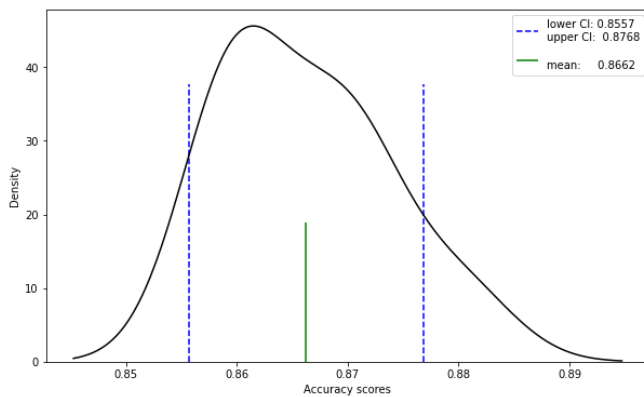


Figure 14: Plot of accuracy scores of the BiTNet model, t-statistics - Confidence Level = 99.00%.

B. Compare the mean of precision between the EfficientNet model and the BiTNet model

1) Null and Alternative Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where

$\mu_1$  = Mean of precision of the EfficientNet model.

$\mu_2$  = Mean of precision of the BiTNet model.

2) The Assumption tests

- There is no relationship of precision between the EfficientNet model and the BiTNet model.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of precision scores for each model.

The EfficientNet model:

Hypothesis:

$H_0$  : Precision scores of the EfficientNet model follow a normal distribution.

$H_1$  : Precision scores of the EfficientNet do not follow a normal distribution.

Table 22: Result of Test of Normality of precision scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.87	0.15
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.87$ ,  $p = 0.15$ , which indicates that the precision scores of the EfficientNet model follow normally distributed.

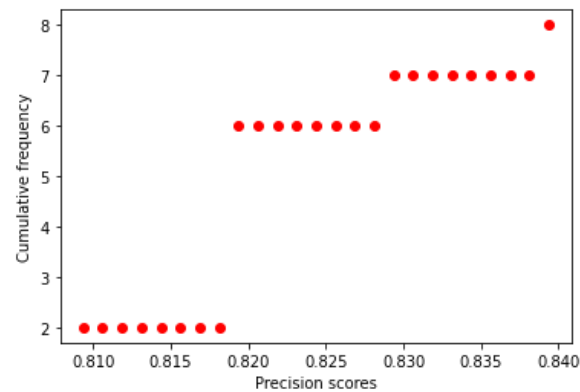


Figure 15: Probability Plots (PP Plot) of precision scores of the EfficientNet model.

The BiTNet model:

Hypothesis:

$H_0$  : Precision scores of the BiTNet model follow a normal distribution.

$H_1$  : Precision scores of the BiTNet do not follow a normal distribution.

Table 23: Result of Test of Normality of precision scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.87	0.15
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.87$ ,  $p = 0.15$ , which indicates that the precision scores of the BiTNet model follow normally distributed.

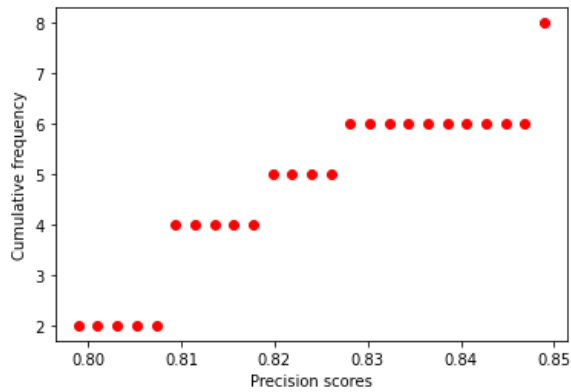


Figure 16: Probability Plots (PP Plots) of precision scores of the BiTNet model.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the precision between the EfficientNet model and the BiTNet model.

*Hypothesis*

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$$

Where

$\sigma_1^2$  = Variances of the precision of the EfficientNet model.

$\sigma_2^2$  = Variances of the precision of the BiTNet model.

Table 24: Result of Test for Equality of Variances of precision between the EfficientNet model and the BiTNet model.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	5.24	0.04
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 5.24$ ,  $p = 0.04$ , which indicates that the population variances of precision between the EfficientNet model and the BiTNet model are equal. When equal variances are assumed, the calculation uses pooled variances to use the Independent Samples T-Test

### 3) Test Statistics

We use the Independent **Samples T-Test**, denoted as t. Equal variances are assumed.

Table 25: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: precision scores.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
1.00	$-1.3 \times 10^{-14}$	$-1.1 \times 10^{-16}$	-0.02	0.02
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two - tailed $p \leq 0.01$ was considered statistically significant).				

### 4) Interval estimates Using T-score with 99.00% CI

Table 26: Result of Interval estimates of precision scores using T-score.

Interval estimates using T-score			
Model	Mean of precision scores	99.00% Confident Interval	
		Lower	Upper
EfficientNet	82.12	80.71	83.54
BiTNet	82.13	79.23	85.02

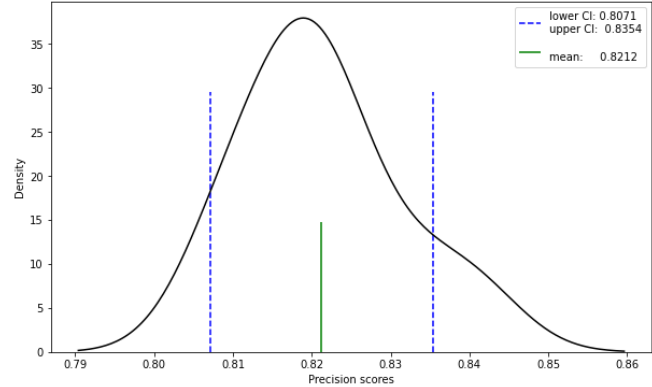


Figure 17: Plot of precision scores of the EfficientNet model, t-statistics - Confidence Level = 99.00%.

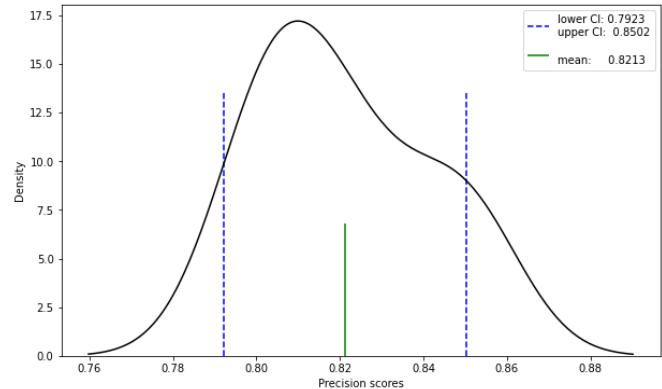


Figure 18: Plot of precision scores of the BiTNet model, t-statistics - Confidence Level = 99.00%.

C. Compare the mean of recall between the EfficientNet model and the BiTNet model

### 1) Null and Alternative Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where

$\mu_1$  = Mean of recall of the EfficientNet model.

$\mu_2$  = Mean of recall of the BiTNet model.

### 2) The Assumption tests

- There is no relationship of recall between the EfficientNet model and the BiTNet model.

- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of recall scores for each model.

The EfficientNet model:

Hypothesis:

$H_0$  : Recall scores of the EfficientNet model follow a normal distribution.

$H_1$  : Recall scores of the EfficientNet do not follow a normal distribution.

Table 27: Result of Test of Normality of recall scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.98	0.96
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.98$ ,  $p = 0.96$ , which indicates that the recall scores of the EfficientNet model follow normally distributed.

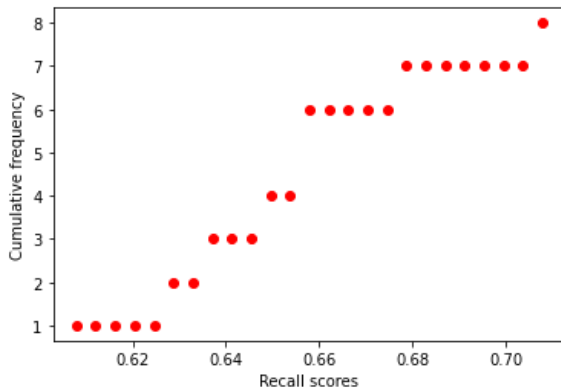


Figure 19: Probability Plots (PP Plot) of recall scores of the EfficientNet model.

The BiTNet model:

Hypothesis:

$H_0$  : Recall scores of the BiTNet model follow a normal distribution.

$H_1$  : Recall scores of the BiTNet model do not follow a normal distribution.

Table 28: Result of Test of Normality of recall scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.95	0.75
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.95$ ,  $p = 0.75$ , which indicates that the recall scores of the BiTNet model follow normally distributed.

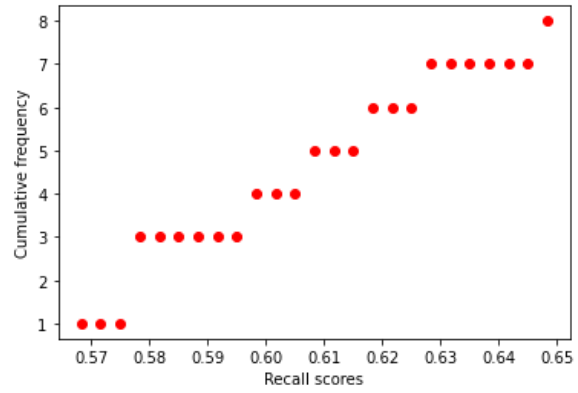


Figure 20: Probability Plots (PP Plot) of recall scores of the BiTNet model.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the recall between the EfficientNet model and the BiTNet model.

Hypothesis

$H_0 : \sigma_1^2 - \sigma_2^2 = 0$

$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$

Where

$\sigma_1^2$  = Variances of the recall of the EfficientNet model.

$\sigma_2^2$  = Variances of the recall of the BiTNet model.

Table 29: Result of Test for Equality of Variances of recall between the EfficientNet model and the BiTNet model.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	$0.76 \times 10^{-30}$	1.0
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 0.76 \times 10^{-30}$ ,  $p = 1.0$ , which indicates that the population variances of recall between the EfficientNet model and the BiTNet model are equal. When equal variances are assumed, the calculation uses pooled variances to use the Independent Samples T-Test

**3) Test Statistics**

We use the **Independent Samples T-Test**, denoted as t. Equal variances are assumed.

Table 30: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: recall scores.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$4.20 \times 10^{-3}$	3.42	0.05	$6.42 \times 10^{-3}$	0.09
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two-tailed $p \leq 0.01$ was considered statistically significant).				

4) Interval estimates Using T-score with 99.00% CI

Table 31: Result of Interval estimates of recall scores using T-score.

Interval estimates using T-score			
Model	Mean of recall scores	99.90% Confident Interval	
		Lower	Upper
EfficientNet	65.50	61.13	69.87
BiTNet	60.50	56.54	64.46

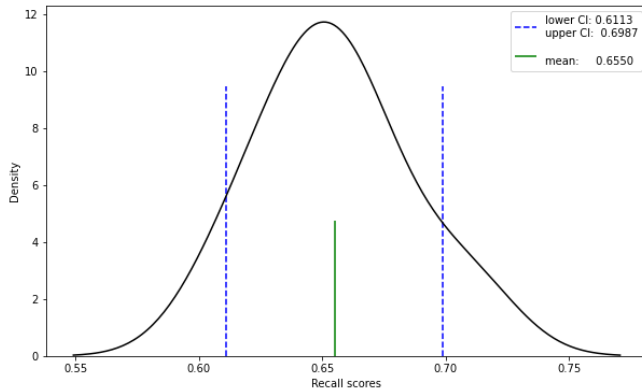


Figure 21: Plot of recall scores of the EfficientNet model, t-statistics - Confidence Level = 99.00%.

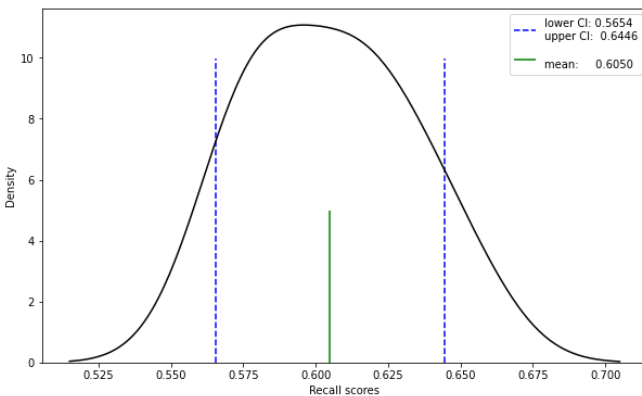


Figure 22: Plot of precision scores of the BiTNet model, t-statistics - Confidence Level = 99.00%.

V. COMPARISON OF THE MEAN DIFFERENCES BETWEEN PREDICTION CONFIDENCE OF THE CORRECT AND INCORRECT GROUPS (PAGE 16).

We use the **Independent Samples T-Test** to compare the means of mean difference in prediction confidence of the correct and incorrect groups between the BiTNet model and the EfficientNet model.

5.1 Null and Alternative Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Where

$\mu_1$  = Mean of mean difference of prediction confidence of the BiTNet model.

$\mu_2$  = Mean of mean difference of prediction confidence of the EfficientNet model.

5.2 The Assumption tests

1) There is no relationship between the mean differences of the BiTNet model and the mean differences of the EfficientNet model.

2) Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of mean difference of prediction confidence for each model.

The BiTNet model:

Hypothesis:

$H_0$  : Mean difference of prediction confidence of the BiTNet model follow a normal distribution.

$H_1$  : Mean difference of prediction confidence of the BiTNet model do not follow a normal distribution.

Table 32: Result of Test of Normality of the mean difference of prediction confidence of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
Mean difference	0.92	$2.72 \times 10^{-2}$
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.92$ ,  $p = 2.72 \times 10^{-2}$ , which indicates that the mean difference of prediction confidence of the BiTNet model is normally distributed.

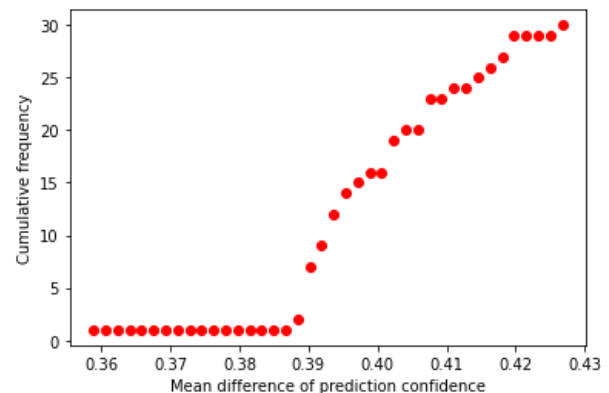


Figure 23: Probability Plots (PP Plots) of the mean difference prediction confidence of the BiTNet model.

The EfficientNet model:

Hypothesis:

$H_0$  : Mean difference of prediction confidence of the EfficientNet model follow a normal distribution.

$H_1$  : Mean difference of prediction confidence of the EfficientNet model do not follow a normal distribution.

Table 33: Result of Test of Normality of the mean difference prediction confidence of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
Mean difference	0.93	$6.27 \times 10^{-2}$
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.93$ ,  $p = 6.27 \times 10^{-2}$ , which indicates that the mean difference of prediction confidence of the EfficientNet model is normally distributed.

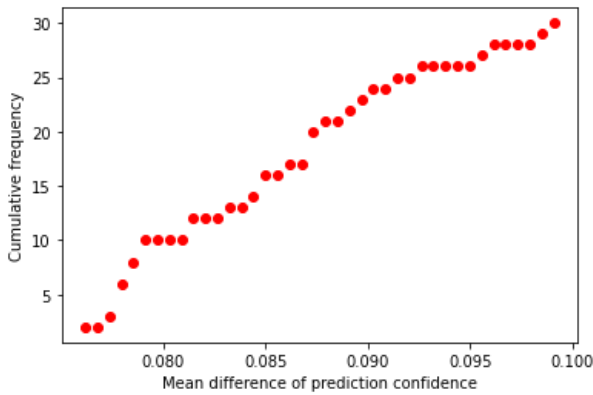


Figure 24: Probability Plots (PP Plots) of the mean difference of prediction confidence of the EfficientNet model.

3) Test of Homogeneity of variances

We use **Levene's Test** to test for the homogeneity of variance of the mean difference of prediction confidence in both models.

*Hypothesis*

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$$

Where

$\sigma_1^2$  = Variances of the mean difference of prediction confidence of the BiTNet model.

$\sigma_2^2$  = Variances of the mean difference of prediction confidence of the EfficientNet model.

Table 34: Result of Test for Equality of Variances of the mean difference of prediction confidence between the BiTNet model and the EfficientNet model.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	8.17	$5.89 \times 10^{-3}$
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 8.17$ ,  $p = 5.89 \times 10^{-3}$ , which indicates that the population variances of the BiTNet model and the EfficientNet model are equal. When equal

variances are assumed, the calculation uses pooled variances to use Independent Samples T-Test.

5.3 Test Statistics

We use the **Independent Samples T-Test**, denoted as t. Equal variances are assumed.

Table 35: Result of the Independent Samples T-Test to compare the means of the mean difference between the BiTNet model and the EfficientNet model.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$2.34 \times 10^{-70}$	-114.60	-0.31	-0.32	-0.30
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one - tailed $p \leq 0.01$ was considered statistically significant).				

5.4 Interval estimates Using T-score with 99.00% CI

Table 36: Result of Interval estimates of the mean differences using T-score.

Interval estimates using T-score			
Model	Mean of mean difference	99.00% Confident Interval	
		Lower	Upper
BiTNet	40.13	39.52	40.74
EfficientNet	8.55	8.25	8.86

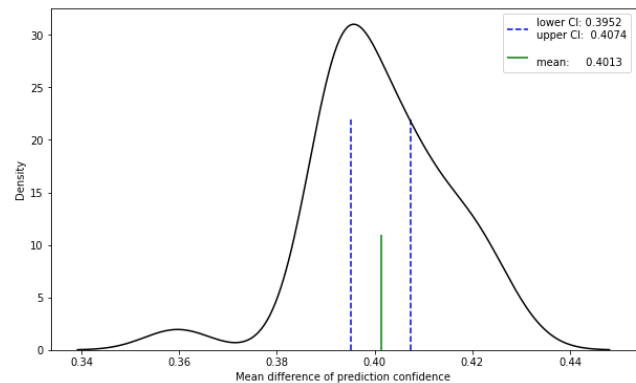


Figure 25: Plot of the mean difference of prediction confidence of the correct and incorrect the BiTNet model, t-statistics - Confidence Level = 99.00%.

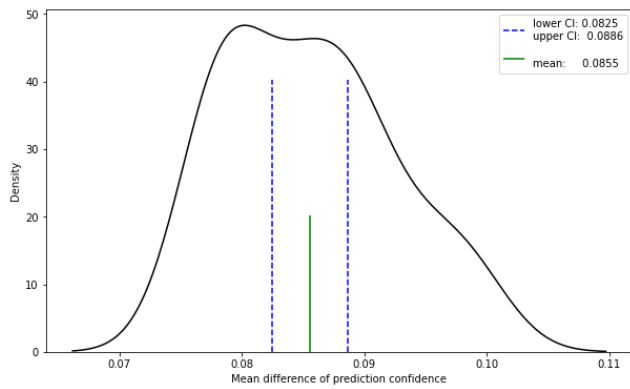


Figure 26: Plot of the mean difference of prediction confidence of the correct and incorrect of the EfficientNet model, t-statistics - Confidence Level = 99.00%.

A. Compare the means of prediction confidence between correct and incorrect the BiTNet model

1) Null and Alternative Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Where

$\mu_1$  = Mean of prediction confidence correct.

$\mu_2$  = Mean of prediction confidence incorrect.

2) The Assumption tests

- There is no relationship of prediction confidence between correct and incorrect.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of the mean for each prediction confidence.

Prediction confidences correct:

Hypothesis:

$H_0$  : Mean of prediction confidence correct follow a normal distribution.

$H_1$  : Mean of prediction confidence correct does not follow a normal distribution.

Table 37: Result of Test of Normality of prediction confidence correct.

	Shapiro-wilk	
	W-test statistic	P-value
Correct	0.96	0.40
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.96$ ,  $p = 0.40$ , which indicates that the mean of confidence correct is normally distributed.

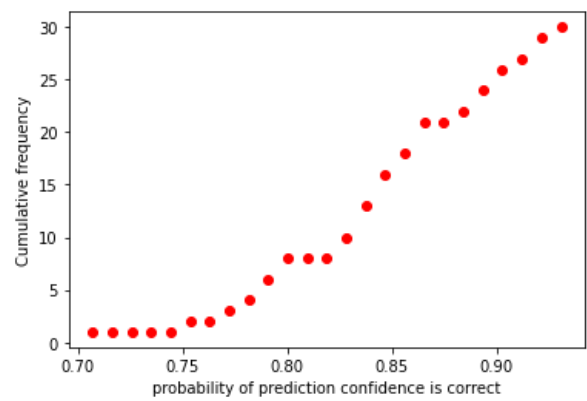


Figure 27: Probability Plots (PP Plot) of prediction confidence is correct.

Prediction confidences incorrect:

Hypothesis:

$H_0$  : Mean of prediction confidence incorrect follows a normal distribution.

$H_1$  : Mean of prediction confidence incorrect does not follow a normal distribution.

Table 38: Result of Test of Normality of prediction confidence incorrect.

	Shapiro-wilk	
	W-test statistic	P-value
Incorrect	0.98	0.72
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.98$ ,  $p = 0.72$ , which indicates that the mean of confidence incorrect is normally distributed.

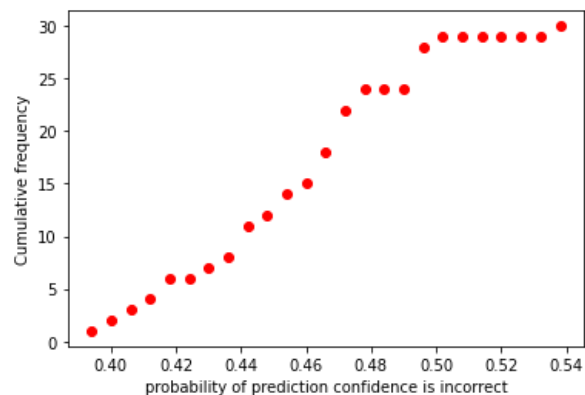


Figure 28: Probability Plots (PP Plot) of prediction confidence incorrect.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the mean of prediction confidence between correct and incorrect.

*Hypothesis*

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$$

Where

$\sigma_1^2$  = Variances of the mean of prediction confidence correct.

$\sigma_2^2$  = Variances of the mean of prediction confidence incorrect.

Table 39: Result of Test for Equality of Variances of the mean of prediction confidence between correct and incorrect.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	4.41	$4.01 \times 10^{-2}$
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 4.41$ ,  $p = 4.01 \times 10^{-2}$ , which indicates that the population variances of correct and incorrect are equal. When equal variances are assumed, the calculation uses pooled variances to use Independent Samples T-Test.

**3) Test Statistics**

We use the **Independent Samples T-Test**, denoted as t. Equal variances are assumed.

Table 40: Result of the Independent Samples T-Test to compare the means of prediction confidence between the correct and incorrect group.

Two sample t-test with equal variance				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$1.03 \times 10^{-39}$	33.17	0.39	0.35	0.42
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one-tailed $p \leq 0.01$ was considered statistically significant).				

*B. Compare the means of prediction confidence between correct and incorrect the EfficientNet model*

**1) Null and Alternative Hypotheses**

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Where

$\mu_1$  = Mean of prediction confidence correct.

$\mu_2$  = Mean of prediction confidence incorrect.

**2) The Assumption tests**

- There is no relationship of mean of prediction confidence between correct and incorrect.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of mean prediction confidence.

Prediction confidences correct:

Hypothesis:

$H_0$  : Mean of prediction confidence correct follow a normal distribution.

$H_1$  : Mean of prediction confidence correct does not follow a normal distribution.

Table 41: Result of Test of Normality of the mean of prediction confidence correct.

	Shapiro-wilk	
	W-test statistic	P-value
Correct	0.87	$2.0 \times 10^{-3}$
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.87$ ,  $p = 2.00 \times 10^{-3}$ , which indicates that the mean of confidence correct is normally distributed.

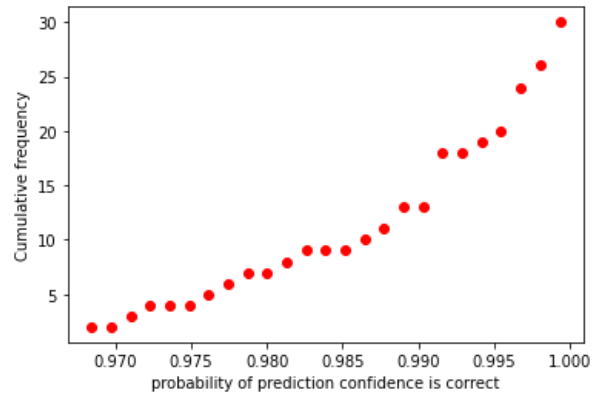


Figure 29: Probability Plots (PP Plot) of the mean prediction confidence correct.

Prediction confidences incorrect:

Hypothesis:

$H_0$  : Mean of prediction confidence incorrect follow a normal distribution.

$H_1$  : Mean of prediction confidence incorrect does not follow a normal distribution.

Table 42: Result of Test of Normality of the mean prediction confidence incorrect.

	Shapiro-wilk	
	W-test statistic	P-value
Incorrect	0.97	0.81
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.97$ ,  $p = 0.81$ , which indicates that the mean of confidence incorrect is normally distributed.

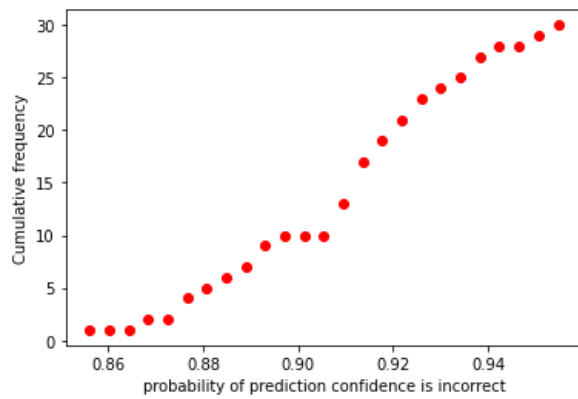


Figure 30: Probability Plots (PP Plot) of the mean prediction confidence incorrect.

- Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of the mean of prediction confidence between correct and incorrect.

*Hypothesis*

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$$

Where

$\sigma_1^2$  = Variances of the mean of prediction confidence correct.

$\sigma_2^2$  = Variances of the mean of prediction confidence incorrect.

Table 43: Result of Test for Equality of Variances of the mean of prediction confidence between correct and incorrect.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance not assumed	15.23	$2.51 \times 10^{-4}$
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $F = 15.23$ ,  $p = 2.51 \times 10^{-4}$ , which indicates that the population variances of correct and incorrect are not equal. When equal variances are not assumed, the calculation utilizes un-pooled variances to use the Independent Samples T-Test.

### 3) Test Statistics

We use **Independent Samples T-Test**, denoted as t. Equal variances are not assumed.

Table 44: Result of the Independent Samples T-Test to compare the means of prediction confidence between the correct and incorrect group.

Two sample t-test with unequal variance (Welch's t-test)				
			99.00% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
$1.22 \times 10^{-18}$	15.74	0.07	0.06	0.08
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one - tailed $p \leq 0.01$ was considered statistically significant).				

## VI. COMPARES PERFORMANCE OF PARTICIPANTS BETWEEN ASSISTED VS UNASSISTED (PAGE 21).

We use **Paired Samples T-Test** to compare the performance of participants with assisting tool and without assisting tool.

*A. Impact of the assisting tool by comparing the performance of participants in accuracy scores*

### 1) Null and Alternative Hypotheses

$$H_0 : \mu_2 = \mu_1$$

$$H_1 : \mu_2 > \mu_1$$

Where

$\mu_1$  = Mean of accuracy among participants without assisting tools.

$\mu_2$  = Mean of accuracy among participants with assisting tool.

### 2) The Assumption tests

- There is the relationship between accuracy scores among participants with assisting tool and without assisting tool.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of accuracy scores difference between assisted and unassisted.

Hypothesis:

$H_0$  : Accuracy scores difference among participants with assisting tool and without the tool follow a normal distribution.

$H_1$  : Accuracy scores difference between among participants with assisting tool and without the tool do not follow a normal distribution.

Table 45: Result of Test of Normality of accuracy scores difference between among participants with assisting tool and without the tool.

	Shapiro-wilk	
	W-test statistic	P-value
Assisted - Unassisted	0.90	0.24
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.90$ ,  $p = 0.24$ , which indicates that the accuracy scores both with assisting tools and without assisting tools are normally distributed.



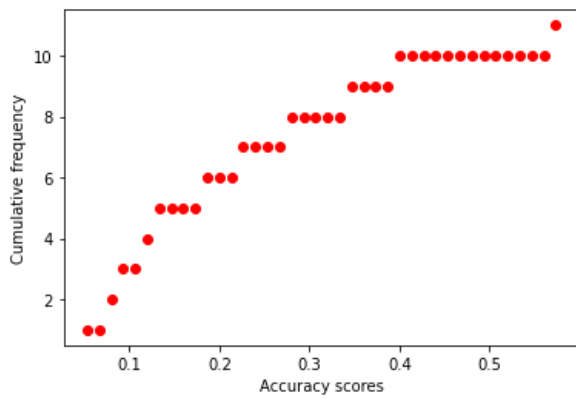


Figure 31: Probability Plots (PP Plot) of accuracy scores difference (assisted - unassisted).

### 3) Test Statistics

To compare the means for assisted and unassisted, we used **Paired Samples T-Test**, denoted as *t*.

Table 46: Result of Paired Samples T-Test between with assisting tool and without assisting tool: accuracy scores.

Paired t-test				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$3.44 \times 10^{-4}$	4.83	35.27	12.14	58.40
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one-tailed $p \leq 0.01$ was considered statistically significant).				

### 4) Interval estimates Using T-score with 99.00% CI

Table 47: Result of Interval estimates of accuracy scores using T-score.

Interval estimates using T-score			
Group	Mean of accuracy scores	99.00% Confident Interval	
		Lower	Upper
Assisted	73.52	62.50	84.53
Unassisted	50.00	30.95	69.05

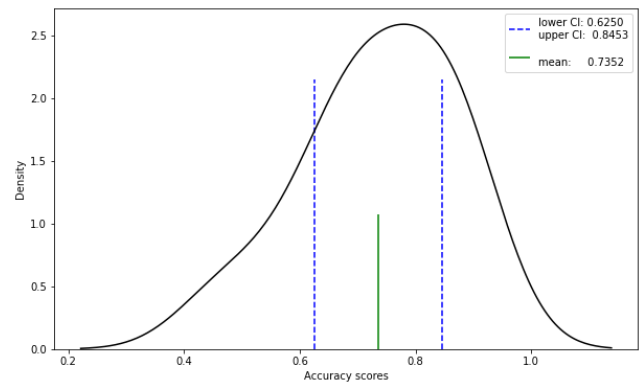


Figure 32: Plot of accuracy scores among participants with assisting tool, t-statistics - Confidence Level = 99.00%.

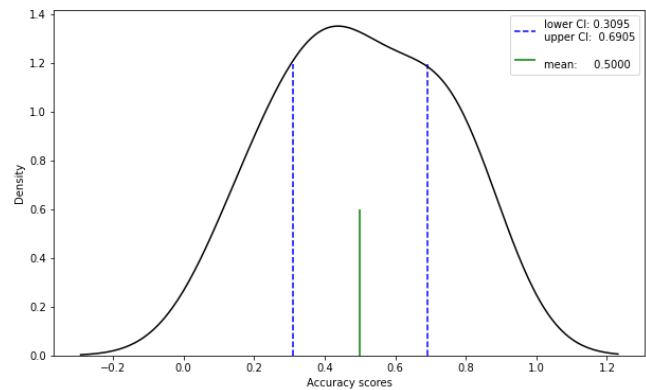


Figure 33: Plot of accuracy scores among participants without assisting tool, t-statistics - Confidence Level = 99.00%.

### B. Impact of the assisting tool by comparing the performance of participants in precision scores

#### 1) Null and Alternative Hypotheses

$$H_0 : \mu_2 = \mu_1$$

$$H_1 : \mu_2 > \mu_1$$

Where

$\mu_1$  = Mean of precision among participants without assisting tool.

$\mu_2$  = Mean of precision among participants with assisting tool.

#### 2) The Assumption tests

- There is the relationship between precision scores among participants with assisting tools and without assisting tools.
- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of precision scores difference between assisted and unassisted.

Hypothesis:

$H_0$  : Precision scores difference among participants with assisting tool and without the tool follow a normal distribution.

$H_1$  : Precision scores difference among participants with assisting tool and without the tool do not follow a normal

distribution.

Table 48: Result of Test of Normality of precision scores difference among participants with assisting tool and without the tool.

	Shapiro-wilk	
	W-test statistic	P-value
Assisted - Unassisted	0.95	0.62
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.95$ ,  $p = 0.62$ , which indicates that the precision scores both with assisting tool and without assisting tool are normally distributed.

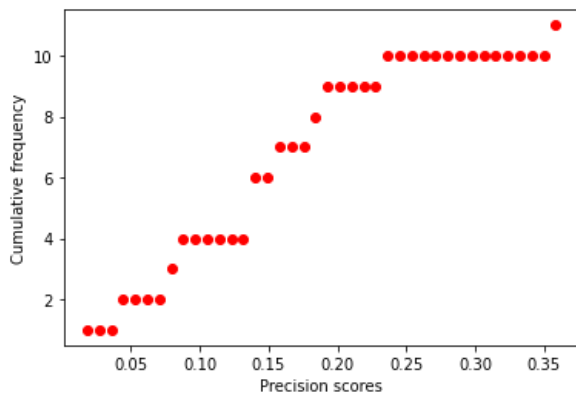


Figure 34: Probability Plots (PP Plot) of precision scores difference (assisted - unassisted).

### 3) Test Statistics

To compare the means for assisted and unassisted, we used **Paired Samples T-Test**, denoted as  $t$ .

Table 49: Result of Paired Samples T-Test between with assisting tool and without assisting tool: precision scores.

Paired t-test				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$1.58 \times 10^{-4}$	5.37	15.39	0.06	0.24
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one - tailed $p \leq 0.01$ was considered statistically significant).				

### 4) Interval estimates Using T-score with 99.00% CI

Table 50: Result of Interval estimates of precision scores using T-score.

Interval estimates using T-score			
Group	Mean of precision scores	99.00% Confident Interval	
		Lower	Upper
Assisted	61.49	49.08	73.90
Unassisted	46.10	32.56	59.63

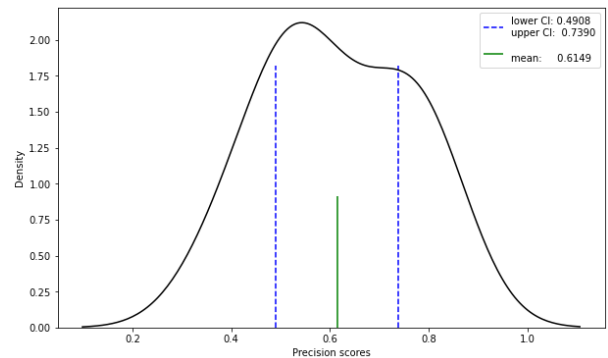


Figure 35: Plot of precision scores among participants with assisting tool, t-statistics - Confidence Level = 99.00%.

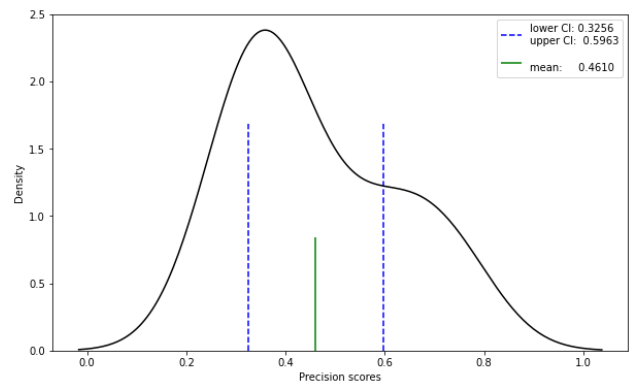


Figure 36: Plot of precision scores among participants without assisting tool, t-statistics - Confidence Level = 99.00%.

### C. Impact of the assisting tool by comparing the performance of participants in recall scores

#### 1) Null and Alternative Hypotheses

$$H_0 : \mu_2 = \mu_1$$

$$H_1 : \mu_2 > \mu_1$$

Where

$\mu_1$  = Mean of recall among participants without assisting tool.

$\mu_2$  = Mean of recall among participants with assisting tool.

#### 2) The Assumption tests

- There is a relationship between recall scores among participants with assisting tools and without assisting tools.

- Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of recall scores difference between assisted and unassisted.

Hypothesis:

$H_0$  : Recall scores difference among participants with assisting tool and without the tool follow a normal distribution.

$H_1$  : Recall scores difference among participants with assisting tool and without the tool do not follow a normal distribution.

Table 51: Result of Test of Normality of recall scores difference between among participants with assisting tool and without the tool.

	Shapiro-wilk	
	W-test statistic	P-value
Assisted - Unassisted	0.94	0.57
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.94$ ,  $p = 0.57$ , which indicates that the recall scores both with assisting tool and without assisting tool are normally distributed.

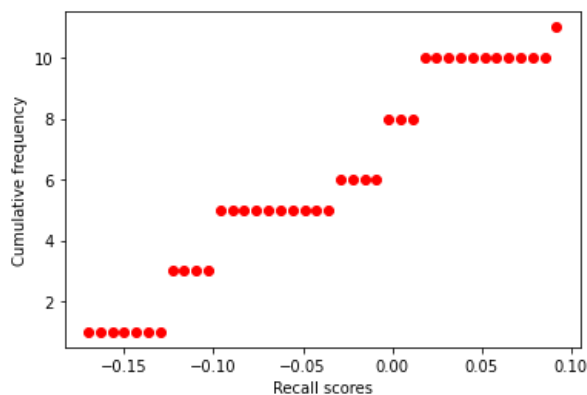


Figure 37: Probability Plots (PP Plot) of recall scores difference (assisted - unassisted).

### 3) Test Statistics

To compare the means for assisted and unassisted, we used **Paired Samples T-Test**, denoted as  $t$ .

Table 52: Result of Paired Samples T-Test between with assisting tool and without assisting tool: recall scores.

Paired t-test				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
0.05	-1.79	-0.04	-0.12	0.03
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one-tailed $p \leq 0.01$ was considered statistically significant).				

### 4) Interval estimates Using T-score with 99.00% CI

Table 53: Result of Interval estimates of recall scores using T-score.

Interval estimates using T-score			
Group	Mean of recall scores	99.00% Confident Interval	
		Lower	Upper
Assisted	88.31	82.33	94.29
Unassisted	92.64	87.74	97.54

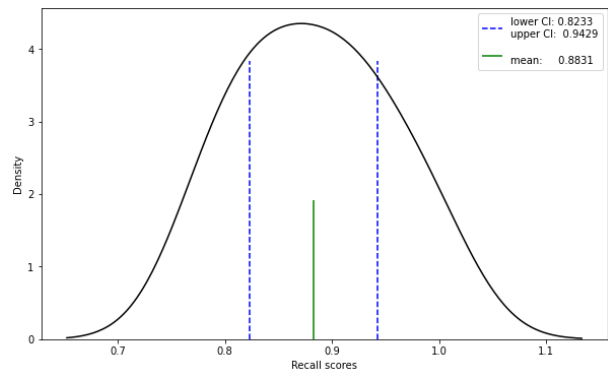


Figure 38: Plot of recall scores among participants with assisting tool, t-statistics - Confidence Level = 99.00%.

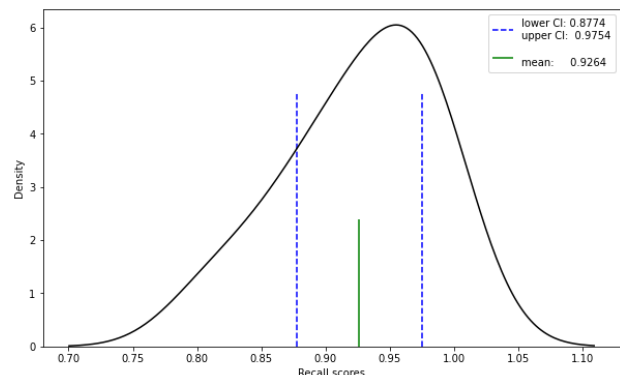


Figure 39: Plot of recall scores among participants without assisting tool, t-statistics - Confidence Level = 99.00%.

## VII. THE PERFORMANCE OF THE PARTICIPANTS BETWEEN THE FIRST ROUND OF EXPERIMENT AND THE SECOND ROUND OF EXPERIMENT (PAGE 21).

We use **Paired Samples T-Test** to compare the accuracy between the first round of the experiment and the second round of the experiment with the participants.

### 7.1 Null and Alternative Hypotheses

$$H_0 : \mu_2 - \mu_1 = 0$$

$$H_1 : \mu_2 - \mu_1 \neq 0$$

Where

$\mu_1$  = Mean of accuracy first round of the experiment.

$\mu_2$  = Mean of accuracy in second round of the experiment.

### 7.2 The Assumption tests

1) There is a relationship of accuracy scores in the rounds of the experiments, between the first session and the second session.

2) Test of Normality: We use the **Shapiro-wilk test** to test normal distribution between the Accuracy scores of 11 participants on the first and the second sessions.

Hypothesis:

$H_0$  : Accuracy scores difference between the first round and the second round of experiment follow normal distribution.

$H_1$  : Accuracy scores difference between the first round and the second round of experiment do not follow normal distribution.

Table 54: Result of Test of Normality of accuracy scores difference between of participants between the first round and the second round of the experiment.

	Shapiro-wilk	
	W-test statistic	P-value
Second experiment – First experiment	0.94	0.55

*\* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$  was considered statistically significant).*

The test is non-significant,  $W = 0.94$ ,  $p = 0.55$ , which indicates that the accuracy scores difference between the first round and the second round of the experiment follow a normal distribution.

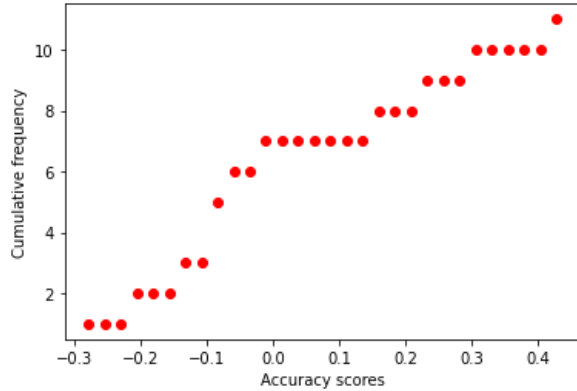


Figure 40: Probability Plots (PP Plot) of accuracy scores difference (second experiment – first experiment).

### 7.3 Test Statistics

To compare the means for the first and the second sessions, we used **Paired Samples T-Test**, denoted as  $t$ .

Table 55: Result of Paired Samples T-Test to compare the means of accuracy in the first round and the second round of the experiment.

Paired t-test				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
0.57	0.59	0.04	-0.17	0.25

*\*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a two-tailed  $p \leq 0.01$  was considered statistically significant).*

### 7.4 Interval estimates Using T-score with 99.00% CI

Table 56: Result of Interval estimates of accuracy scores using T-score.

Interval estimates using T-score			
Group	Mean of accuracy scores	99.00% Confident Interval	
		Lower	Upper
First experiment	68.24	46.89	89.59
Second experiment	72.24	54.71	89.78

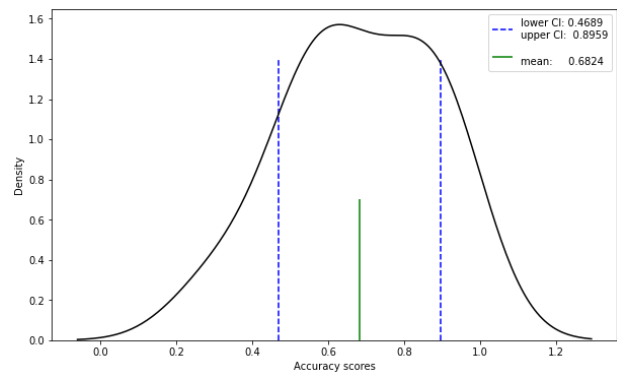


Figure 41: Plot of accuracy scores of participants on the first experiment, t-statistics - Confidence Level = 99.00%.

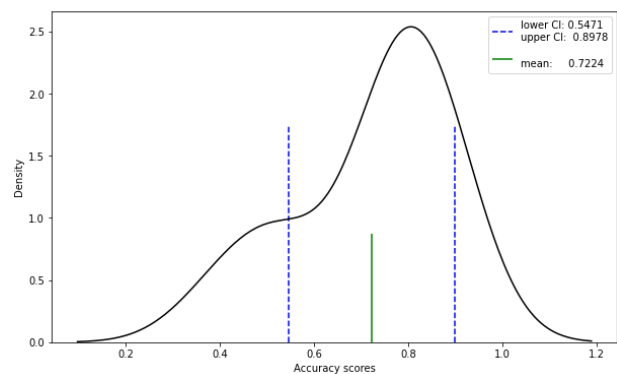


Figure 42: Plot of accuracy scores of participants on the second experiment, t-statistics - Confidence Level = 99.00%.

**VIII. INFLUENCE OF AI SUGGESTION ON PARTICIPANT DECISIONS WHEN ASSISTED/UNASSISTED (PAGE 21).**

We use **Paired Samples T-Test** to compare similarity scores between AI suggestion (prediction) and the final decision of the participants when assisted/unassisted.

**8.1 Null and Alternative Hypotheses**

$$H_0 : \mu_2 = \mu_1$$

$$H_1 : \mu_2 > \mu_1$$

Where

$\mu_1$  = Mean of similarity between AI suggestion and participant decisions without assisting tool.

$\mu_2$  = Mean of similarity between AI suggestion and participant decisions with assisting tool.

**8.2 The Assumption tests**

- 1) There is a relationship of similarity scores between AI suggestion and decision of 11 participants when assisted/unassisted.
- 2) Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution between the similarity scores between AI suggestion and participant decisions when assisted/unassisted.

Hypothesis

$H_0$  : Similarity scores difference between AI suggestion and participant decisions when assisted/unassisted follow a normal distribution.

$H_1$  : Similarity scores difference between AI suggestion and participant decisions when assisted/unassisted do not follow a normal distribution.

Table 57: Result of Test of Normality of similarity scores difference between AI suggestion and participant decisions when assisted/unassisted.

	Shapiro-wilk	
	W-test statistic	P-value
Assisted - Unassisted	0.94	0.49
* 99.00% confidence intervals (99.00% CI) and p-values from testing ( $p \leq 0.01$ was considered statistically significant).		

The test is non-significant,  $W = 0.94$ ,  $p = 0.49$ , which indicates that the similarity scores difference between AI suggestion and participant decisions when assisted/unassisted follow a normal distribution.

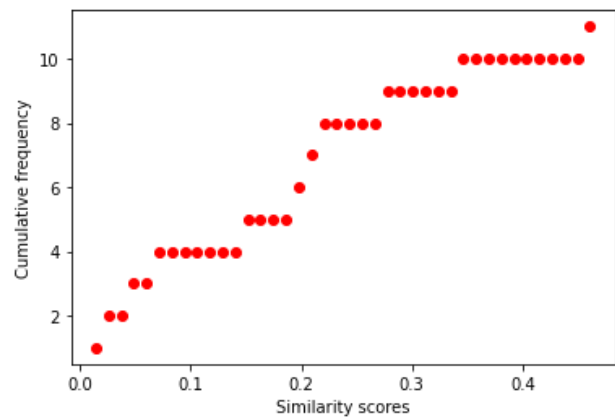


Figure 43: Probability Plots (PP Plot) of similarity scores difference between AI suggestion and participant decisions.

**8.3 Test Statistics**

To compare the means for assisted and unassisted, we used **Paired Samples T-Test**, denoted as t.

Table 58: Result of Paired Samples T-Test to compare the means of similarity between AI suggestion and participant decisions when assisted/unassisted.

Paired t-test				
P - value	t	Mean difference	99.00% Confident Interval of the difference	
			Lower	Upper
$6.90 \times 10^{-4}$	4.38	0.18	0.05	0.32
*With 99.00% confidence intervals (99.00% CI) and p-values from testing (a one-tailed $p \leq 0.01$ was considered statistically significant).				

**8.4 Interval estimates Using T-score with 99.00% CI**

Table 59: Result of Interval estimates of similarity scores using T-score.

Interval estimates using T-score			
Group	Mean of similarity scores	99.00% Confident Interval	
		Lower	Upper
Assisted	77.64	68.18	87.09
Unassisted	58.85	42.32	75.38

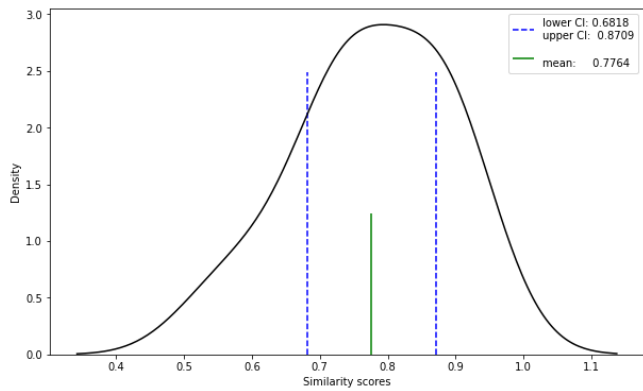


Figure 44: Plot of similarity scores between AI suggestion and participant decisions when assisted, t-statistics - Confidence Level = 99.00%.

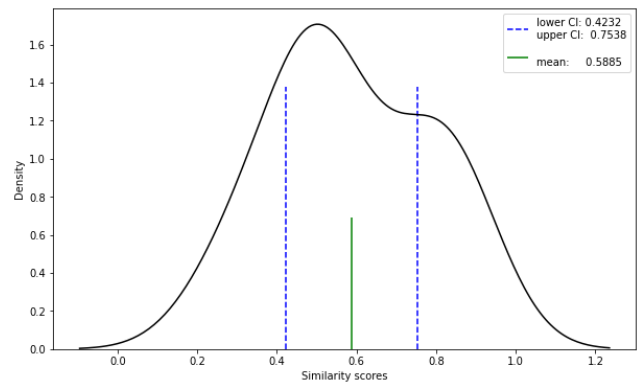


Figure 45: Plot of similarity scores between AI suggestion and participant decisions when unassisted, t-statistics - Confidence Level = 99.00%.

IX. CONFUSION MATRICES OF THE PERFORMANCE OF PARTICIPANTS ON DIFFERENT ABNORMALITIES (PAGE 20).

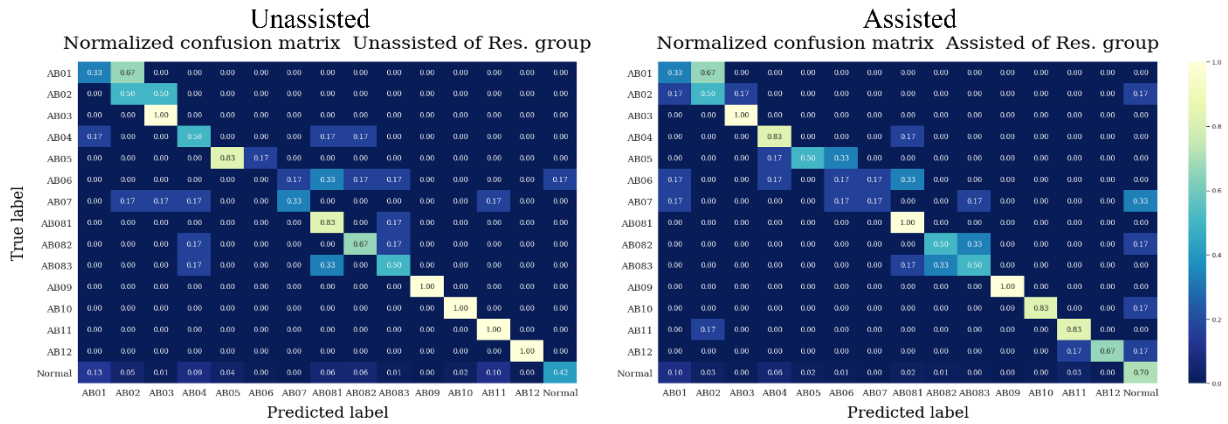


Figure 46: The confusion matrix of the performance of the residence radiologist group without the assisting tool (left) and with assisting tool (right), the numbers are row-wise normalized.

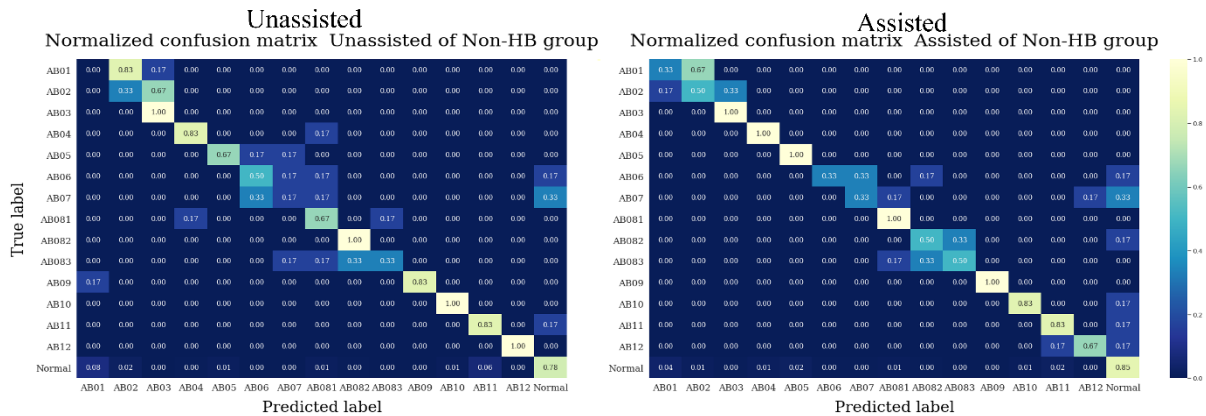


Figure 47: The confusion matrix of the performance of the non-hepatobiliary radiologist group without the assisting tool (left) and with assisting tool (right), the numbers are row-wise normalization.

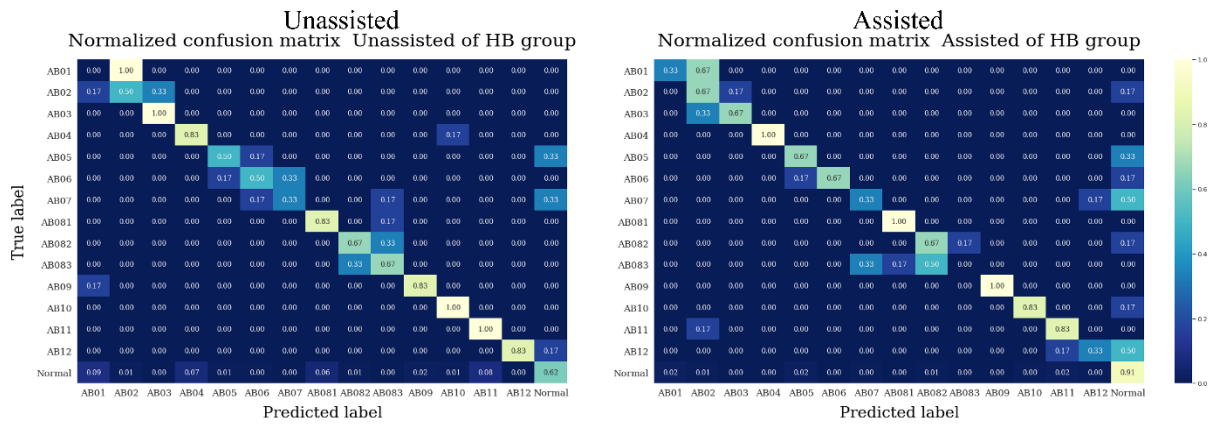


Figure 48: The confusion matrix of the performance of the hepatobiliary radiologist group without the assisting tool (left) and with assisting tool (right), the numbers are row-wise normalization.