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. Performance comparison of different CNN’s with different numbers of parameters and input image sizes (page 9).

Table 1: 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

II. Comparison of the performance between the EfficientNet model and the BiTNet model (page 18).

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

θ_1 = Median of accuracy of the EfficientNet model.

θ_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 2: Result of Test of Normality of accuracy scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.86	0.12
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

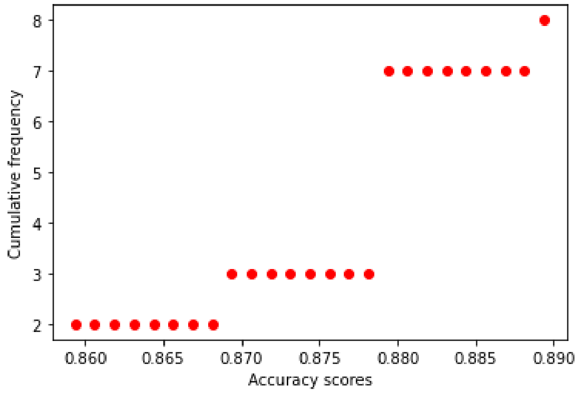

Figure 1: Probability Plots (PP Plots) of accuracy scores of the EfficientNet model.

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 3: Result of Test of Normality of accuracy scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.66	0.00
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

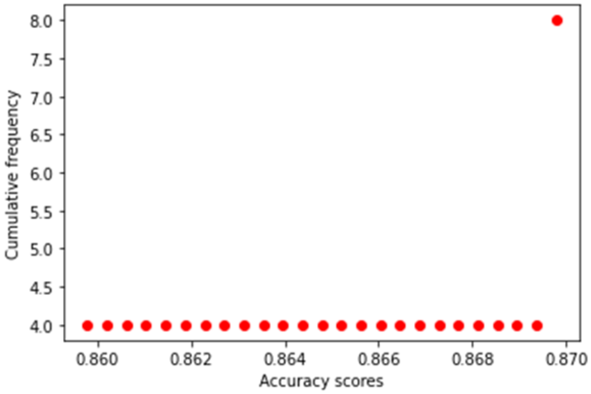

Figure 2: Probability Plots (PP Plots) of accuracy scores of the BiTNet model.

3) Test Statistics

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

Table 4: Result of Mann Whitney U-Test between the EfficientNet model and the BiTNet model: accuracy scores.

Mann-Whitney Test	
	EfficientNet × BiTNet
U	50.00
P-value	5.32×10^{-2}
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 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
μ_1 = Mean of precision of the EfficientNet model.
μ_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 5: Result of Test of Normality of precision scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.89	0.23
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

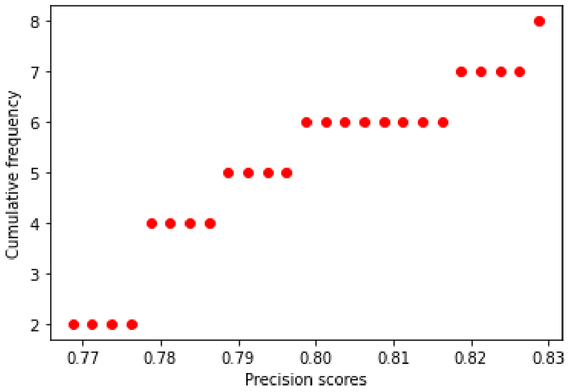

Figure 3: Probability Plots (PP Plots) of precision scores of the EfficientNet model.

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 6: Result of Test of Normality of precision scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.88	0.21

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

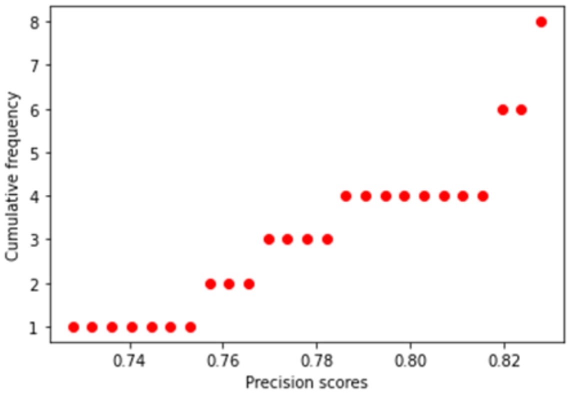

Figure 4: Probability Plots (PP Plots) of precision scores of the BiTNet model.

- 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

σ_1^2 = Variances of the precision of the EfficientNet model.

σ_2^2 = Variances of the precision of the BiTNet model.

Table 7: 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

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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 8: 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	99.90% Confident Interval of the difference	
			Lower	Upper
0.94	-0.08	-1.25×10^{-3}	-0.06	0.06

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*

4) Interval estimates Using T-score with 99.90% CI

Table 9: Result of Interval estimates of precision scores using T-score.

Interval estimates using T-score			
Model	Mean of precision scores	99.90% Confident Interval	
		Lower	Upper
EfficientNet	79.25	74.41	84.09
BiTNet	79.37	71.33	87.42

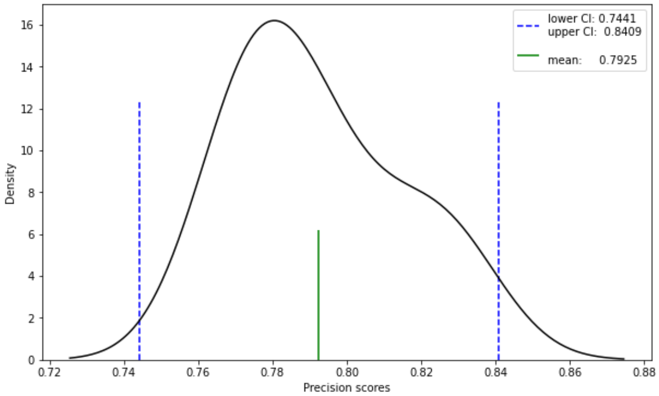

Figure 5: Plot of precision scores of the EfficientNet model, t-statistics - Confidence Level = 99.90%.

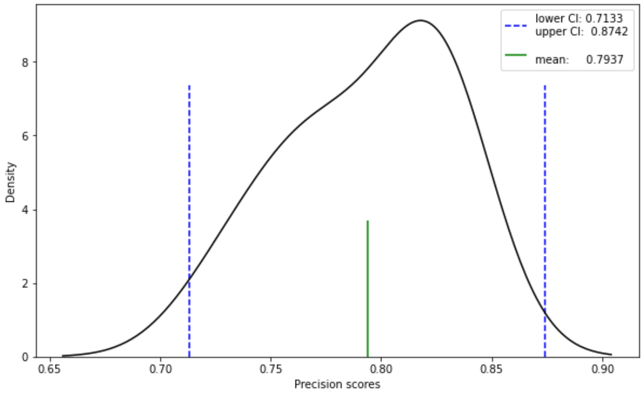

Figure 6: Plot of precision scores of the BiTNet model, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of recall of the EfficientNet model.

μ_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 a follow normal distribution.

Table 10: Result of Test of Normality of recall scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.96	0.85
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

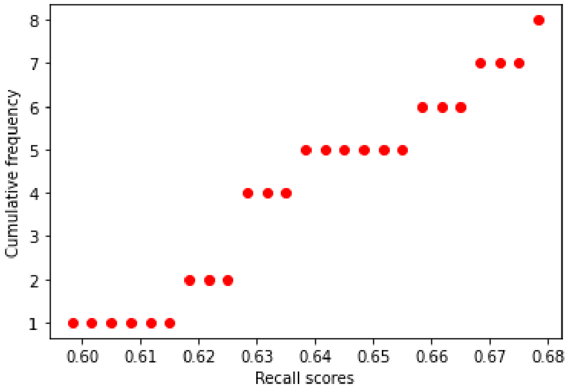

Figure 7: Probability Plots (PP Plot) of recall scores of the EfficientNet model.

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 11: Result of Test of Normality of recall scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.97	0.93
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

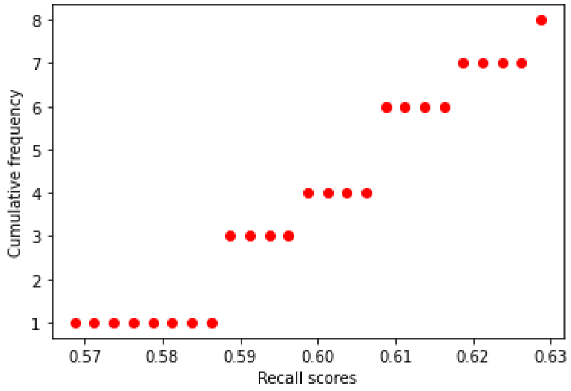

Figure 8: Probability Plots (PP Plot) of recall scores of the BiTNet model.

- 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

σ_1^2 = Variances of the recall of the EfficientNet model.

σ_2^2 = Variances of the recall of the BiTNet model.

Table 12: 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

** 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ 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 13: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: recall scores.

Two sample t-test with equal variance				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
5.07×10^{-3}	3.32	0.04	-9.60×10^{-3}	0.09

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed $p \leq 0.001$ was considered statistically significant).*

4) Interval estimates Using T-score with 99.90% CI

Table 14: Result of Interval estimates of recall scores using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
Model	Mean of recall scores	Lower	Upper
EfficientNet	0.64	0.60	0.70
BiTNet	0.60	0.56	0.64

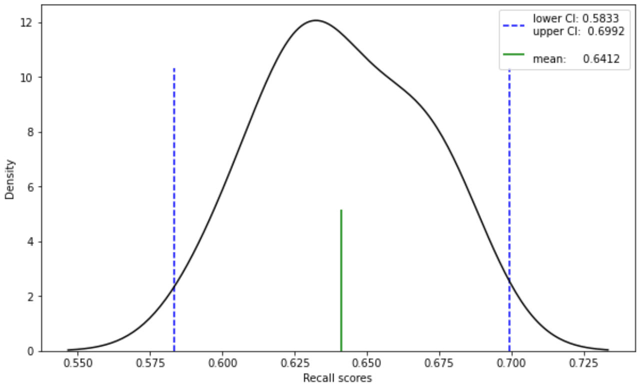

Figure 9: Plot of recall scores of the EfficientNet model, t-statistics - Confidence Level = 99.90%.

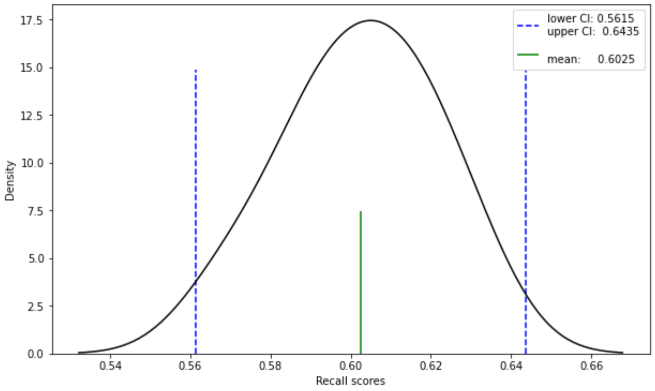

Figure 10: Plot of recall scores of the BiTNet model, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of the accuracy of the EfficientNet model.

μ_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 15: Result of Test of Normality of accuracy scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.83	0.05
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

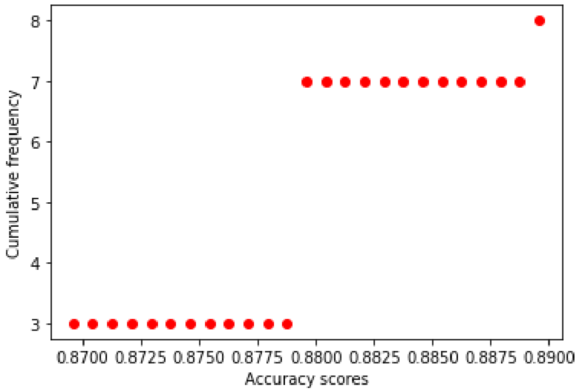

Figure 11: Probability Plots (PP Plots) of accuracy scores of the EfficientNet model.

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 16: Result of Test of Normality of accuracy scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.80	0.03
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

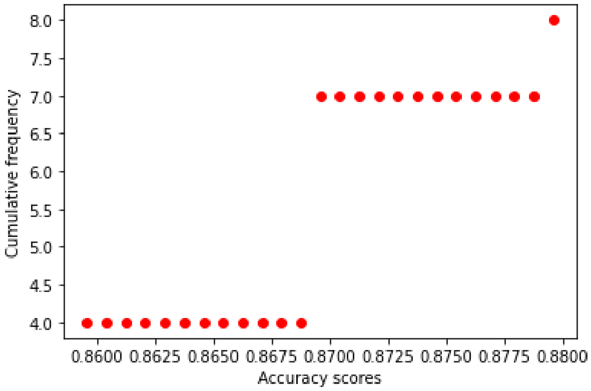

Figure 12: Probability Plots (PP Plots) of accuracy scores of the BiTNet model.

- 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

σ_1^2 = Variances of the accuracy of the EfficientNet model.

σ_2^2 = Variances of the accuracy of the BiTNet model.

Table 17: 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
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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 18: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: accuracy scores.

Two sample t-test with equal variance				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
0.01	3.10	0.01	-3.77×10^{-3}	0.03
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*				

4) Interval estimates Using T-score with 99.90% CI

Table 19: Result of Interval estimates of accuracy scores using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
Model	Mean of accuracy scores	Lower	Upper
EfficientNet	87.75	86.23	89.27
BiTNet	86.63	85.03	88.22

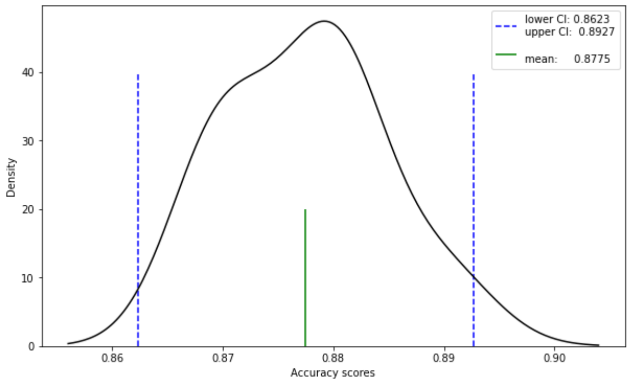

Figure 13: Plot of accuracy scores of the EfficientNet model, t-statistics - Confidence Level = 99.90%.

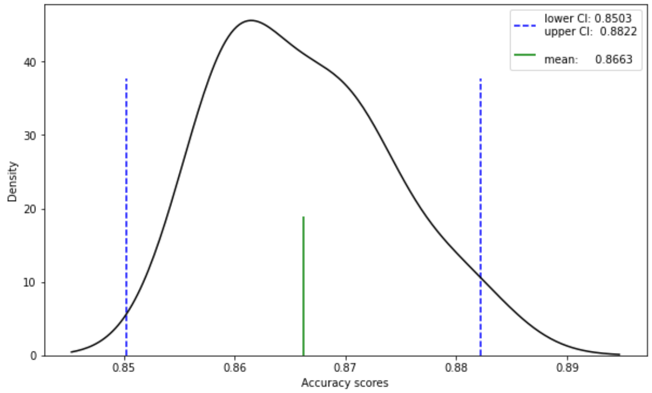

Figure 14: Plot of accuracy scores of the BiTNet model, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of precision of the EfficientNet model.

μ_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 20: Result of Test of Normality of precision scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.87	0.15
* 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ 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.

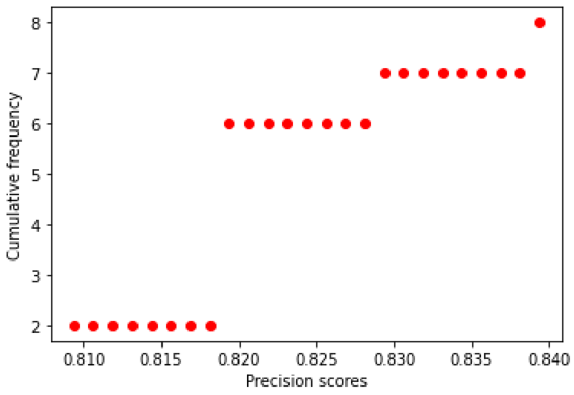

Figure 15: Probability Plots (PP Plot) of precision scores of the EfficientNet model.

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 21: Result of Test of Normality of precision scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.87	0.15
* 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ 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.

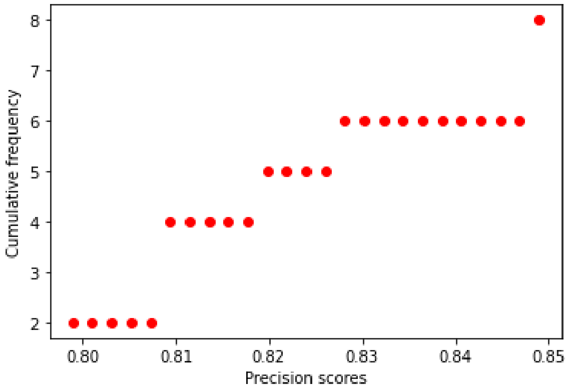

Figure 16: Probability Plots (PP Plots) of precision scores of the BiTNet model.

- 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

σ_1^2 = Variances of the precision of the EfficientNet model.

σ_2^2 = Variances of the precision of the BiTNet model.

Table 22: 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

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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 23: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: precision scores.

Two sample t-test with equal variance				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
1.00	0.00	0.00	-0.03	0.03

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*

4) Interval estimates Using T-score with 99.90% CI

Table 24: Result of Interval estimates of precision scores using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
Model	Mean of precision scores	Lower	Upper
EfficientNet	82.13	79.99	84.26
BiTNet	82.13	77.76	56.49

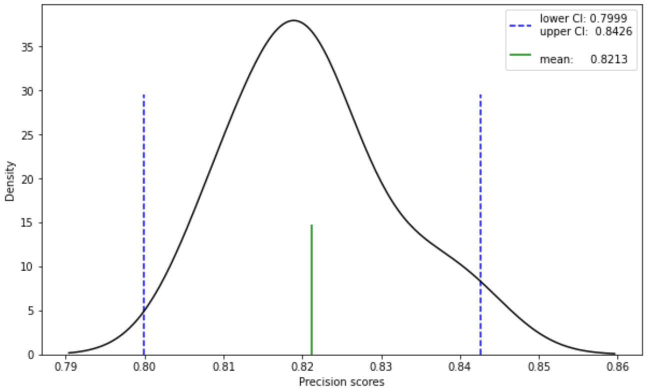

Figure 17: Plot of precision scores of the EfficientNet model, t-statistics - Confidence Level = 99.90%.

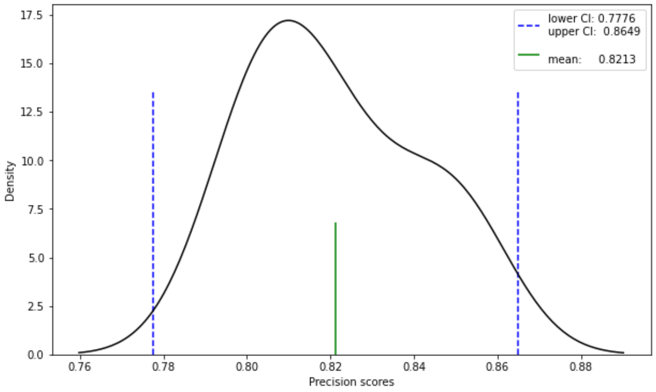

Figure 18: Plot of precision scores of the BiTNet model, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of recall of the EfficientNet model.

μ_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 25: Result of Test of Normality of recall scores of the EfficientNet model.

	Shapiro-wilk	
	W-test statistic	P-value
EfficientNet	0.98	0.96
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

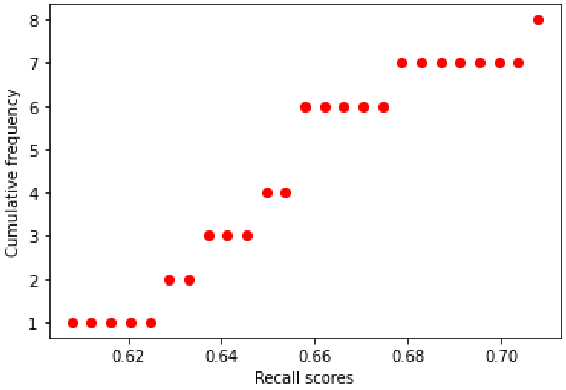

Figure 19: Probability Plots (PP Plot) of recall scores of the EfficientNet model.

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 26: Result of Test of Normality of recall scores of the BiTNet model.

	Shapiro-wilk	
	W-test statistic	P-value
BiTNet	0.95	0.75
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

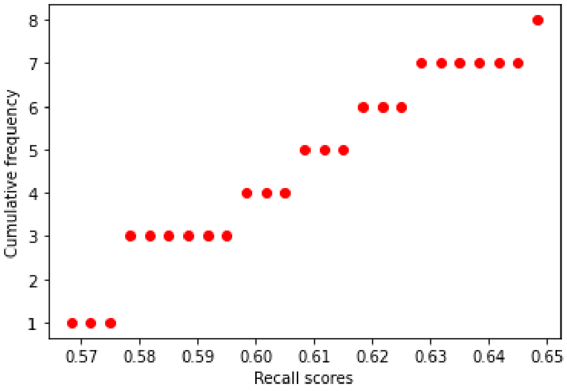

Figure 20: Probability Plots (PP Plot) of recall scores of the BiTNet model.

- 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

σ_1^2 = Variances of the recall of the EfficientNet model.

σ_2^2 = Variances of the recall of the BiTNet model.

Table 27: 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	0.76×10^{-30}	1.0
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*		

The test is non-significant, F = 0.76×10^{-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 28: Result of Independent Samples T-Test between the EfficientNet model and the BiTNet model: recall scores.

Two sample t-test with equal variance				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
4.20×10^{-3}	3.42	0.05	-0.01	0.11
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*				

4) Interval estimates Using T-score with 99.90% CI

Table 29: 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	58.90	72.10
BiTNet	60.50	54.53	66.47

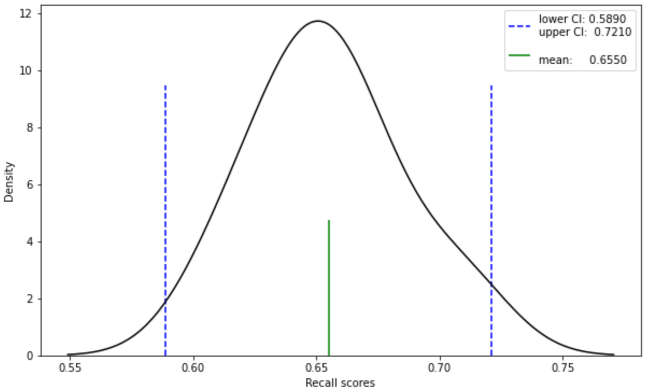

Figure 21: Plot of recall scores of the EfficientNet model, t-statistics - Confidence Level = 99.90%.

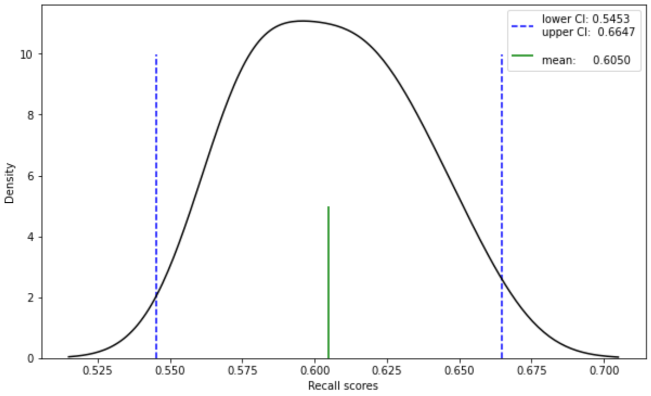

Figure 22: Plot of precision scores of the BiTNet model, t-statistics - Confidence Level = 99.90%.

III. COMPARISON OF THE MEAN DIFFERENCES BETWEEN PREDICTION CONFIDENCE OF THE CORRECT AND INCORRECT GROUPS (PAGE 19).

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.

3.1 Null and Alternative Hypotheses

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Where

μ_1 = Mean of mean difference of prediction confidence of the BiTNet model.

μ_2 = Mean of mean difference of prediction confidence of the EfficientNet model.

3.2 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.

<u>*The BiTNet model:*</u>

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 30: 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	2.72×10^{-2}

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*

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

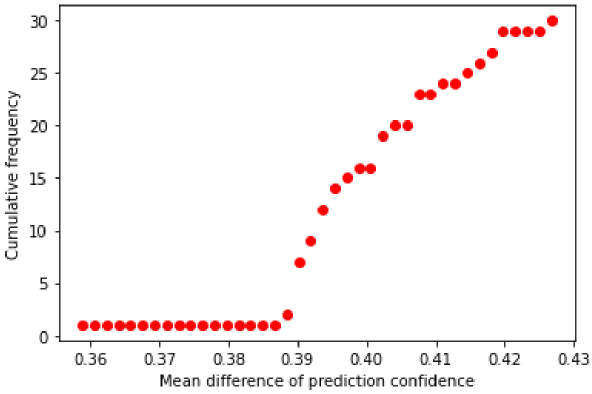

Figure 23: Probability Plots (PP Plots) of the mean difference prediction confidence of the BiTNet model.

<u>*The EfficientNet model:*</u>

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 31: 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	6.27×10^{-2}
* 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ was considered statistically significant).		

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

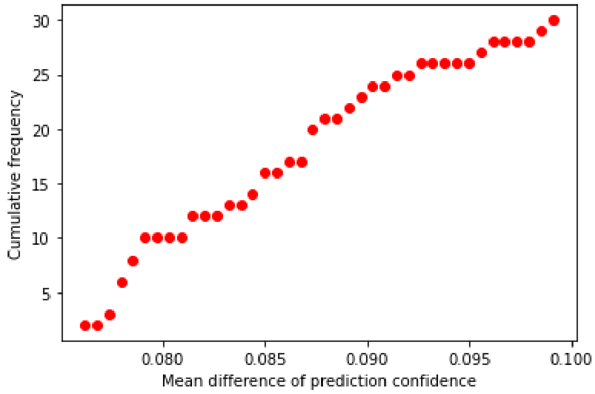

Figure 24: Probability Plots (PP Plots) of the mean difference of prediction confidence of the EfficientNet model.

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

σ_1^2 = Variances of the mean difference of prediction confidence of the BiTNet model.

σ_2^2 = Variances of the mean difference of prediction confidence of the EfficientNet model.

Table 32: 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	5.89×10^{-3}
* 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ was considered statistically significant).		

The test is non-significant, F= 8.17, p = 5.89×10^{-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.

3.3 Test Statistics

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

Table 33: 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				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
2.34×10^{-70}	114.60	31.58	30.62	32.53
*With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed $p \leq 0.001$ was considered statistically significant).				

3.4 Interval estimates Using T-score with 99.90% CI

Table 34: Result of Interval estimates of the mean differences using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
Model	Mean of mean difference	Lower	Upper
BiTNet	40.13	39.29	40.97
EfficientNet	8.55	8.13	8.98

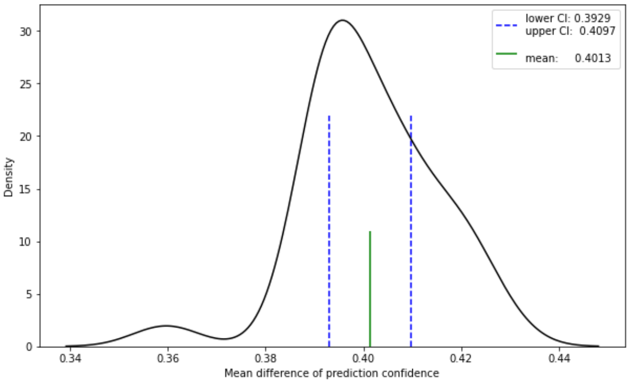

Figure 25: Plot of the mean difference of prediction confidence of the correct and incorrect the BiTNet model, t-statistics - Confidence Level = 99.90%.

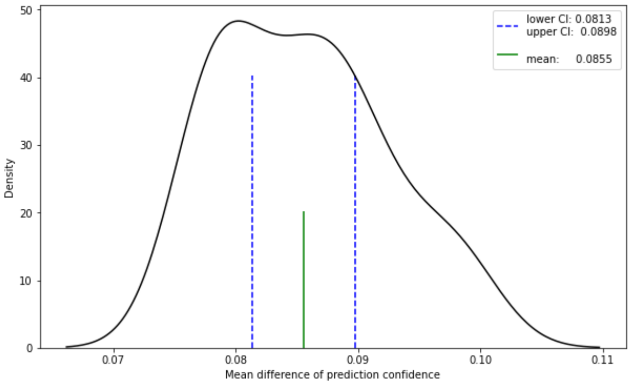

Figure 26: Plot of the mean difference of prediction confidence of the correct and incorrect of the EfficientNet model, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of prediction confidence correct.

μ_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.

<u>*Prediction confidences correct:*</u>

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 35: Result of Test of Normality of prediction confidence correct.

	Shapiro-wilk	
	W-test statistic	P-value
Correct	0.96	0.40
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

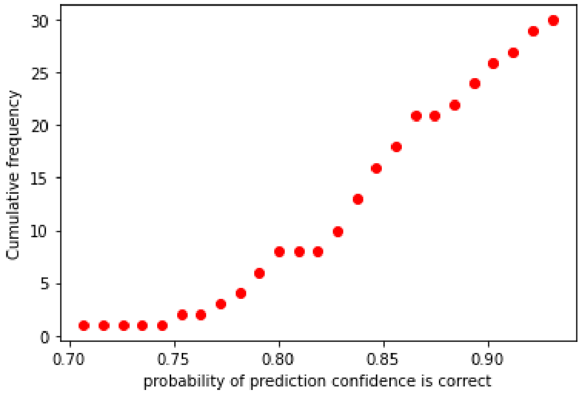

Figure 27: Probability Plots (PP Plot) of prediction confidence is correct.

<u>*Prediction confidences incorrect:*</u>

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 36: Result of Test of Normality of prediction confidence incorrect.

	Shapiro-wilk	
	W-test statistic	P-value
Incorrect	0.98	0.72
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

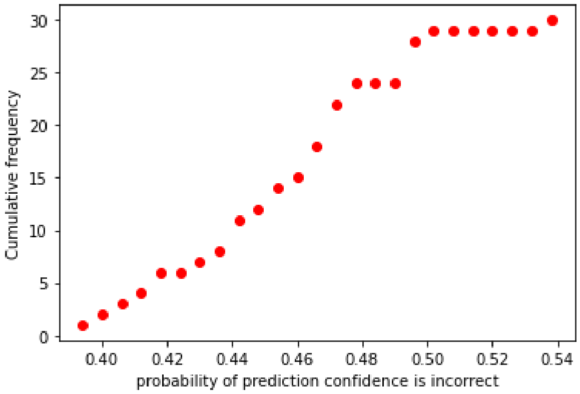

Figure 28: Probability Plots (PP Plot) of prediction confidence incorrect.

- 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

σ_1^2 = Variances of the mean of prediction confidence correct.

σ_2^2 = Variances of the mean of prediction confidence incorrect.

Table 37: 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	4.01×10^{-2}

** 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ 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 38: 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				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
1.0×10^{-39}	33.17	39.06	34.98	43.14

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed $p \leq 0.001$ 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

μ_1 = Mean of prediction confidence correct.

μ_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.

<u>*Prediction confidences correct:*</u>

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 39: Result of Test of Normality of the mean of prediction confidence correct.

	Shapiro-wilk	
	W-test statistic	P-value
Correct	0.87	2.0×10^{-3}

** 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ 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.

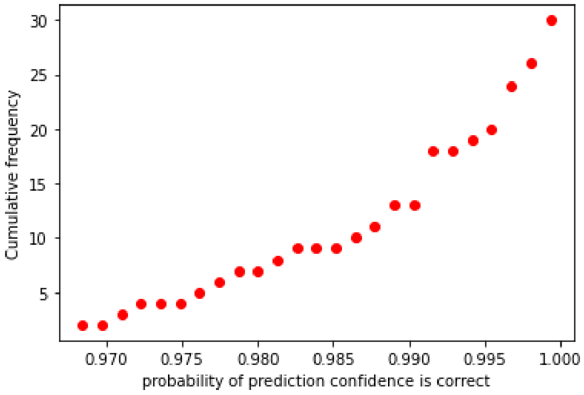

Figure 29: Probability Plots (PP Plot) of the mean prediction confidence correct.

<u>*Prediction confidences incorrect:*</u>

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 40: Result of Test of Normality of the mean prediction confidence incorrect.

	Shapiro-wilk	
	W-test statistic	P-value
Incorrect	0.97	0.81

** 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ 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.

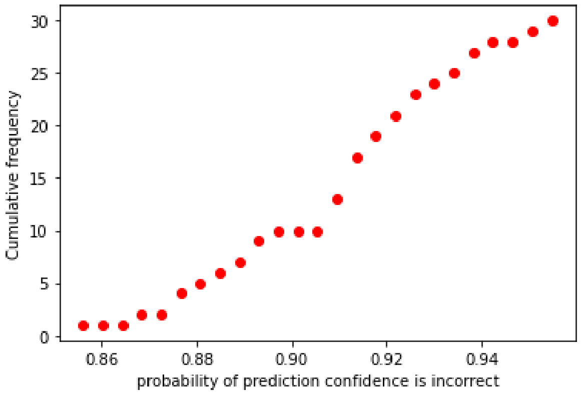

Figure 30: Probability Plots (PP Plot) of the mean prediction confidence incorrect.

- 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

σ_1^2 = Variances of the mean of prediction confidence correct.

σ_2^2 = Variances of the mean of prediction confidence incorrect.

Table 41: 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	2.51×10^{-4}

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*

The test is non-significant, F= 15.23, p = 2.51×10^{-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 42: 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)				
P - value	t	Mean difference	99.90% Confident Interval of the difference	
			Lower	Upper
1.22×10^{-18}	15.74	7.67	5.93	9.41

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).*

IV. PERFORMANCE OF THE BITNET MODEL ON DIFFERENT CLASSES (PAGE 20).

We use the **Independent Samples T-Test** to compare accuracy between the group with more training images and the group with a lesser number of training images.

4.1 Null and Alternative Hypotheses

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Where

μ_1 = Mean of the accuracy of the group with more training images.

μ_2 = Mean of the accuracy of the group with lesser training images.

4.2 The Assumption tests

1) There is no relationship between the group with more training images and the group with a lesser number of training images.

2) Test of Normality: We use the **Shapiro-wilk test** to test the normal distribution of accuracy scores for each group training images.

The group with more training images:

Hypothesis:

H_0 : Accuracy scores of the group with more training images follow a normal distribution.

H_1 : Accuracy scores of the group with more training images does not follow a normal distribution.

Table 43: Result of Test of Normality of accuracy scores of the group with more training images.

	Shapiro-wilk	
	W-test statistic	P-value
The group with more training images	0.94	0.65

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*

The test is non-significant, W= 0.94, p = 0.65 which indicates that the accuracy scores of the group with more training images follow a normal distribution.

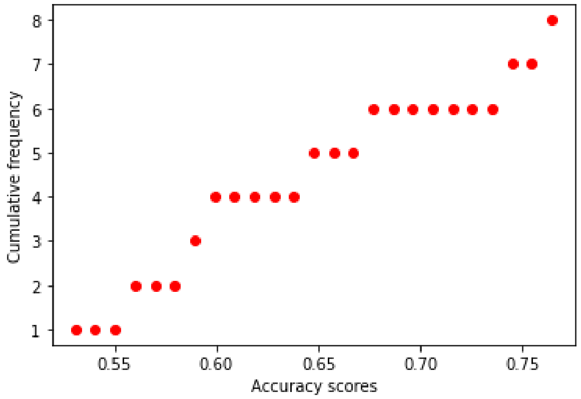

Figure 31: Probability Plots (PP Plots) of accuracy scores of the group with more training images.

The group with a lesser number of training images:

Hypothesis:

H_0 : Accuracy scores of the group with a lesser number of training images follow a normal distribution.

H_1 : Accuracy scores of the group with a lesser number of training images do not follow a normal distribution.

Table 44: Result of Test of Normality of accuracy scores of the group with a lesser number of training images.

	Shapiro-wilk	
	W-test statistic	P-value
The group with lesser training images	0.78	5.29×10^{-2}
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*		

The test is non-significant, W = 0.78, p = 5.29×10^{-2}, which indicates that the accuracy scores of the group with a lesser number of training images follow a normal distribution.

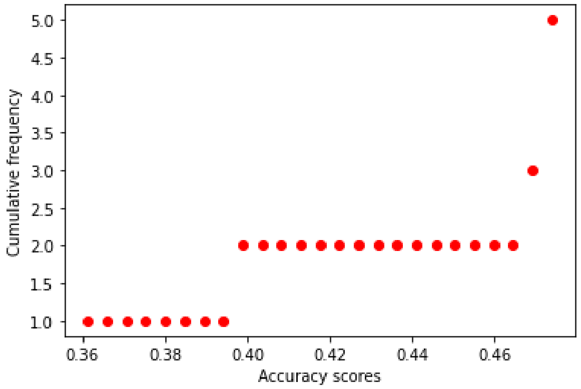

Figure 32: Probability Plots (PP Plots) of accuracy scores of the group with a lesser number of training images.

3) Test of Homogeneity of variances: We use **Levene's Test** to test for the homogeneity of variance of accuracy scores for each group training images.

Hypothesis

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

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

Where

σ_1^2 = Variances of the accuracy of the group with more training images.

σ_2^2 = Variances of the accuracy of the group with a lesser training images.

Table 45: Result of Test for Equality of Variances of accuracy between the group with more training images and the group with lesser number of training images.

	Levene's Test for Equality of Variances	
	F	P-value
Equal variance assumed	1.64	0.23
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*		

The test is non-significant, F = 1.64, p = 0.23, which indicates that the population variances of both group training images are equal. When equal variances are assumed, the calculation uses pooled variances to use Independent Samples T-Test.

4.3 Test Statistics

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

Table 46: Result of the Independent Samples T-Test to compare the means of accuracy score between the group with more training images and the group with lesser training images.

Two sample t-test with equal variance				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
2.63×10^{-4}	4.83	20.68	1.68	39.68
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).*				

4.4 Interval estimates Using T-score with 99.90% CI

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

Interval estimates using T-score			
		99.90% Confident Interval	
Group	Mean of accuracy scores	Lower	Upper
The group with more training images.	64.48	50.05	78.90
The group with lesser training images.	43.79	26.85	60.74

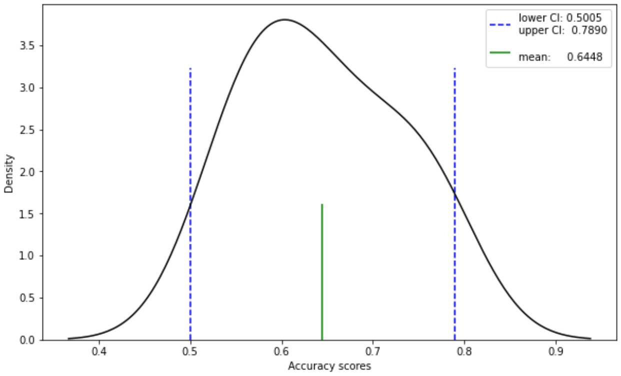

Figure 33: Plot of accuracy scores of the group with more training, t-statistics - Confidence Level = 99.90%.

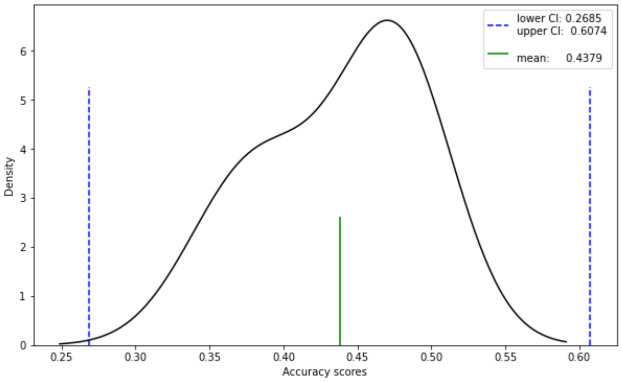

Figure 34: Plot of accuracy scores of the group with lesser training images, t-statistics - Confidence Level = 99.90%.

V. Compares Performance of Participants between Assisted VS Unassisted (Page 24).

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

μ_1 = Mean of accuracy among participants without assisting tools.

μ_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 48: 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
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

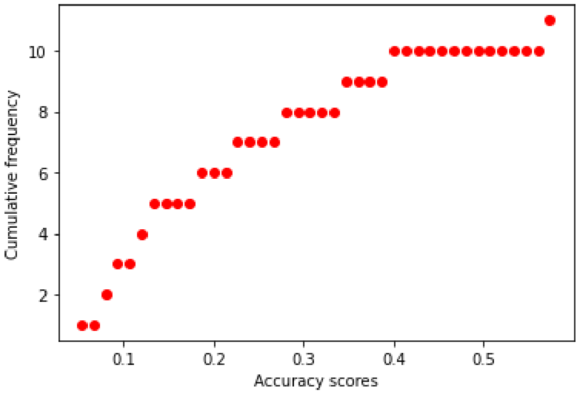

Figure 35: Probability Plots (PP Plot) of accuracy scores difference (assisted - unassisted).

3) Test Statistics

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

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

Paired t-test				
P - value	t	Mean difference	99.90% Confident Interval of the difference	
			Lower	Upper
3.44×10^{-4}	4.83	35.27	1.80	68.75
*With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).				

4) Interval estimates Using T-score with 99.90% CI

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

Interval estimates using T-score			
Group	Mean of accuracy scores	99.90% Confident Interval	
		Lower	Upper
Assisted	73.52	57.01	90.02
Unassisted	50.00	78.57	21.43

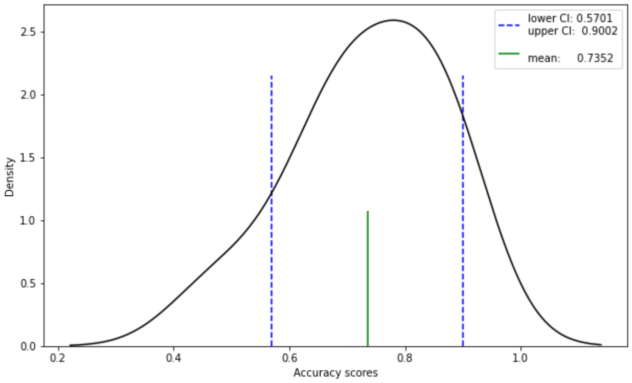

Figure 36: Plot of accuracy scores among participants with assisting tool, t-statistics - Confidence Level = 99.90%.

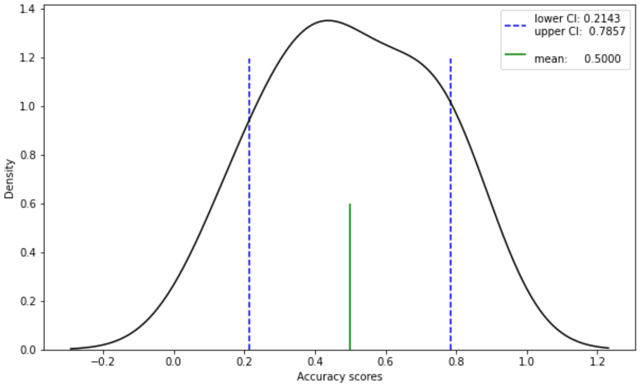

Figure 37: Plot of accuracy scores among participants without assisting tool, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of precision among participants without assisting tool.

μ_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 51: 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
* 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

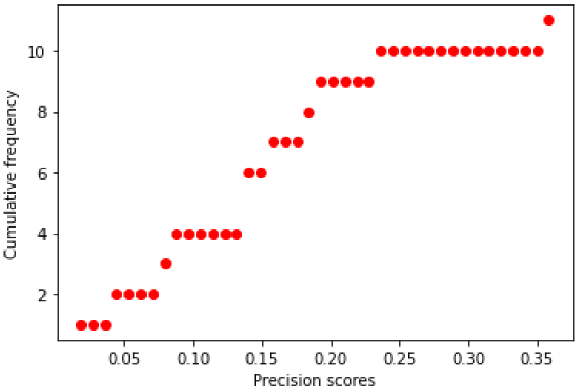

Figure 38: Probability Plots (PP Plot) of precision scores difference (assisted - unassisted).

3) Test Statistics

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

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

Paired t-test				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
1.58×10^{-4}	5.37	15.39	2.24	28.54

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).*

4) Interval estimates Using T-score with 99.90% CI

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

Interval estimates using T-score			
		99.90% Confident Interval	
Group	Mean of precision scores	Lower	Upper
Assisted	61.49	42.88	80.10
Unassisted	46.10	25.81	66.38

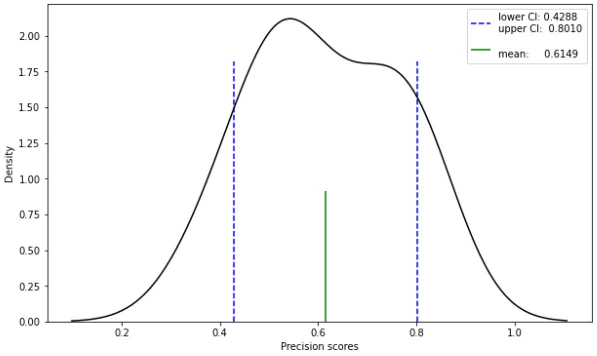

Figure 39: Plot of precision scores among participants with assisting tool, t-statistics - Confidence Level = 99.90%.

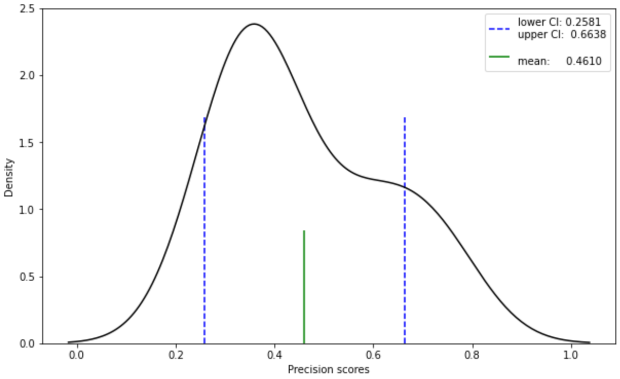

Figure 40: Plot of precision scores among participants without assisting tool, t-statistics - Confidence Level = 99.90%.

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

μ_1 = Mean of recall among participants without assisting tool.

μ_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 54: 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

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

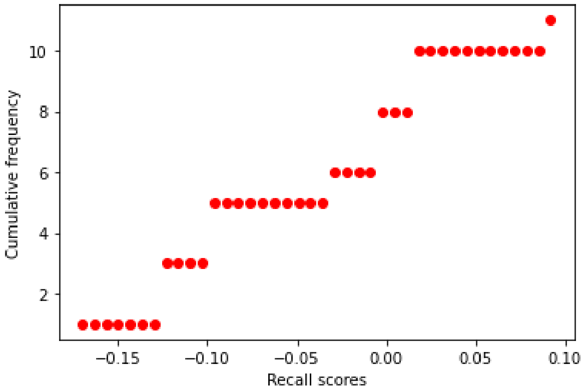

Figure 41: Probability Plots (PP Plot) of recall scores difference (assisted - unassisted).

3) Test Statistics

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

Table 55: Result of Paired Samples T-Test between with assisting tool and without assisting tool: recall scores.

Paired t-test				
		Mean	99.90% Confident Interval of the difference	
P - value	t	difference	Lower	Upper
0.05	-1.79	-4.33	-15.42	6.77

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).*

4) Interval estimates Using T-score with 99.90% CI

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

Interval estimates using T-score			
	Mean of recall	99.90% Confident Interval	
Group	scores	Lower	Upper
Assisted	88.31	79.34	97.28
Unassisted	92.64	85.30	99.98

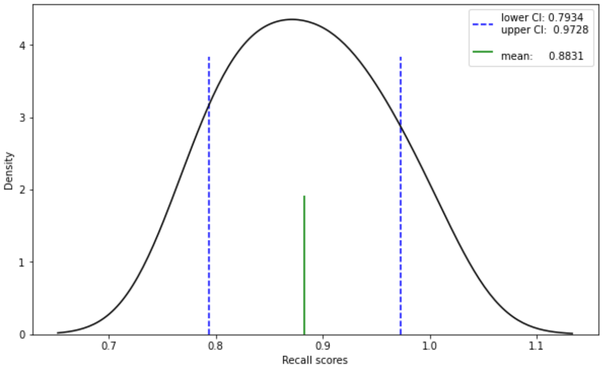

Figure 42: Plot of recall scores among participants with assisting tool, t-statistics - Confidence Level = 99.90%.

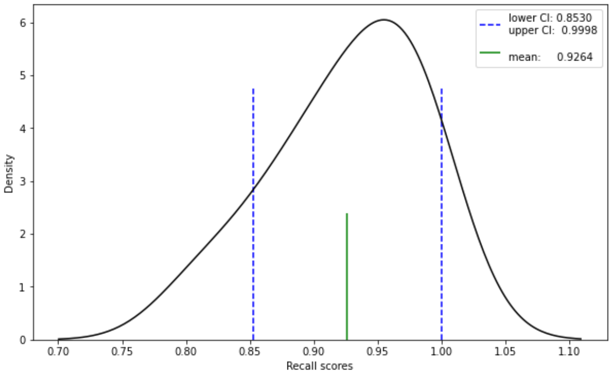

Figure 43: Plot of recall scores among participants without assisting tool, t-statistics - Confidence Level = 99.90%.

VI. The performance of the participants between the first round of experiment and the second round of experiment (page 25).

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.

6.1 Null and Alternative Hypotheses

$H_0 : \mu_2 - \mu_1 = 0$

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

Where

μ_1 = Mean of accuracy first round of the experiment.

μ_2 = Mean of accuracy in second round of the experiment.

6.2 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 57: 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

** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

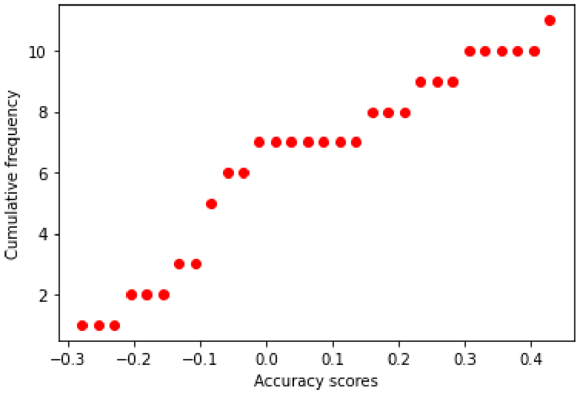

Figure 44: Probability Plots (PP Plot) of accuracy scores difference (second experiment – first experiment).

6.3 Test Statistics

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

Table 58: 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	99.90% Confident Interval of the difference	
			Lower	Upper
0.57	0.59	4.00	27.04	35.04
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*				

6.4 Interval estimates Using T-score with 99.90% CI

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

Interval estimates using T-score			
Group	Mean of accuracy scores	99.90% Confident Interval	
		Lower	Upper
First experiment	68.24	38.14	98.34
Second experiment	72.24	47.52	96.97

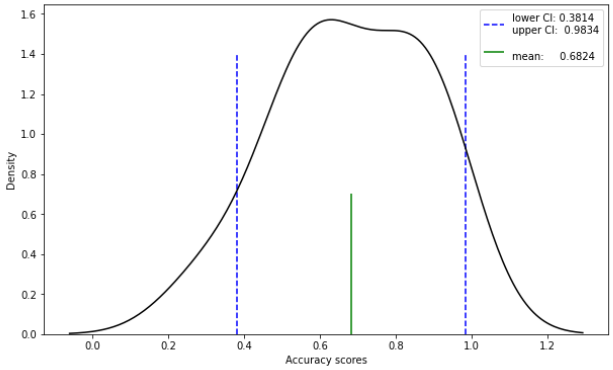

Figure 45: Plot of accuracy scores of participants on the first experiment, t-statistics - Confidence Level = 99.90%.

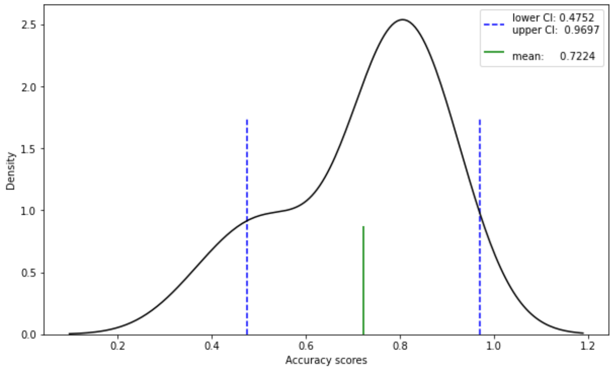

Figure 46: Plot of accuracy scores of participants on the second experiment, t-statistics - Confidence Level = 99.90%.

VII. Influence of AI suggestion on participant decisions when assisted/unassisted (page 26).

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

7.1 Null and Alternative Hypotheses

$H_0 : \mu_2 = \mu_1$

$H_1 : \mu_2 > \mu_1$

Where

μ_1 = Mean of similarity between AI suggestion and participant decisions without assisting tool.

μ_2 = Mean of similarity between AI suggestion and participant decisions with assisting tool.

7.2 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 60: 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
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 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.

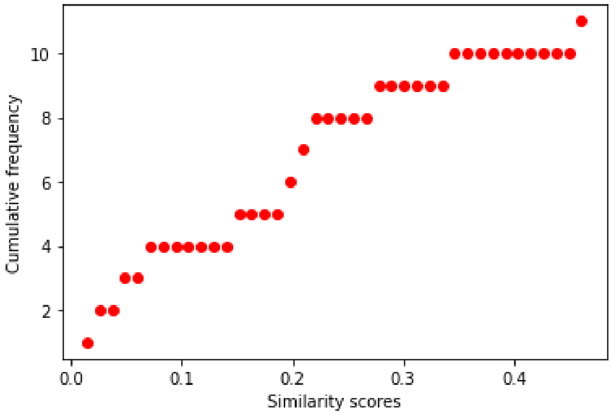

Figure 47: Probability Plots (PP Plot) of similarity scores difference between AI suggestion and participant decisions.

7.3 Test Statistics

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

Table 61: Result of Paired Samples T-Test to compare the means of similarity between AI suggestion and participant decisions when assisted/unassisted.

Paired t-test				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
6.90×10^{-4}	4.38	18.78	-0.89	38.47
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).*				

7.4 Interval estimates Using T-score with 99.90% CI

Table 62: Result of Interval estimates of similarity scores using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
Group	Mean of similarity scores	Lower	Upper
Assisted	77.64	63.47	91.81
Unassisted	58.85	34.07	83.63

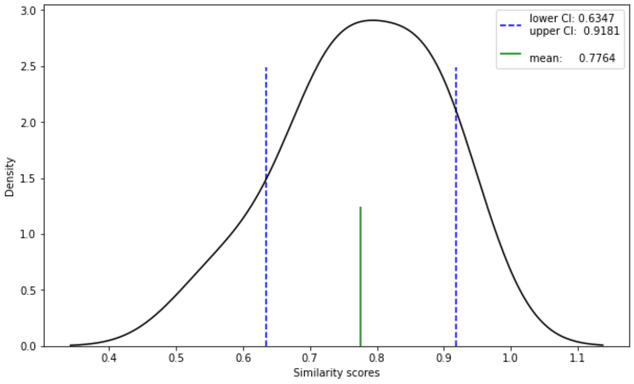

Figure 48: Plot of similarity scores between AI suggestion and participant decisions when assisted, t-statistics - Confidence Level = 99.90%.

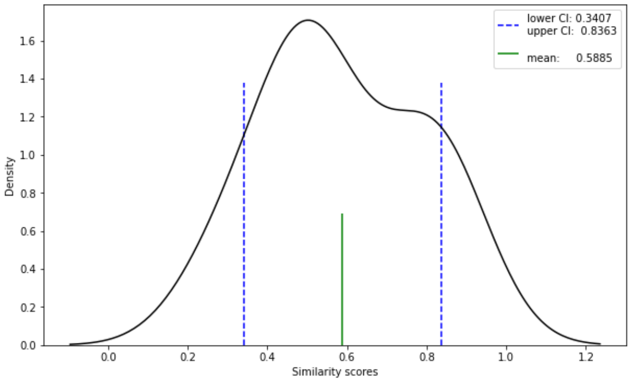

Figure 49: Plot of similarity scores between AI suggestion and participant decisions when unassisted, t-statistics - Confidence Level = 99.90%.

VIII. COMPARE THE RELATIONSHIP BETWEEN HIGH-LOW PREDICTION CONFIDENCE AND THE SIMILARITY OF THE PARTICIPANT ANSWER (PAGE 27).

We use the **Pearson Chi-Square test** to hypothesize by testing the correlation between high-low prediction confidence (confidence ≤ 50 and confidence > 50) and the similarity of the participant's answer to the prediction suggestion suggested.

8.1 Our cross-tabulation table

Table 63: Cross-tabulation between high-low prediction confidence and similarity of the participant answer to the prediction suggested.

	The answer of the participant		
Prediction confidence	Does not have a similar answer	Have similar answer	Total
High	331	956	1,287
Low	181	182	363
Total	**512**	**1,138**	**1,650**

8.2 Null and Alternative Hypotheses

H_0 : Prediction confidence is not associated with the answer of the participant.

H_1 : Prediction confidence is associated with the answer of the participants.

8.3 The Assumption tests

1) Prediction confidence and the answer of the participant were collected independently of each other.
2) Whole expected cell counts greater than 10.
We can be checked by looking at the expected frequency table.

Table 64: Expected frequency table between high-low prediction confidence and similarity of the participant answer to the prediction suggested.

Prediction confidence	The answer of the participant	
	Does not have a similar answer	Have similar answer
High	399.36	887.64
Low	112.64	250.36

8.4 Test Statistics

The test statistic for the **Chi-Square Test of Independence** is denoted χ^2, the research question is the following, is there a relationship between prediction confidence and the answer of participant.

Table 65: Result of Chi-Square Test of Independence between prediction confidence and the answer of the participant.

	Value	P - value
Pearson Chi-Square	76.00	2.84×10^{-18}
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*		

IX. COMPARE THE RELATIONSHIP BETWEEN CORRECT-INCORRECT ROI AND THE PARTICIPANT DECISIONS (PAGE 27).

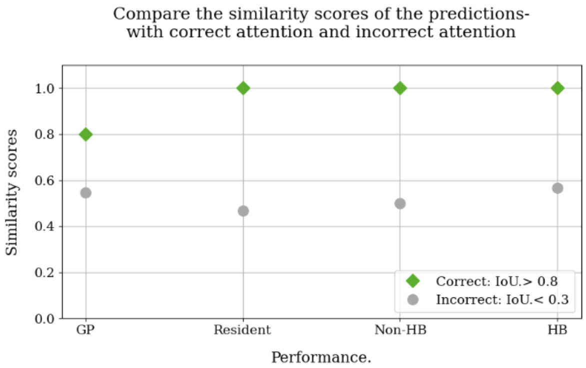

Figure 50: Shows the similarity score of 4 participant groups.

We use the **Pearson Chi-Square test** to hypothesize testing correlation between the decisions when IoU of the GradCam and the ROI are greater than 0.8 (correct) and the decisions when IoU of the GradCam and the ROI are less than 0.3 (incorrect).

9.1 Our cross-tabulation table

Table 66: Cross-tabulation between correct - incorrect IoU and decisions of the participant.

Rating of IOU	The decisions of the participant		Total
	Does not have similar decisions	Have similar decisions	
Correct	2	20	22
Incorrect	96	69	165
Total	**98**	**89**	**187**

9.2 Null and Alternative Hypotheses

H_0 : IoU of the GradCam and the ROI is not associated with the decisions of the participant.

H_1 : IoU of the GradCam and the ROI is associated with the decisions of the participant.

9.3 The Assumption tests

1) IoU values and decisions of the participant were collected independently of each other.
2) Whole expected cell counts greater than 10.
We can be checked by looking at the expected frequency table.

Table 67: Expected frequency table between correct - incorrect IoU and decisions of the participant.

Rating of IOU	The decisions of the participant	
	Does not have similar decisions	Have similar decisions
Correct	11.53	10.47
Incorrect	86.47	78.53

9.4 Test Statistics

The test statistic for the **Chi-Square Test of Independence** is denoted χ^2, the research question is the following, is there a relationship between the IoU of the GradCam and the ROI and the decisions of the participant.

Table 68: Result of Chi-Square Test of Independence between correct - incorrect IoU and decisions of the participant.

	Value	P - value
Pearson Chi-Square	16.84	4.07×10^{-5}
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*		

X. COMPARE THE RELATIONSHIP BETWEEN CORRECT-INCORRECT VIEWING ANGLE PREDICTION AND THE PARTICIPANT DECISIONS (PAGE 28).

We use the **Pearson Chi-Square test** to hypothesize by testing the correlation between the decisions when the viewing angle predictions are correct and the decisions when the viewing angle predictions are incorrect.

10.1 Our cross-tabulation table

Table 69: Cross-tabulation between correct – incorrect viewing angle predictions and decisions of the participant.

Viewing angle predictions	The decisions of the participant		Total
	Does not have similar decisions	Have similar decisions	
Correct	299	779	1,078
Incorrect	196	376	572
Total	**495**	**1,155**	**1,650**

10.2 Null and Alternative Hypotheses

H_0 : Viewing angle predictions are not associated with the decisions of the participant.

H_1 : Viewing angle predictions are associated with the decisions of the participant.

10.3 The Assumption tests

1) Viewing angle predictions and decisions of the participant were collected independently of each other.
2) Whole expected cell counts greater than 10. We can be checked by looking at the expected frequency table.

Table 70: Expected frequency table between correct - incorrect viewing angle predictions and decisions of the participant.

Viewing angle predictions	The decisions of the participant	
	Does not have similar decisions	Have similar decisions
Correct	323.40	754.60
Incorrect	171.60	400.40

10.4 Test Statistics

The test statistic for the **Chi-Square Test of Independence** is denoted χ^2, the research question is the following, is there a relationship between viewing angle predictions and the decisions of the participant.

Table 71: Result of Chi-Square Test of Independence between correct - incorrect viewing angle predictions and decisions of the participant.

	Value	P - value
Pearson Chi-Square	7.28	7.00×10^{-3}
**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a two - tailed p ≤ 0.001 was considered statistically significant).*		

XI. INFLUENCE OF TOP-3 PREDICTION ON PARTICIPANT DECISIONS (PAGE 28).

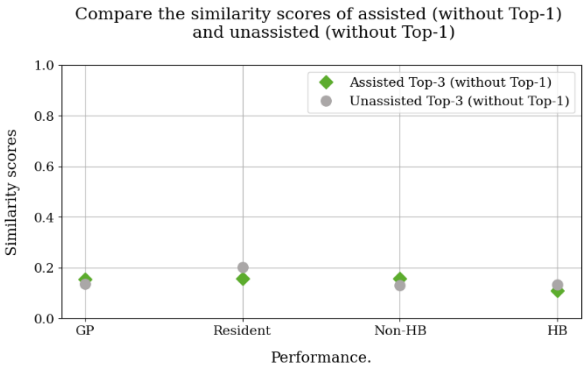

Figure 51: The influence of the top-3 prediction on the participant decisions confirms that the top-3 prediction did not influence the participant's decisions.

We use **Paired Samples T-Test** to compare similarity scores between the participant decisions versus the model top second predictions or the model top third predictions, assisted and unassisted.

11.1 Null and Alternative Hypotheses

$H_0 : \mu_2 = \mu_1$

$H_1 : \mu_2 > \mu_1$

Where

μ_1 = Mean of similarity between top-3 prediction and participant decisions without assisting tool.

μ_2 = Mean of similarity between top-3 prediction and participant decisions with assisting tool.

11.2 The Assumption tests

1) There is a relationship of similarity scores between top-3 prediction 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 top-3 prediction and participant decisions when assisted/unassisted.

Hypothesis:

H_0 *:* Similarity scores difference between top-3 prediction and participant decisions when assisted/unassisted follow a normal distribution.

H_1 : Similarity scores difference between top-3 prediction and participant decisions when assisted/unassisted do not follow a normal distribution.

Table 72: Result of Test of Normality of similarity scores difference between top-3 prediction and participant decisions when assisted/unassisted.

	Shapiro-wilk	
	W-test statistic	P-value
Assisted - Unassisted	0.92	0.31
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*		

The test is non-significant, W = 0.92, p = 0.31, which indicates that the similarity scores difference between top-3 prediction and participant decisions when assisted/unassisted follow a normal distribution.

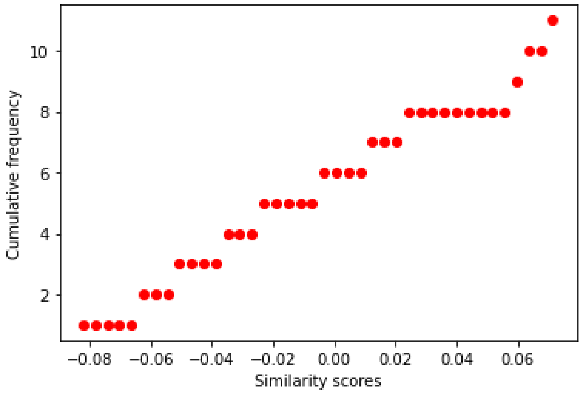

Figure 52: Probability Plots (PP Plot) of similarity scores difference between top-3 prediction and participant decisions (assisted - unassisted).

11.3 Test Statistics

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

Table 73: Result of Paired Samples T-Test to compare the means of similarity between top-3 prediction and participant decisions when assisted/unassisted.

Paired t-test				
			99.90% Confident Interval of the difference	
P - value	t	Mean difference	Lower	Upper
0.50	0.00	-2.52×10^{-16}	-8.61	8.61

**With 99.90% confidence intervals (99.90% CI) and p-values from testing (a one - tailed p ≤ 0.001 was considered statistically significant).*

11.4 Interval estimates Using T-score with 99.90% CI

Table 74: Result of Interval estimates of similarity scores using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
Group	Mean of similarity score	Lower	Upper
Assisted	14.73	8.17	21.28
Unassisted	14.73	10.33	19.13

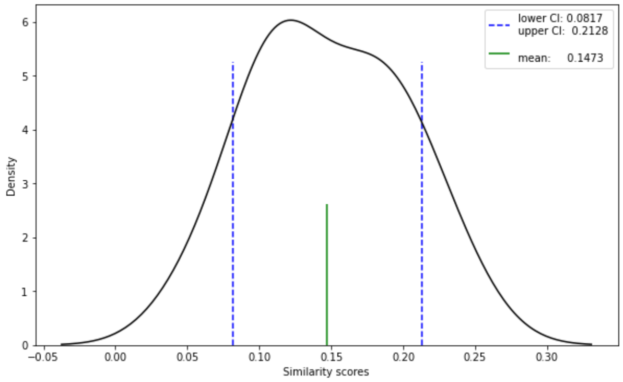

Figure 54: Plot of similarity scores between top-3 prediction and participant decisions when assisted, t-statistics - Confidence Level = 99.90%.

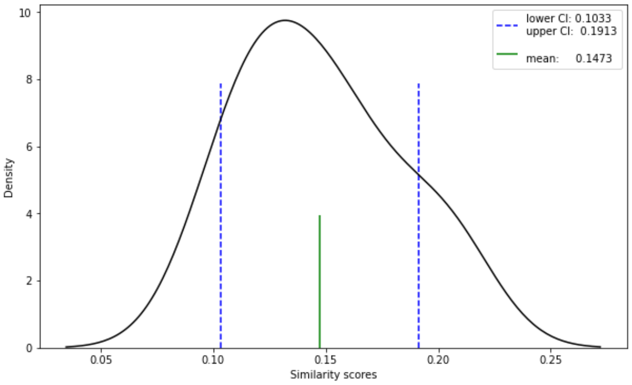

Figure 55: Plot of similarity scores between top-3 prediction and participant decisions when unassisted, t-statistics - Confidence Level = 99.90%.

XII. Confusion Matrices of the performance of participants on different abnormalities (page 24).

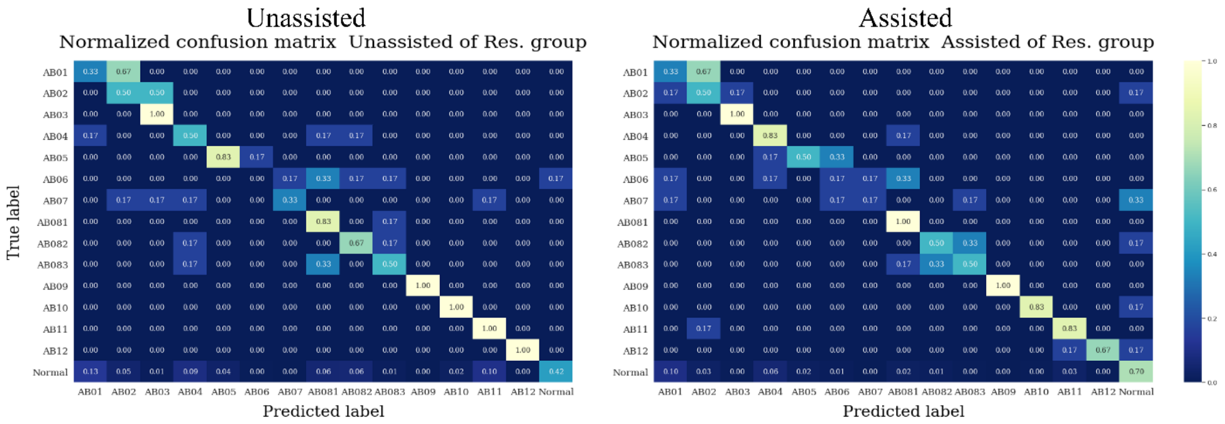

Figure 56: The confusion matrix of the performance of the residence radiologist group without the assisting tool (left) and with assisting tool (right), the numbers are row-wise normalized.

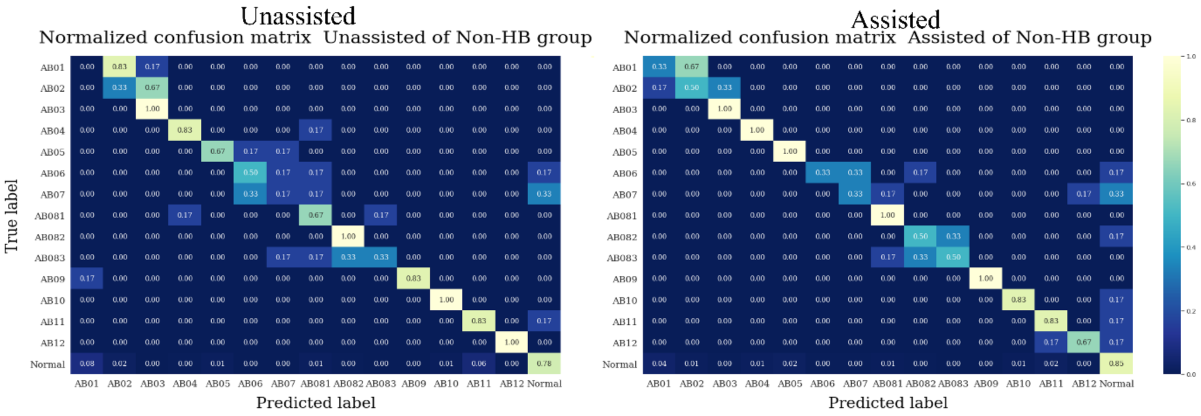

Figure 57: The confusion matrix of the performance of the non-hepatobiliary radiologist group without the assisting tool (left) and with assisting tool (right), the numbers are row-wise normalization.

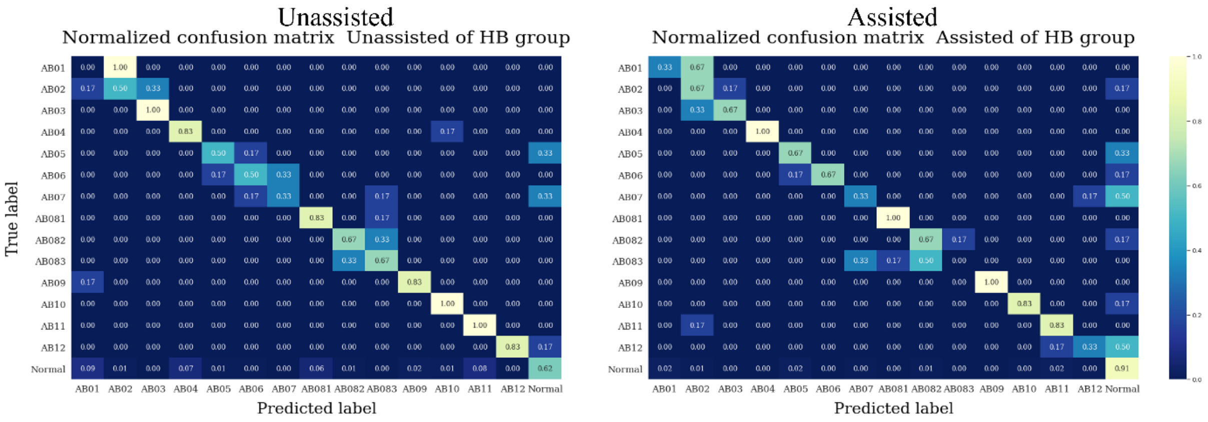

Figure 58: The confusion matrix of the performance of the hepatobiliary radiologist group without the assisting tool (left) and with assisting tool (right), the numbers are row-wise normalization.