EXPLORATORY DATA ANALYSIS

CONTENT

ABSTRACT

This report contains indebt data audit and analysis for foot ulcer diagnosis for diabetes using infrared technologies. Full data exploration was performed on total of 167 volunteers where 45 are non-diabetic and 122 are diabetic patients.

The data was gotten from the plantar thermogram database where thermogram images of each patients left and right foot was taken and also 4 angiosomes (LPA, MPA, LCA, MCA) was taken from each foot which makes a total of 10 images per volunteers. The database also comes with a csv for each of the image which all was used for data exploration purpose.

This document contains information on how we data was loaded to the notebook environment, statistical comparison analysis performed, Thermal Change Index Analysis, Noise and outlier investigation and the correlation matrix

DATA LOADING

The plantar thermogram database csv data was loaded to the exploratory environment “Python Jupyter Notebook” using a system file manipulation function.

The function goes into the document path on the PC, it then goes into the plantar file and retrieves 11 data from each patients, these are:

1. Patients_id: The patient folder name
2. Left_foot: average of patients left foot scanned data
3. Right_foot: average of patients right foot data
4. Left_LCA: average left foot LCA data
5. Left_LPA: average left foot LPA data
6. Left_MCA: average left foot MCA data
7. Left_MPA: average left foot MPA data
8. Right_LCA: average right foot LCA data
9. Right_LPA: average right foot LPA data
10. Right_MCA: average right foot MCA data
11. Right_MPA: average right foot MPA data

The CSV data of each data point was ageregated on average, we avoided the 0s in the CSV as this was the background noise from the image and will distort the original data. Same process was carried out on the 2 categories of volunteers, i.e (control & diabetic group).

STATISTICAL COMPARISON

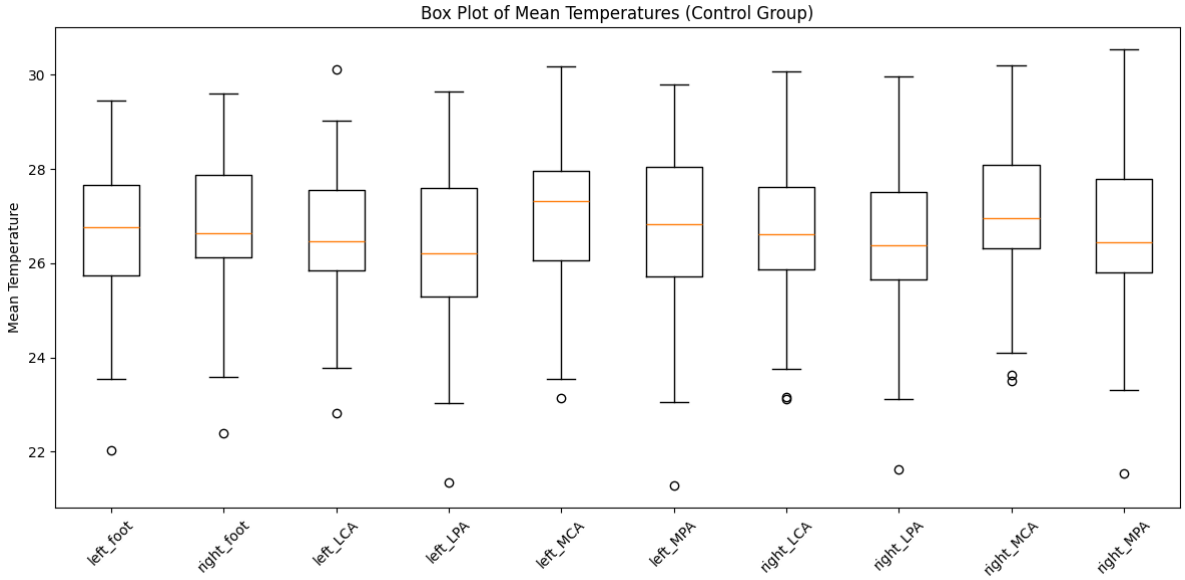
A full analysis and comparison was carried out on both group. The focus was finding a distinct difference from both group.

For successful comparison, we performed 2 analysis to understand the distribution of the data.

1. Boxplots
2. Symmetry Analysis (L - R)

1. Boxplots: We used the boxplots to understand how well the data is distributed, The goal was to understand outliers/noise, whiskers, median and frequency of data of each data points of each patients in the 2 category of patients

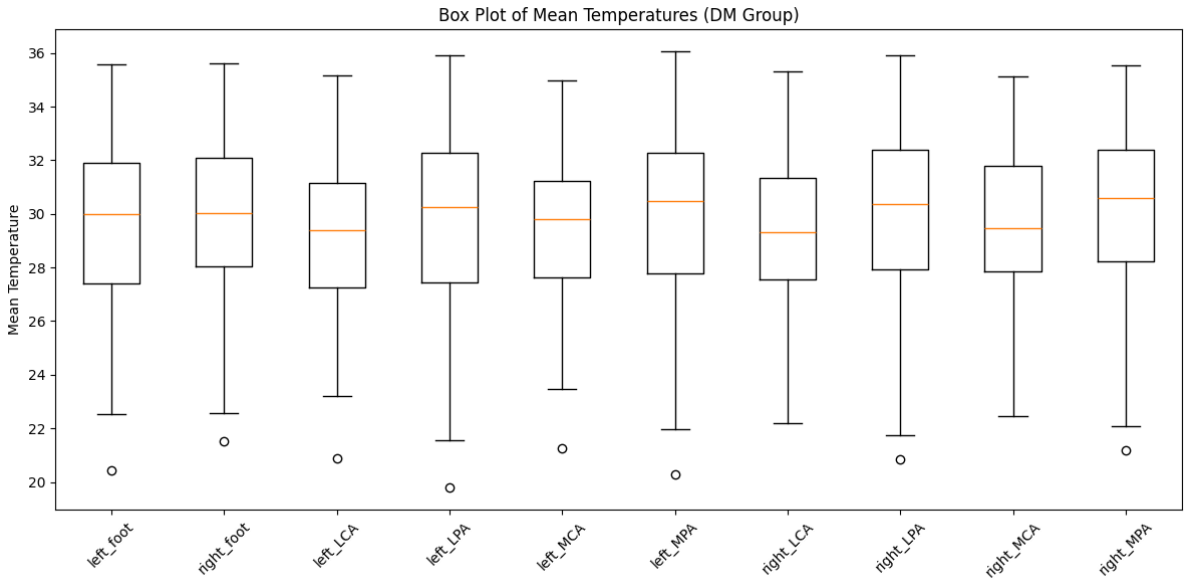
   A. The Control Group:



File ref: (control_boxlot.png)

Analysis Report: The average temperature of the control group data points seats between 26 to 27.5 which is expected to be this figure or lower. Although there are some extreme data points that reach up to 30; looking at the left_LCA and some of the angiosomes with stretched whiskers and outliers. But further experiments proves that these are likely noise from the site background

B. The DM Group:



File ref: (DM_boxlot.png)

Analysis Report: The interquartile range of the data is around 27 to 32 degree temperature, and the average seats between 29 to 31. Some angiosomes have up to 36 degree which are extreme cases and possibley cause foot ulcer formation

2. Symmetry Analysis (L - R): The goal was to compare the temperature of one foot to the other for each patients.

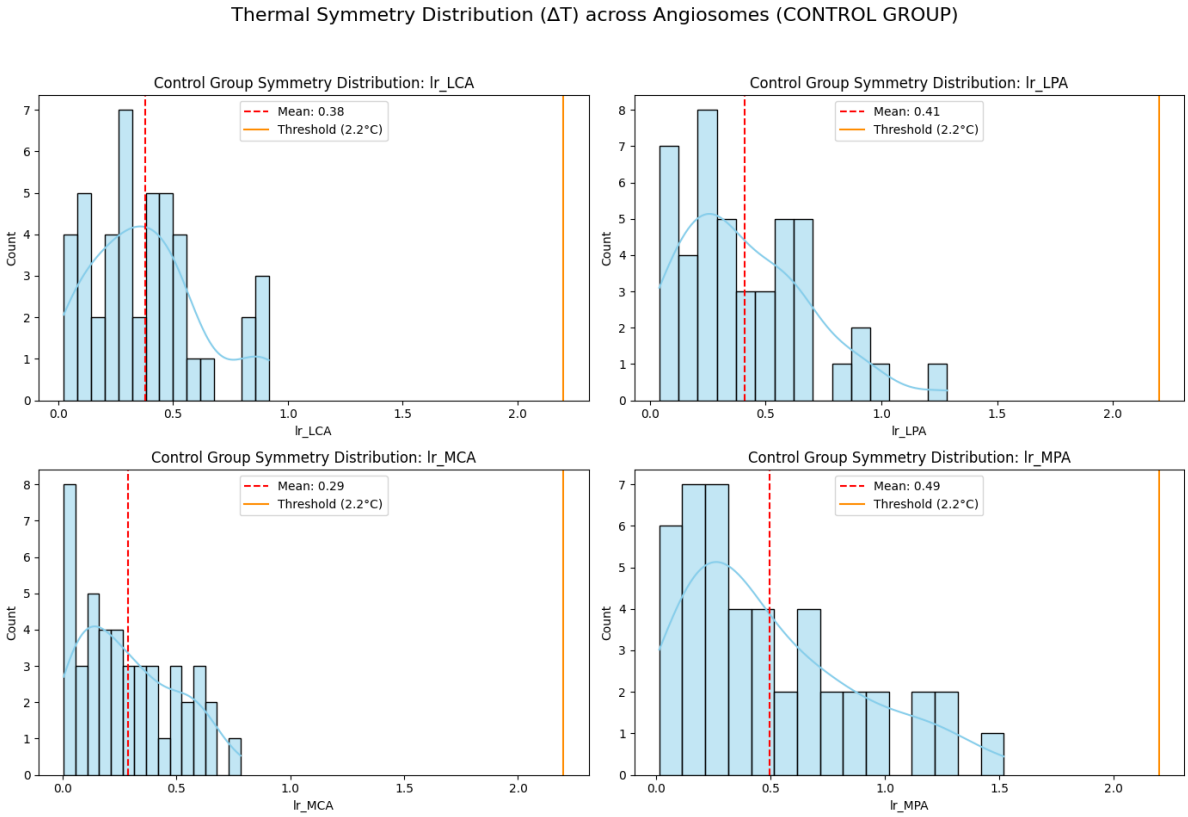
This was done by finding out the temperature difference between the patients right foot angiosomes to the left foot. We use the numpy library to find an absolute difference so we can avoid negative values

Code Snippet:

```
def lr(data):
    data["lr_LCA"] = np.abs(data["left_LCA"] - data["right_LCA"])
    data["lr_LPA"] = np.abs(data["left_LPA"] - data["right_LPA"])
    data["lr_MCA"] = np.abs(data["left_MCA"] - data["right_MCA"])
    data["lr_MPA"] = np.abs(data["left_MPA"] - data["right_MPA"])

    return data
```

To understand the symmetry distribution of each angiosomes in each group, we used an histogram plot and also gave a mean line (red dashed line) and a threshold of 2.2 degree (dark orange line) to understand the foot deviation.

A. Control Group:



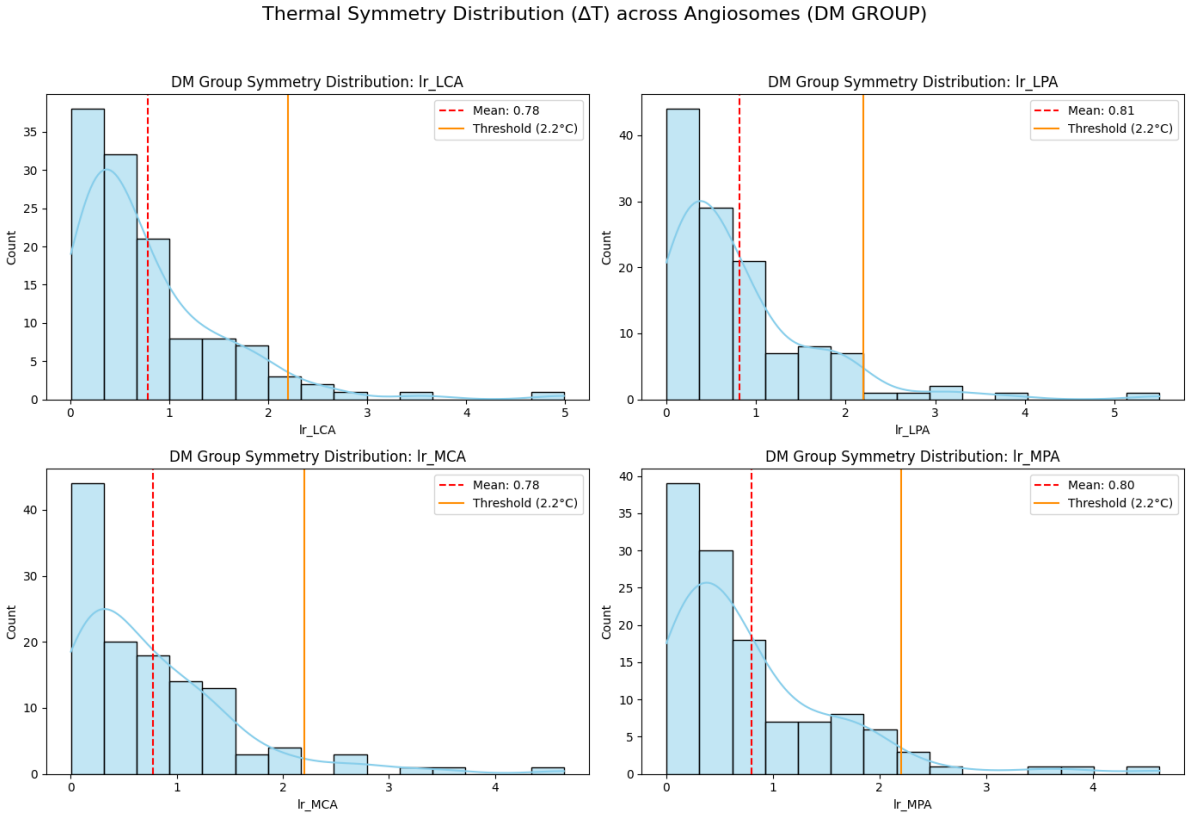
File_ref: (control_symmetry.png)

Analysis Report: Mean of the 4 angiosomes are not up to half (0.5) and the datapoints are far away from the set threshold of 2.2, this statistically means thatthere is little to no deviation of the control group patients foot and the plot confirms the data validation.

```
]: control_df[features].std()

]: lr_LCA    0.241561
   lr_LPA    0.283291
   lr_MCA    0.210145
   lr_MPA    0.382294
   dtype: float64
```

Above is the standard deviation of the group. And this further proves the insignificance of this symmetry deviation.

B. DM Group:



File_ref: (DM_symmetry.png)

Analysis Report: The mean of the diabetic group angiosomes is much closer to 1 than to half (0.5) and much higher than the corresponding control group mean, Some datapoints are above the threshold of 2.2 degree and this proves there is a common large deviation of foot temperature of these patients

```
[15]: DM_df[features].std()

[15]: lr_LCA    0.764346
      lr_LPA    0.855188
      lr_MCA    0.782484
      lr_MPA    0.800413
      dtype: float64
```

Above is the standard deviation of the diabetic group, and this information further proves how the group foot deviation is more significant than the control group data.

THERMAL CHANGE INDEX VALIDATION

This is also a deviation measurement metric, it simply measures how deviated each patient (both the control and the diabetic group) is from the control group mean. Its gotten by subtracting a patient angiosomes from the control group angiosomes and dividing by 4

To achieve this, we wrote a function that calculates each patients (both in the control group and the diabetic group) TCI.

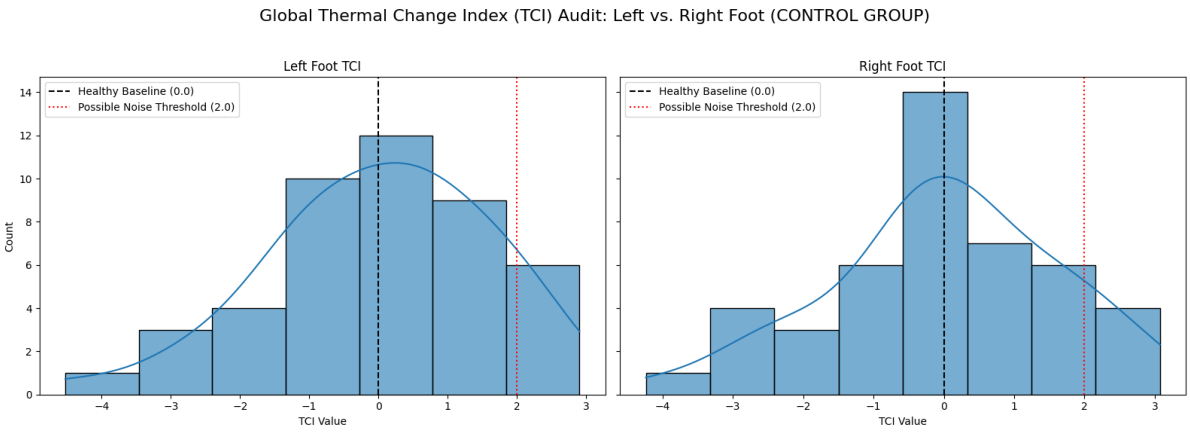
Code Snippet:

```
[16]: df_cols = ['left_LCA', 'left_LPA',
          'left_MCA', 'left_MPA', 'right_LCA', 'right_LPA', 'right_MCA',
          'right_MPA']
      control_means = control_df[df_cols].mean().values
      def calculate_tci(data, control_means):
          data["left_TCI"] = (
              (data["left_LCA"] - control_means[0]) +
              (data["left_LPA"] - control_means[1]) +
              (data["left_MCA"] - control_means[2]) +
              (data["left_MPA"] - control_means[3])
          ) / 4

          data["right_TCI"] = (
              (data["right_LCA"] - control_means[4]) +
              (data["right_LPA"] - control_means[5]) +
              (data["right_MCA"] - control_means[6]) +
              (data["right_MPA"] - control_means[7])
          ) / 4

          return data
```

To further understand the distribution of each patient's TCI values, we used an histogram plot for the left and right foot TCI. The focus was on the left and right foot.

We set a healthy baseline of 0 and a Threshold of 2 so we can further identify and classify extreme cases (for the diabetic group) or noise (for the control group)

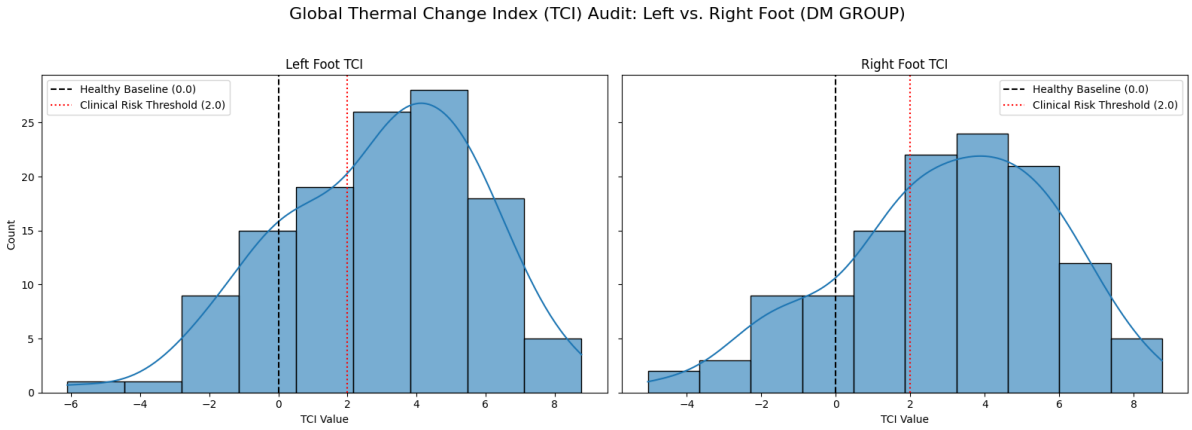
A. Control Group:



File_ref: (control_tci.png)

Analysis Report: the TCI data for the group left and right foot further shows that we have most patients fall inside the healthy baseline (which is expected) and about 4 fall outside the noise threshold which will be investigated in the future analysis

B. DM Group:



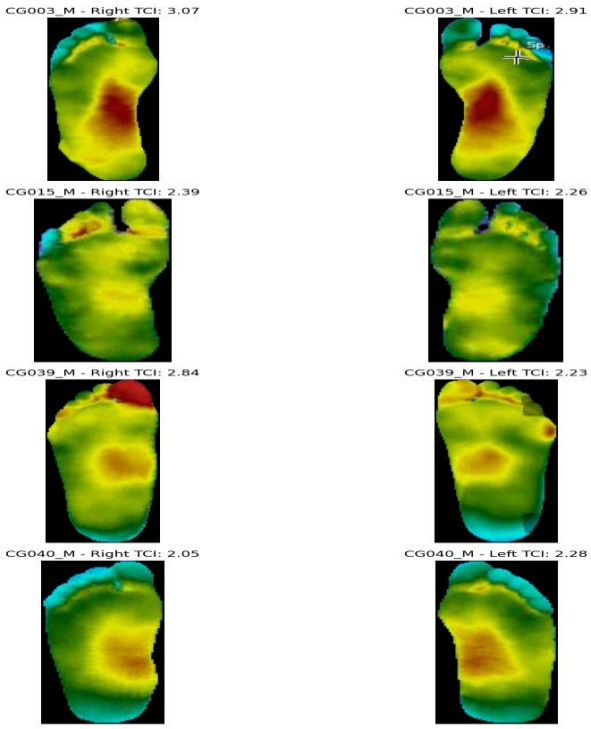
File_ref: (DM_tci.png)

Analysis Report: Unlike the Control group, the majority of the DM population sits well above the Clinical Risk Threshold (2.0). For the left foot, the primary density sits in the 4+ TCI range (over 25 subjects), while the right foot shows even more extreme cases reaching 8.0. This significant positive skew mathematically validates our dataset as representing pathological diabetic states.

NOISE & OUTLIER INVESTIGATION

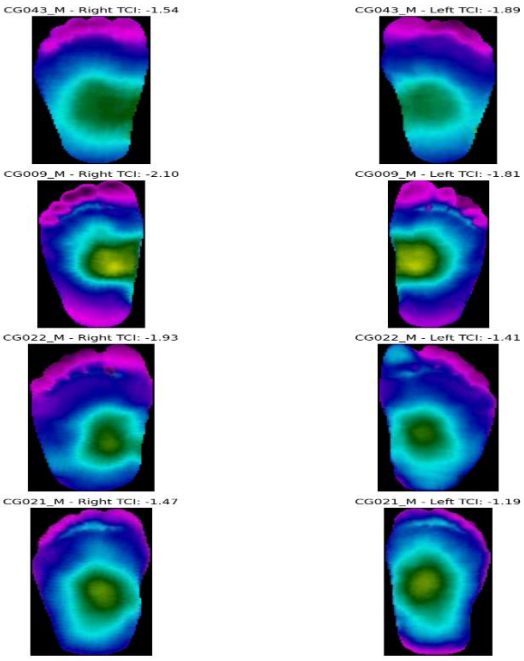
A validation test was carried out for the noise reason in the control group and also visually differentiate it from the normal temperature control group and severe temperature DM Group

1. Noise Image:



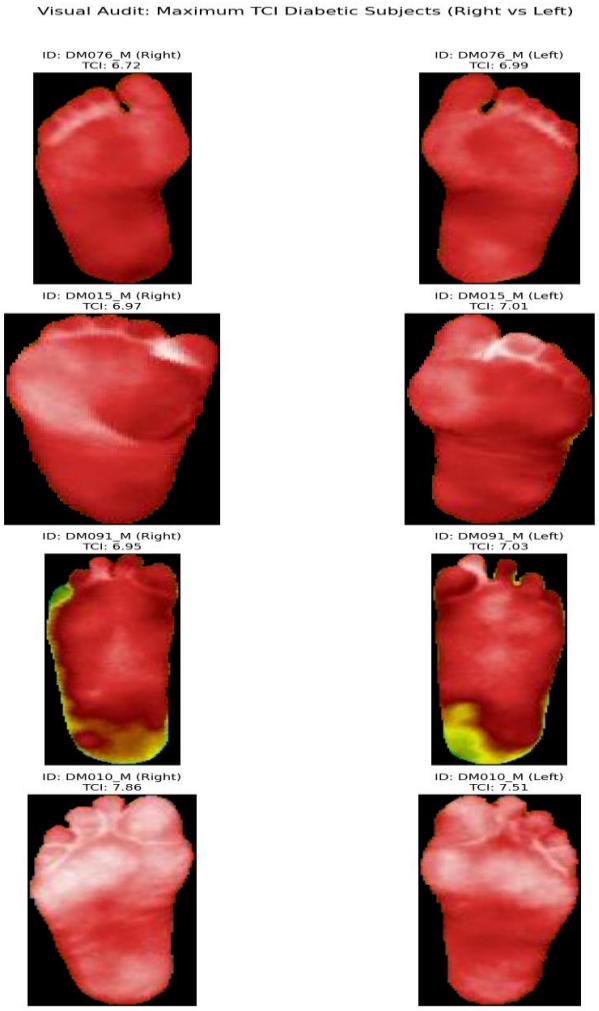
File_ref: (noise_val.png)

2. Typical Control Group Image:



File_ref: (control_val.png)

3. DM Group Severe case:



File_ref: (DM_val.png)

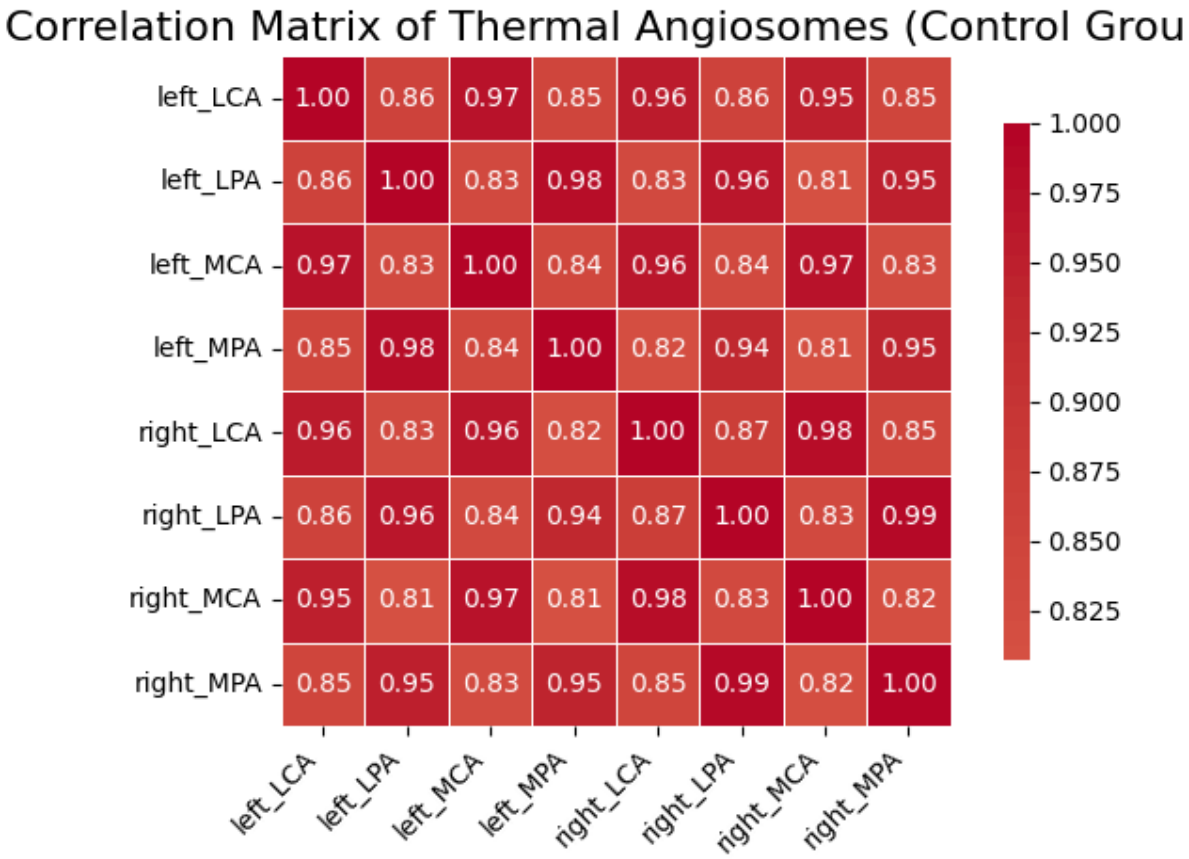
Analysis Report: The noise investigation confirms that while the Healthy Baseline is characterized by a deep purple and cyan distribution with TCI scores typically between -1.0 and -2.2, a subset of the control group exhibits a (greenish-yellow) signature that mimics early-stage diabetic pathology. Visual audit of outliers like e.g CG003_M reveals that this (green) heat is diffuse and often bleeds into the background, Suspecting the environmental noise such as a warm or cold floor rather than the localized (hot spots) typical of biological inflammation. This is statistically evidenced by the Global TCI Audit, where 4 volunteers sit beyond the 2.0 noise threshold, creating a potential for model bias if these images are not removed. By contrast, the DM Group distribution is shifted significantly to the right, with high-density peaks at TCI 4.0 to 6.0, proving that true diabetic heat is of a much higher intensity (red) and more organized than the greenish environmental interference found in the control group.

CORRELATION MATRIX

This is a form of symmetry in the feet, it makes us understand how a change in a particular area (angiosomes) affect the other area (angiosomes).

This was carried out so we can understand which angiosome temperature influences which.

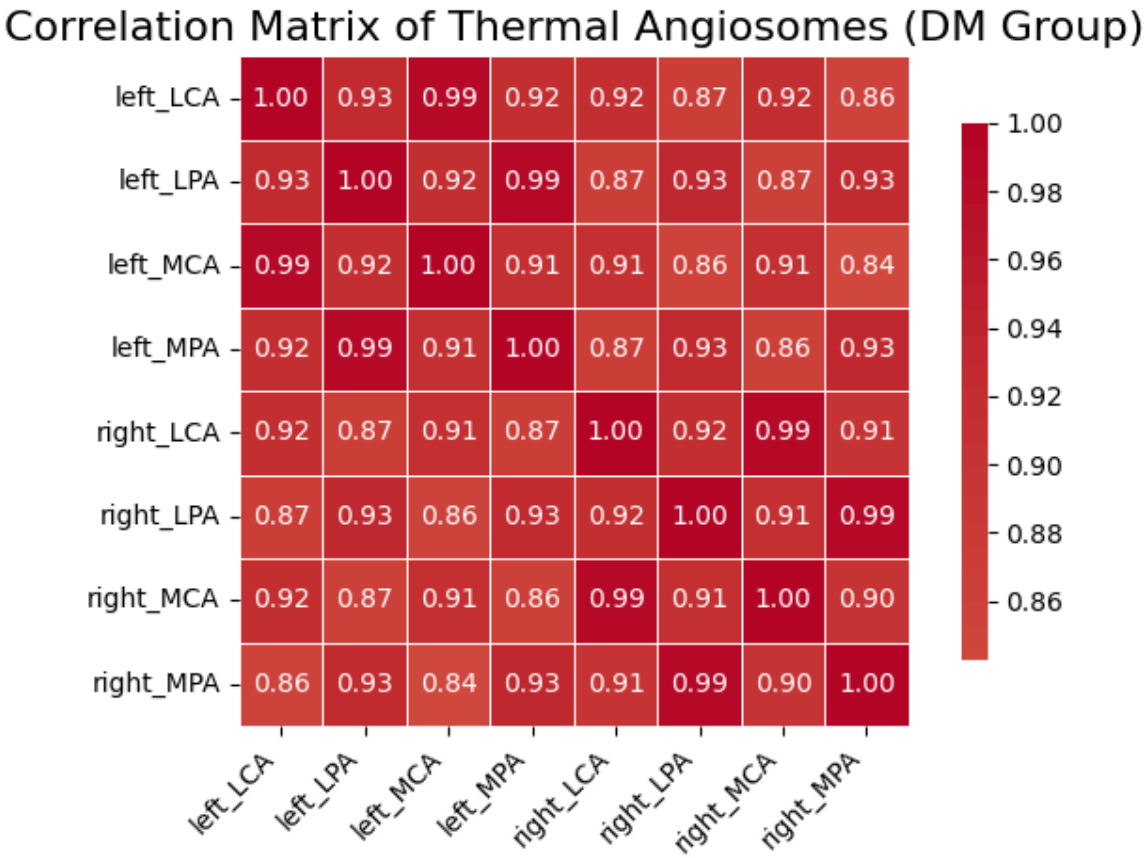
1. Control Group:



File_ref: (CONTROL_angiosome_correlation_heatmap.png)

Analysis Report: The chart above shows that the LCA correlates much significantly with the other foot's LCA and also a high correlation with the MCA. Same behavior as the LPA. But all 8 angiosomes are still highly correlated with each other

2. DM Group:



File_ref: (DM_angiosome_correlation_heatmap.png)

Analysis Report: There is also a similar behaviour here, the 8 angiosomes are well correlated significantly.

CONCLUSION

The establishment of a clear TCI baseline and the identification of environmental noise outliers mark the completion of our initial data audit. We have successfully validated the thermal contrast between healthy and diabetic populations, providing a definitive target for our upcoming image preprocessing and segmentation. We are now prepared to refine our dataset to ensure high model sensitivity toward true pathological signals