|  |  |
| --- | --- |
| A chart with different colored squares  Description automatically generated | A chart with different colored squares  Description automatically generated |
| A graph of different colored boxes  Description automatically generated | A chart with yellow and blue squares  Description automatically generated |
| A diagram of a variety of colored boxes  Description automatically generated | A chart with different colored boxes  Description automatically generated |

Figure 1: Performance Metrics of IMCM and ADCM Across Centres C1 to C5

Table 1: Summary statistics of quantitative parameters for different approaches on cross center (Ga dataset)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Method** | **ME** | **MAE** | **RE** | **ARE** | **RMSE** | **PSNR** | **SSI** |
| **Mean ± SD** | **ADCM** | 0.67 ± 1.10 | 2.87 ± 0.75 | -2.17 ± 20.85 | 57.23 ± 7.41 | 11.79 ± 7.03 | 36.83 ± 3.17 | 0.85 ± 0.03 |
| **IMCM** | -1.38 ± 0.93 | 1.94 ± 0.83 | -12.38 ± 20.98 | 43.62 ± 11.56 | 4.40 ± 2.66 | 34.42 ± 3.92 | 0.91 ± 0.04 |
| **CI95%** | **ADCM** | [0.15, 1.18] | [2.52, 3.22] | [-11.93, 7.59] | [53.77, 60.70] | [8.50, 15.08] | [35.35, 38.31] | [0.84, 0.86] |
| **IMCM** | [-1.81, -0.94] | [1.55, 2.33] | [-22.20, -2.56] | [38.21, 49.04] | [3.16, 5.65] | [32.58, 36.25] | [0.89, 0.92] |

**Statistical tests**

**Normality Testing**

Before selecting an appropriate statistical test for our analysis, we first assessed the normality of the distribution of each metric within both datasets using the Shapiro-Wilk test. This step was crucial to determine whether parametric or non-parametric statistical methods were suitable. Our findings indicated that several metrics did not follow a normal distribution, particularly in the IMCM dataset where metrics such as Relative Error (SUV%) and Absolute Relative Error (SUV%) showed significant deviations from normality with p-values below 0.05. Similarly, Root Mean Squared Error and Peak Signal-to-Noise Ratio in the ADCM dataset also deviated significantly from a normal distribution.

Table 2: Evaluation of normality of all metric variables across both ADCM and IMCM datasets, by performing a Shapiro-Wilk test for each metric.

|  | **Metric** | **ADCM Statistic** | **ADCM P-value** | **IMCM Statistic** | **IMCM P-value** |
| --- | --- | --- | --- | --- | --- |
| 0 | Mean Error (SUV) | 0.962684 | 0.598745 | 0.964505 | 0.637189 |
| 1 | Mean Absolure Error (SUV) | 0.973161 | 0.819726 | 0.902938 | 0.046832 |
| 2 | Relative Error (SUV%) | 0.926644 | 0.133062 | 0.903215 | 0.047397 |
| 3 | Absolure Relative Error (SUV%) | 0.934748 | 0.190480 | 0.813324 | 0.001375 |
| 4 | Root Mean Squared Error | 0.875041 | 0.014425 | 0.670732 | 0.000018 |
| 5 | Peak Signal-to-Noise Ratio | 0.826691 | 0.002222 | 0.944862 | 0.295736 |
| 6 | Structual Similarity Index | 0.963606 | 0.618108 | 0.973200 | 0.820480 |

**Choice of Statistical Test**

Given the non-normality observed in several key metrics across the datasets, we opted to use the Wilcoxon signed-rank test, a non-parametric method, for our analysis. This test is particularly advantageous as it does not assume normality of the data and is ideal for comparing two related samples or repeated measurements on a single sample. This choice was reinforced by the need to handle the paired nature of our data, where each centre was analysed under both ADCM and IMCM conditions.

Our analysis revealed significant differences between the ADCM and IMCM methodologies in several metrics. Notably, the Mean Error (SUV) and Absolute Relative Error (SUV%) showed considerable variations, suggesting distinct impacts of the two methodologies on these particular metrics. The Wilcoxon test results indicated statistically significant differences with low p-values, underscoring the effectiveness of one method over the other in specific conditions.

Table 3: Summarized results of the Wilcoxon test with the False Discovery Rate (FDR) corrections applied to the p-values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **U-statistic** | **Corrected P-value** | **Rejected (Yes/No)** |
| Mean Error (SUV) | 371.0 | 2.79e-05 | Yes |
| Mean Absolute Error (SUV) | 330.0 | 0.000644 | Yes |
| Relative Error (SUV%) | 267.0 | 0.072045 | No |
| Absolute Relative Error (SUV%) | 357.0 | 4.03e-05 | Yes |
| Root Mean Squared Error | 364.0 | 3.41e-05 | Yes |
| Peak Signal-to-Noise Ratio | 286.0 | 0.024190 | Yes |
| Structural Similarity Index | 42.0 | 4.03e-05 | Yes |
| Rejected "Yes" means the null hypothesis (that there is no difference between the ADCM and IMCM datasets for the given metric) is rejected, indicating statistically significant differences.  Rejected "No" means the null hypothesis is not rejected, suggesting no significant difference for that metric based on the adjusted p-value. | | | |

The results from the Wilcoxon test with FDR correction show that there are statistically significant differences between the ADCM and IMCM datasets for most of the image-derived metrics, except for the "Relative Error (SUV%)" where the corrected p-value does not indicate a statistically significant difference.

Table 4: Summary statistics of quantitative parameters for different approaches on cross tracer (FDG dataset)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Method** | **ME** | **MAE** | **RE** | **ARE** | **RMSE** | **PSNR** | **SSI** |
| **Mean ± SD** | **ADCM** | 0.29 ± 0.58 | 1.08 ± 0.35 | 34.08 ± 48.96 | 80.22 ± 34.25 | 3.71 ± 4.14 | 37.38 ± 3.89 | 0.77 ± 0.09 |
| **TL-MC** | -0.54 ± 0.15 | 0.69 ± 0.13 | -39.70 ± 9.13 | 52.11 ± 7.61 | 1.18 ± 0.61 | 35.27 ± 6.18 | 0.78 ± 0.11 |
| **CI95%** | **ADCM** | [0.02, 0.55] | [0.92, 1.24] | [11.80, 56.37] | [64.63, 95.81] | [1.82, 5.59] | [35.61, 39.15] | [0.72, 0.81] |
| **TL-MC** | [-0.60, -0.47] | [0.63, 0.75] | [-43.86, -35.55] | [48.64, 55.57] | [0.91, 1.46] | [32.46, 38.09] | [0.73, 0.83] |

Table 5: Summary statistics of quantitative parameters for different centers tuned for each radiotracer separately (TL-MC) and tested on all test sets (centers 1-7).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Quantitative metric** | **Centre 1-4** | **Centre 5** | **Centre 6** | **Centre 7** | **All Test Set** | |
| **ME** | -0.56 ± 0.74 | -1.92 ± 0.58 | -0.46 ± 0.16 | -0.61 ± 0.09 | -0.95 ± 0.78 | |
| **MAE** | 1.28 ± 0.37 | 2.38 ± 0.76 | 0.64 ± 0.13 | 0.73 ± 0.12 | 1.30 ± 0.86 | |
| **RE** | -1.15 ± 18.77 | -19.87 ± 19.58 | -35.66 ± 11.69 | -43.38 ± 3.55 | -26.38 ± 21.03 | |
| **ARE** | 36.38 ± 7.12 | 48.45 ± 11.62 | 49.56 ± 8.11 | 54.42 ± 6.64 | 47.97 ± 10.53 | |
| **RMSE** | 2.90 ± 0.58 | 5.41 ± 3.05 | 1.00 ± 0.25 | 1.35 ± 0.78 | 2.75 ± 2.49 | |
| **PSNR** | 37.66 ± 2.67 | 32.25 ± 3.04 | 37.74 ± 6.59 | 33.03 ± 5.07 | 34.86 ± 5.16 | |
| **SSIM** | 0.93 ± 0.03 | 0.89 ± 0.03 | 0.80 ± 0.13 | 0.76 ± 0.092 | 0.84 ± 0.11 | |
| **CI 95%** | | | | |  |
| **ME** | [-1.18, 0.06] | [-2.29, -1.56] | [-0.57, -0.34] | [-0.67, -0.55] | [-1.19, -0.70] | |
| **MAE** | [0.97, 1.59] | [1.90, 2.87] | [0.55, 0.73] | [0.65, 0.81] | [1.03, 1.57] | |
| **RE** | [-16.84, 14.55] | [-32.31, -7.43] | [-44.02, -27.29] | [-45.77, -41.00] | [-33.01, -19.74] | |
| **ARE** | [30.42, 42.34] | [41.07, 55.84] | [43.75, 55.37] | [49.96, 58.89] | [44.65, 51.29] | |
| **RMSE** | [2.42, 3.38] | [3.47, 7.35] | [0.82, 1.18] | [0.83, 1.88] | [1.97, 3.54] | |
| **PSNR** | [35.43, 39.90] | [30.32, 34.18] | [37.62, 37.85] | 32.97 to 33.09 | [33.23, 36.48] | |
| **SSIM** | [0.90, 0.96] | [0.87, 0.91] | 0.68 to 0.91 | 0.70 to 0.82 | [0.81, 0.87] | |
| Column “Centre 1-4” represents the results of testing on the whole test set when training is performed on centre 1 to 4 data set. “Centre 5” represents as external centre with same radiotracer and Centre 6 & 7 test sets represent the results of tuned models, in which training and testing are performed for different radiotracer (whole 20% of the clean dataset). | | | | | | |