Supplementary material for the paper “BiTNet: Deep Hybrid Model for Ultrasonography Image Analysis of Human Biliary Tract and Its Applications”

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

I. COMPARES THE MEAN DIFFERENCES BETWEEN PREDICTION CONFIDENCE OF THE CORRECT AND INCORRECT GROUPS

We are use **Independent Two-sample t Test** to hypothesis testing compares the means of mean difference of prediction confidence of the correct and incorrect groups between BiTNet model and EfficientNet model.

1.1 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 BiTNet model.

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

1.2 The Assumption tests

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

2) Test of Normality: We are use **Shapiro wilk – test** to testing normal distribution of Mean difference of prediction confidence each model.

BiTNet model:

Hypothesis:

H_0 *:* Mean difference of prediction confidence of BiTNet model follows normal distribution.

H_1 *:* Mean difference of prediction confidence of BiTNet model does not follow normal distribution.

Table 1: Result of Test of Normality of prediction confidence of 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 BiTNet model are normally distributed.

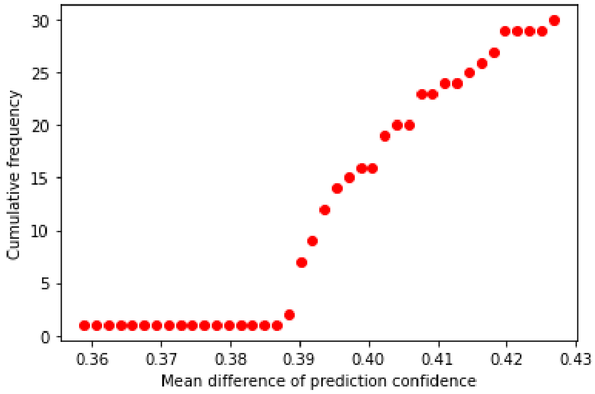

Fig 1: Probability Plots (PP Plot) of prediction confidence of BiTNet model.

EfficientNet model:

Hypothesis:

H_0 *:* Mean difference of prediction confidence of EfficientNet model follows normal distribution.

H_1 *:* Mean difference of prediction confidence of EfficientNet model does not follow normal distribution.

Table 2: Result of Test of Normality of prediction confidence of BiTNet 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 ≤ 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 EfficientNet model are normally distributed.

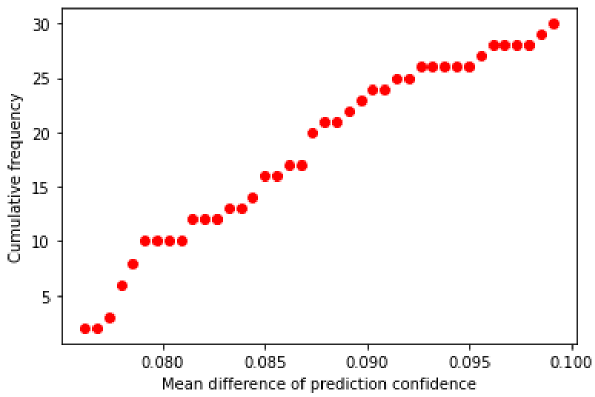

Fig 2: Probability Plots (PP Plot) of prediction confidence of EfficientNet model.

3) Test of Homogeneity of variances

We are use **Levene's Test** to test for the homogeneity of variance of the Mean difference of prediction confidence 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 BiTNet model.

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

Table 3: Result of Test for Equality of Variances of the Mean difference of prediction confidence between BiTNet model and 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 BiTNet model and EfficientNet model are equal. When equal variances are assumed, the calculation uses pooled variances to use Independent Two-sample t Test.

1.3 Test Statistics

The test statistic for this **Independent Samples t Test** is denoted t, for equal variances are assumed.

Table 4: Result of Independent Two-sample t Test between BiTNet model and EfficientNet model.

Two sample t-test with equal variance				
			99.90% Confident Interval of the difference	
P - value	t	Mean Difference	Lower	Upper
2.3×10^{-70}	114.60	31.58	30.62	32.53
* 99.90% confidence intervals (99.90% CI) and p-values from testing ($p \leq 0.001$ was considered statistically significant).				

1.4 Interval estimates Using T-score with 99.90% CI

Table 5: Result of Interval estimates of 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

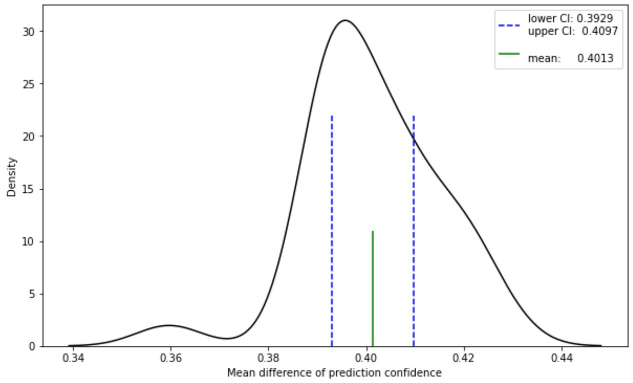

Fig 3: Plot of Mean difference of prediction confidence of the correct and incorrect of BiTNet model, t-statistics - Confidence Level = 99.90%.

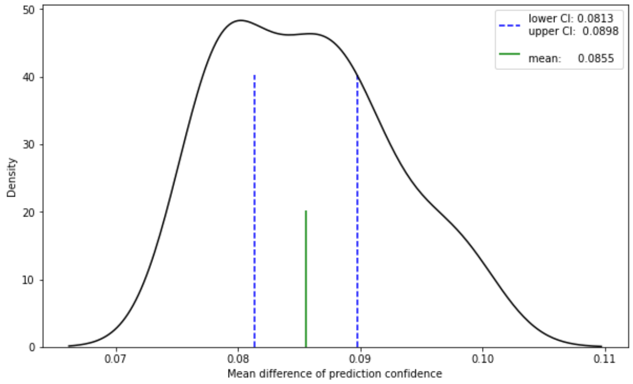

Fig 4: Plot of Mean difference of prediction confidence of the correct and incorrect of EfficientNet model, t-statistics - Confidence Level = 99.90%.

A. Compares the means of prediction confidence between correct and incorrect of 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 are use **Shapiro wilk – test** to testing normal distribution of Mean each prediction confidence.

Prediction confidences correct:

Hypothesis:

H_0 *:* Mean of prediction confidence correct follows Normal Distribution.

H_1 *:* Mean of prediction confidence correct does not follows normal distribution.

Table 6: 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 are normally distributed.

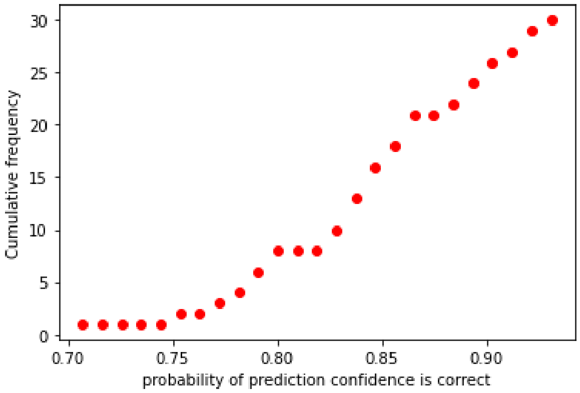

Fig 5: Probability Plots (PP Plot) of prediction confidence is correct.

Prediction confidences incorrect:

Hypothesis:

H_0 *:* Mean of prediction confidence incorrect follows normal distribution.

H_1 *:* Mean of prediction confidence incorrect does not follows normal distribution.

Table 7: 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 are normally distributed.

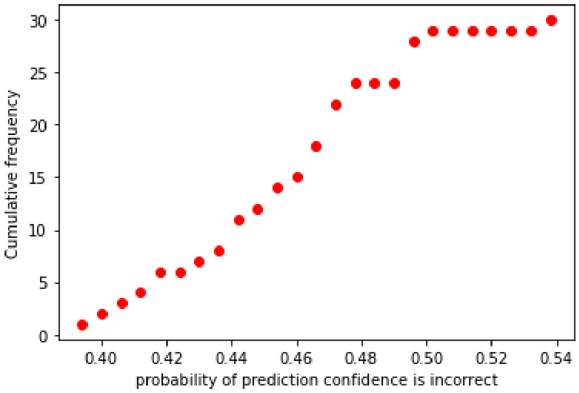

Fig 6: Probability Plots (PP Plot) of prediction confidence incorrect.

- Test of Homogeneity of variances: We are 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 8: 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 ≤ 0.001 was considered statistically significant).*

The test is non-significant, F= 4.41, p = 4.01×10^{-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 Two-sample t Test.

3) Test Statistics

The test statistic for this **Independent Samples t Test** is denoted t, for equal variances are assumed.

Table 9: Result of Independent Two-sample t Test between 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
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*				

B. Compares the means of prediction confidence between correct and incorrect of 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 prediction confidence between correct and incorrect.
- Test of Normality: We are use **Shapiro wilk – test** to testing normal distribution of Mean each prediction confidence.

Prediction confidences correct:

Hypothesis:

H_0 : Mean of prediction confidence correct follows Normal Distribution.

H_1 : Mean of prediction confidence correct does not follows Normal Distribution.

Table 10: Result of Test of Normality 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 ≤ 0.001 was considered statistically significant).*		

The test is non-significant, W = 0.87, p = 2.00×10^{-3}, which indicates that the Mean of confidence correct are normally distributed.

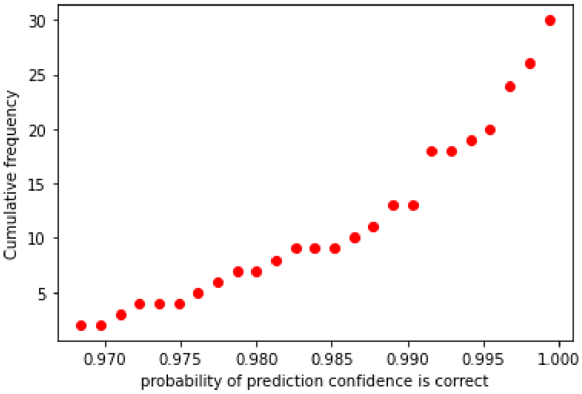

Fig 7: Probability Plots (PP Plot) of prediction confidence correct.

Prediction confidences incorrect:

Hypothesis:

H_0 : Mean of prediction confidence incorrect follows Normal Distribution.

H_1 : Mean of prediction confidence incorrect does not follows Normal Distribution.

Table 11: Result of Test of Normality of 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 ≤ 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 are normally distributed.

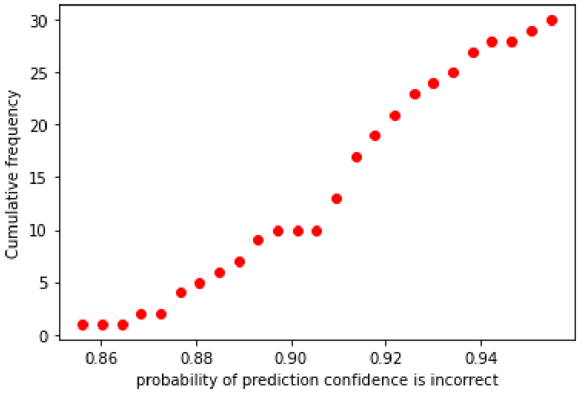

Fig 8: Probability Plots (PP Plot) of prediction confidence incorrect.

- Test of Homogeneity of variances: We are 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 12: 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 not assumed, the calculation utilizes un-pooled variances to use Independent Two-sample t Test.

3) Test Statistics

The test statistic for this **Independent Samples t Test** is denoted t, for equal variances not assumed.

Table 13: Result of Independent Two-sample t Test between 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

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

II. COMPARES PERFORMANCE OF PARTICIPANTS BETWEEN ASSISTED VS UNASSISTED

We are use **Paired Samples t Test** to hypothesis testing comparing performance of participants in accuracy and precision with assisting tool and without assisting tool.

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

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 among participants without assisting tool.

μ_2 = Mean of accuracy among participants with assisting tool.

2) The Assumption tests

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

Hypothesis:

H_0 *:* Accuracy scores difference between among participants with assisting tool and without the tool follows normal distribution.

H_1 *:* Accuracy scores difference between among participants with assisting tool and without the tool does not follows normal distribution.

Table 14: 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 tool and without assisting tool are normally distributed.

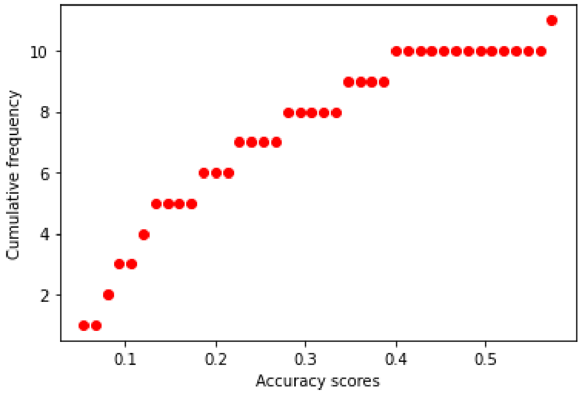

Fig 9: Probability Plots (PP Plot) of accuracy scores difference (assisted - unassisted).

3) Test Statistics

The test statistic for the **Paired Samples t Test**, denoted t, for compare the means for assisted – unassisted.

Table 15: 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

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

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

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

Interval estimates using T-score			
	Mean of Accuracy Scores	99.90% Confident Interval	
group		Lower	Upper
Assisted	73.52	57.01	90.02
Unassisted	50.00	78.57	21.43

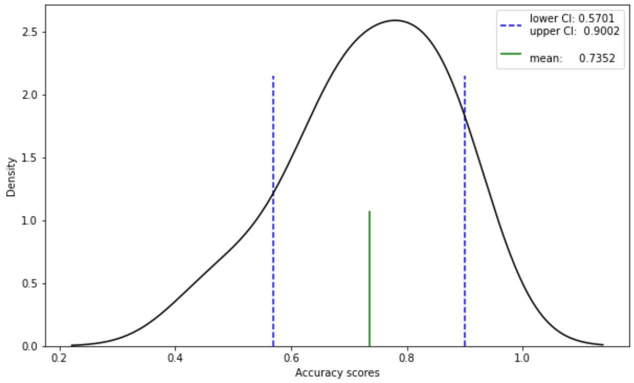

Fig 10: Plot of accuracy scores among participants with assisting tool, t-statistics - Confidence Level = 99.90%.

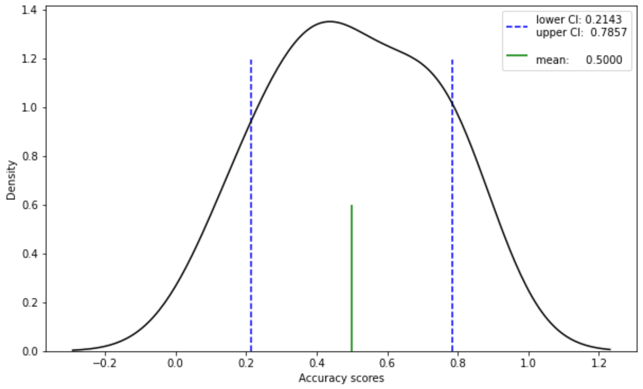

Fig 11: Plot of accuracy scores among participants without assisting tool, t-statistics - Confidence Level = 99.90%.

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

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 precision among participants without assisting tool.

μ_2 = Mean of precision among participants with assisting tool.

2) The Assumption tests

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

Hypothesis:

H_0 : Precision scores difference between among participants with assisting tool and without the tool follows normal distribution.

H_1 : Precision scores difference between among participants with assisting tool and without the tool does not follows normal distribution.

Table 17: Result of Test of Normality of precision scores difference between 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.

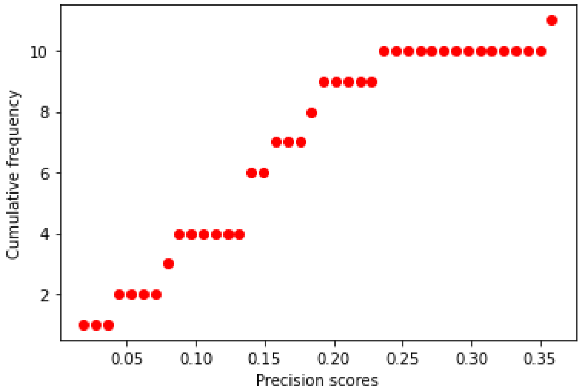

Fig 12: Probability Plots (PP Plot) of precision scores difference (assisted - unassisted).

3) Test Statistics

The test statistic for the **Paired Samples t Test**, denoted t, for compare the means for assisted – unassisted.

Table 18: 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

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

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

Table 19: Result of Interval estimates of precision score 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

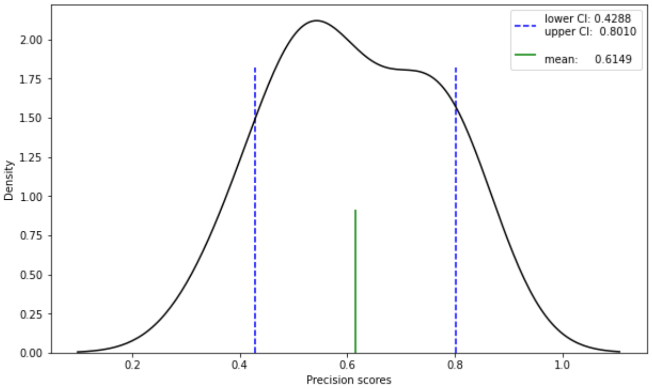

Fig 13: Plot of precision scores among participants with assisting tool, t-statistics - Confidence Level = 99.90%.

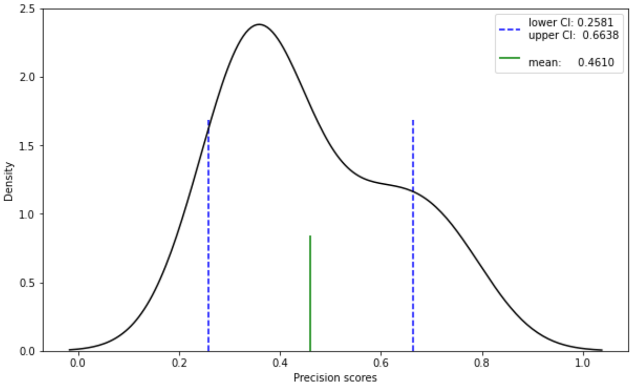

Fig 14: Plot of precision scores among participants without assisting tool, t-statistics - Confidence Level = 99.90%.

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

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 recall among participants without assisting tool.

μ_2 = Mean of recall among participants with assisting tool.

2) The Assumption tests

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

Hypothesis:

H_0 : Recall scores difference between among participants with assisting tool and without the tool follows normal distribution.

H_1 : Recall scores difference between among participants with assisting tool and without the tool does not follows normal distribution.

Table 20: 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.

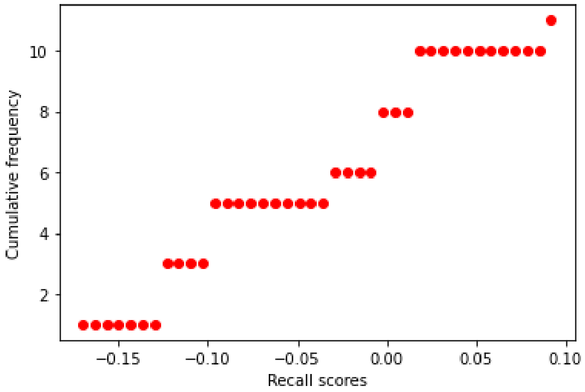

Fig 15: Probability Plots (PP Plot) of recall scores difference (assisted - unassisted).

3) Test Statistics

The test statistic for the **Paired Samples t Test**, denoted t, for compare the means for assisted – unassisted.

Table 21: Result of Paired Samples t Test between with assisting tool and without assisting tool: recall scores.

Paired t-test				
			99.90% Confident Interval of the difference	
P - value	t	Mean Difference	Lower	Upper
0.05	-1.79	-4.33	-15.42	6.77
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*				

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

Table 22: Result of Interval estimates of recall scores using T-score.

Interval estimates using T-score			
		99.90% Confident Interval	
group	Mean of Recall Scores	Lower	Upper
Assisted	88.31	79.34	97.28
Unassisted	92.64	85.30	99.98

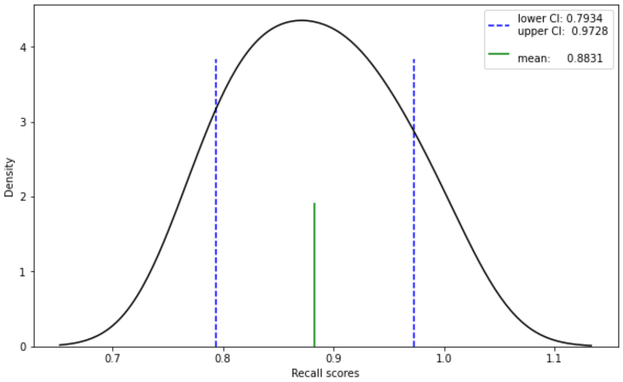

Fig 16: Plot of recall scores among participants with assisting tool, t-statistics - Confidence Level = 99.90%.

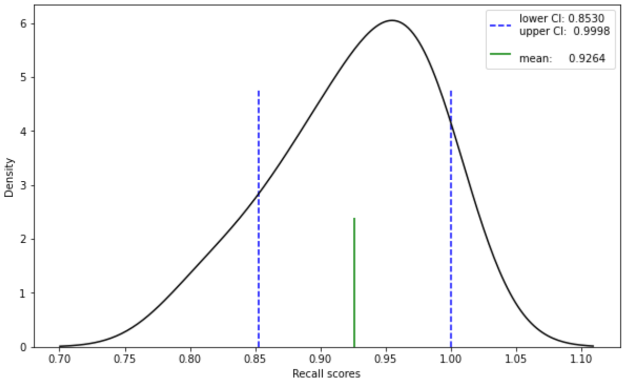

Fig 17: Plot of recall scores among participants without assisting tool, t-statistics - Confidence Level = 99.90%.

III. The performance of the participants between the first round of experiment and the second round of experiment

We are use **Paired Samples t Test** to hypothesis testing comparing accuracy between the first round of experiment and the second round of experiment of the participants.

3.1 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 experiment.

μ_2 = Mean of accuracy second round of experiment.

3.2 The Assumption tests

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

2) Test of Normality: We are use **Shapiro wilk – test** to testing 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 of experiment and the second round of experiment follows normal distribution.

H_1 : Accuracy scores difference between the first round of experiment and the second round of experiment does not follows normal distribution.

Table 23: Result of Test of Normality of accuracy scores difference between of participants between the first round of experiment and the second round.

	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 of experiment and the second round of experiment follows normal distribution.

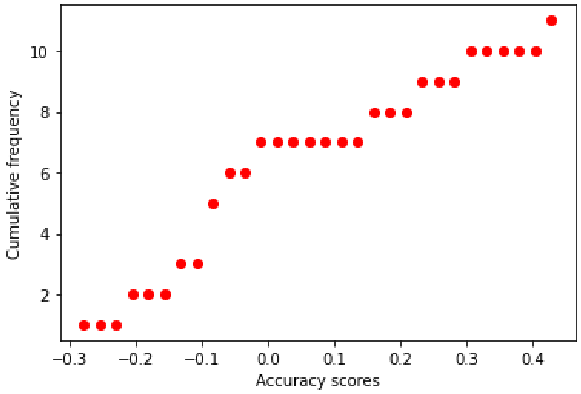

Fig 18: Probability Plots (PP Plot) of accuracy scores difference (Second experiment – First experiment).

3.3 Test Statistics

The test statistic for the **Paired Samples t Test**, denoted t, for compare the means of accuracy for the first and the second sessions.

Table 24: Result of Paired Samples t Test the first round of experiment and the second round of experiment: accuracy scores.

Paired t-test				
			99.90% Confident Interval of the difference	
P - value	t	Mean Difference	Lower	Upper
0.57	0.59	4.00	27.04	35.04

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

3.4 Interval estimates Using T-score with 99.90% CI

Table 25: Result of Interval estimates of accuracy score 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

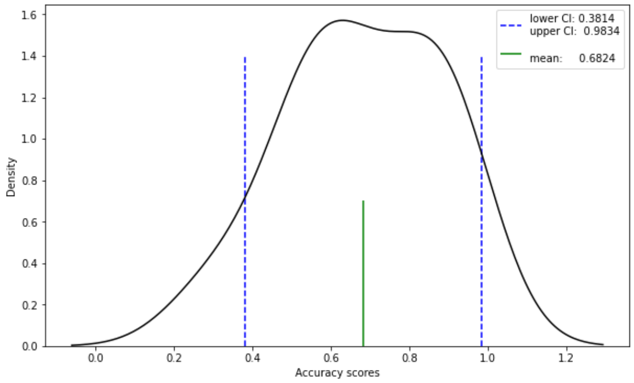

Fig 19: Plot of accuracy scores of participants on the first experiment, t-statistics - Confidence Level = 99.90%.

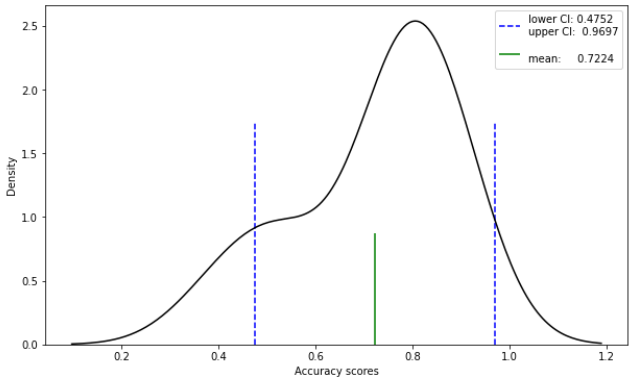

Fig 20: Plot of accuracy scores of participants on the second experiment, t-statistics - Confidence Level = 99.90%.

IV. Influence of AI suggestion on participant decisions when assisted/unassisted

We are use **Paired Samples t Test** to hypothesis testing comparing similarity scores between AI suggestion (prediction) and the final decision of the participants when assisted/unassisted.

4.1 Null and Alternative Hypotheses

$H_0 : \mu_2 - \mu_1 = 0$

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

Where

μ_1 = Mean of similarity between AI suggestion and participant decisions with assisting tool.

μ_2 = Mean of similarity between AI suggestion and participant decisions without assisting tool.

4.2 The Assumption tests

1) There is relationship of similarity scores between AI suggestion and decision of 11 participants when assisted/unassisted.

2) Test of Normality: We are use **Shapiro wilk – test** to testing 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 follows normal distribution.

H_1 : Similarity scores difference between AI suggestion and participant decisions when assisted/unassisted does not follows normal distribution.

Table 26: 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 ($p \leq 0.001$ 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 follows normal distribution.

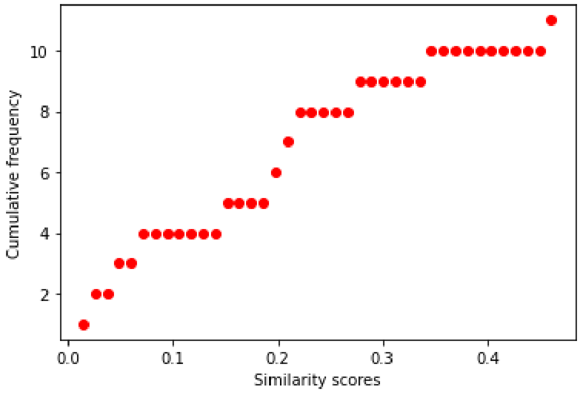

Fig 21: Probability Plots (PP Plot) of similarity scores difference between AI suggestion and participant decisions (assisted - unassisted).

4.3 Test Statistics

The test statistic for the **Paired Samples t Test**, denoted t, for compare the means of similarity for participant decisions when assisted/unassisted.

Table 27: Result of Paired Samples t Test for compare the means of similarity between AI suggestion and participant decisions when assisted/unassisted.

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

4.4 Interval estimates Using T-score with 99.90% CI

Table 28: Result of Interval estimates of similarity score using T-score: AI suggestion.

Interval estimates using T-score			
group	Mean of Similarity Scores	99.90% Confident Interval	
		Lower	Upper
Assisted	77.64	63.47	91.81
Unassisted	58.85	34.07	83.63

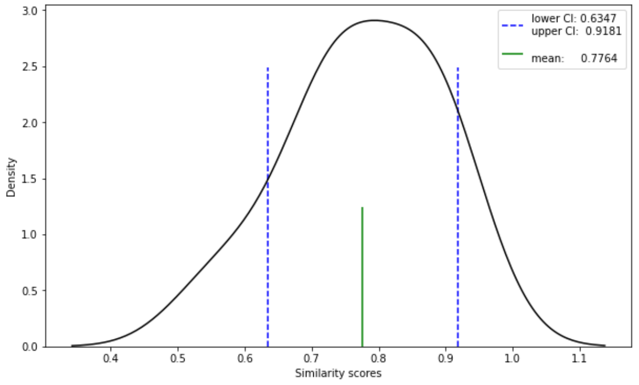

Fig 22: Plot of similarity scores between AI suggestion and participant decisions when assisted, t-statistics - Confidence Level = 99.90%.

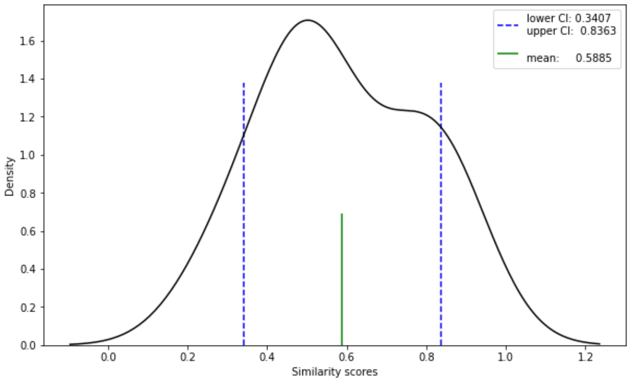

Fig 23: Plot of similarity scores between AI suggestion and participant decisions when unassisted, t-statistics - Confidence Level = 99.90%.

V. COMPARE THE RELATIONSHIP BETWEEN HIGH-LOW PREDICTION CONFIDENCE AND SIMILARITY OF THE PARTICIPANT ANSWER

We are use **Pearson Chi-square test** to hypothesis testing correlation between high-low prediction confidence (confidence ≤ 50 and confidence > 50) and similarity of the participant answer to the prediction suggestion suggested.

5.1 Our cross-tabulation table

Table 29: Cross tabulation between high-low prediction confidence and similarity of the participant answer to the prediction suggested.

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

5.2 Null and Alternative Hypotheses

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

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

5.3 The Assumption tests

1) Prediction confidence and the answer of 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 30: Expected frequency table between high-low prediction confidence and similarity of the participant answer to the prediction suggested.

Prediction confidence	The answer of participant	
	Does not have similar answer	Have similar answer
High	399.36	887.64

Table 30: Continued

Prediction confidence	The answer of participant	
	Does not have similar answer	Have similar answer
Low	112.64	250.36

5.4 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 31: Result of Chi-Square Test of Independence between prediction confidence and the answer of participant.

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

VI. COMPARE THE RELATIONSHIP BETWEEN CORRECT-INCORRECT INTERSECTION OVER UNION (IOU) AND THE PARTICIPANT DECISIONS.

We are use **Pearson Chi-square test** to hypothesis 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).

6.1 Our cross-tabulation table

Table 32: Cross tabulation between correct - incorrect IoU and decisions of the participant.

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

6.2 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.

6.3 The Assumption tests

1) IoU value 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 33: Expected frequency table between correct - incorrect IoU and decisions of the participant.

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

6.4 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 IoU of the GradCam and the ROI and the decisions of the participant.

Table 34: 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}
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*		

VII. COMPARE THE RELATIONSHIP BETWEEN CORRECT-INCORRECT VIEWING ANGLE PREDICTION AND THE PARTICIPANT DECISIONS.

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

7.1 Our cross-tabulation table

Table 35: Cross tabulation between correct – incorrect viewing angle predictions and decisions of the participant.

Viewing angle predictions	The decisions of 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**

7.2 Null and Alternative Hypotheses

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

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

7.3 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 36: Expected frequency table between correct - incorrect viewing angle predictions and decisions of the participant.

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

7.4 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 37: 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}
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*		

VIII. Influence of Top-3 Prediction on Participant Decisions When Assisted/Unassisted

We are use **Paired Samples t Test** to hypothesis testing comparing similarity scores between the participant decisions versus the model top second predictions and the model top third predictions, assisted and unassisted.

8.1 Null and Alternative Hypotheses

$H_0 : \mu_2 - \mu_1 = 0$

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

Where

μ_1 = Mean of similarity between top-3 prediction and participant decisions with assisting tool.

μ_2 = Mean of similarity between top-3 prediction and participant decisions without assisting tool.

8.2 The Assumption tests

1) There is relationship of similarity scores between top-3 prediction and decision of 11 participants when assisted/unassisted.

2) Test of Normality: We are use **Shapiro wilk – test** to testing 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 follows normal distribution.

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

Table 38: 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 follows normal distribution.

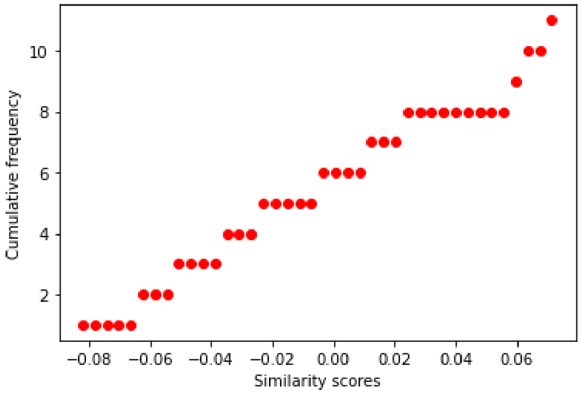

Fig 24: Probability Plots (PP Plot) of similarity scores difference between top-3 prediction and participant decisions (assisted - unassisted).

8.3 Test Statistics

The test statistic for the **Paired Samples t Test**, denoted t, for compare the means of similarity for participant decisions when assisted/unassisted.

Table 39: Result of Paired Samples t Test for compare the means of similarity between top-3 prediction and participant decisions when assisted/unassisted.

Paired t-test				
P - value	t	Mean Difference	99.90% Confident Interval of the difference	
			Lower	Upper
0.50	0.00	-2.52×10^{-16}	-8.61	8.61
** 99.90% confidence intervals (99.90% CI) and p-values from testing (p ≤ 0.001 was considered statistically significant).*				

8.4 Interval estimates Using T-score with 99.90% CI

Table 40: Result of Interval estimates of similarity score using T-score: Top-3 prediction.

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

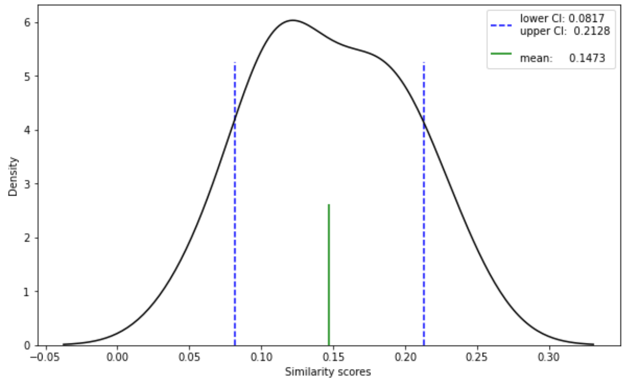

Fig 25: Plot of similarity scores between top-3 prediction and participant decisions when assisted, t-statistics - Confidence Level = 99.90%.

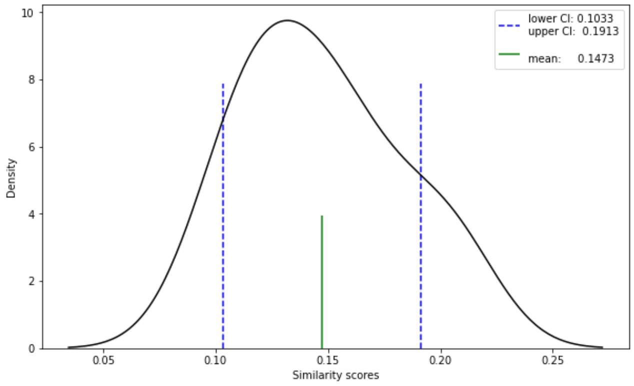

Fig 26: Plot of similarity scores between top-3 prediction and participant decisions when unassisted, t-statistics - Confidence Level = 99.90%.