Automatic Hand Landmark Detection for Leprosy Diagnosis

Comparison of Output Adaptation Techniques for Hand Keypoint Prediction

Marek Tran[1]

Supervisor(s): Thomas Markhorst[1]**, Zhi-Yi Lin**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Marek Tran
Final project course: CSE3000 Research Project
Thesis committee: Jan van Gemert, Thomas Markhorst, Zhi-Yi Lin, Kaitai Liang

Abstract

Early detection of leprosy, a neglected tropical disease, is crucial to preventing irreversible nerve damage and disability. Analyzing temperature variations in hands using infrared (IR) cameras offers a potential low-cost alternative to existing medical equipment for early detection of leprosy. This study explores the adaptation of hand landmark detection models, commonly used for hand pose tracking, to infer the hypothenar area, a critical region for leprosy diagnosis. The research addresses the challenge of limited ground-truth data for the hypothenar keypoint by developing annotated datasets and evaluating machine learning models like Lasso Regression and XGBoost. These models significantly outperform the existing method of linear interpolation, demonstrating the feasibility of accurate hypothenar keypoint prediction even with limited training data. The findings contribute to the development of accessible, automated tools for early leprosy diagnosis, particularly in resource-constrained settings.

1 Introduction

Leprosy, a preventable and treatable condition, remains a significant global concern [1]. Its persistence is strongly associated with regions facing socioeconomic hardship. If untreated, it causes nerve dysfunction, also known as neuropathy, resulting in loss of sensation, muscular weakness, and potential deformities [2]. The World Health Organization (WHO) highlights delayed case detection, limited access to healthcare services, and insufficient research as critical challenges in the global effort to eliminate leprosy [1].

Leprosy detection can be facilitated by monitoring hand temperature regulation [3], [4], [5]. Subclinical leprosy impairs the vasomotor reflex, which regulates blood flow for temperature control [4], [3], and this can be assessed using Laser Doppler Flowmetry (LDF). Alternatively, Abbot et al. [4] highlight measuring skin temperature as a potential clinical indicator, as cold fingers strongly correlate with impaired vasomotor control. This approach is particularly valuable since LDF requires costly specialized equipment.

Infrared (IR) imaging (thermography) has emerged as a potential tool for leprosy assessment [5],[6]. Cavalheiro et al. [5] demonstrated the feasibility of using thermography to monitor temperature in the hands of patients with leprosy. The process involves measuring temperatures at regions of interest (ROIs).

The regions of interest are along the median and ulnar nerves. The median nerve runs from the wrist through the thenar area and innervates the thumb, index and middle finger. On the other side of the hand, the ulnar nerve runs from the wrist, through the hypothenar area to the ring and pinky finger. The specific definition of ROIs is different among studies, for example, Tiago et al. [6] measure temperature at fingertips, knuckles and rectangular areas on both thenar and hypothenar sides of the hand. Cavalheiro et al [5] define ROIs as points shown in Figure 2. Moreover, Cavalheiro et al [5] have observed that the hypothenar side was more affected.

Hand tracking is a key area of study, enabling advancements in human-computer interaction such as virtual reality, sign language recognition, and surgical assistance [7]. It commonly involves representing the hand via 21 keypoints that represent the hands' joints, as shown in Figure 1. Models like Google's MediaPipe Hands are widely used for their efficiency and ability to run in real-time on mobile devices [8].

Schemkes [9] investigated the use of hand keypoint detection models in infrared leprosy diagnosis, utilizing the same ROIs as Cavalheiro et al. [5]. In their process, the hands were photographed from above, with the palms facing the camera, as shown in Figure 2. The research [9] highlighted the time-efficiency and reproducibility benefits of automatization of leprosy detection. Collaborating with Dr. Arjan Knulst from the Green Pastures Hospital in Nepal, Schemkes [9] emphasized the need for accurate hand keypoint detection and the challenge posed by the lack of ground truth data.

A significant limitation of current hand landmark detection models is their inability to detect a keypoint near the hypothenar eminence (hereafter referred to as the hypothenar keypoint), annotated as R6 and L6 in Figure 2. Similarly, existing datasets do not include the annotation of the hypothenar keypoint. Schemkes [9] addressed this issue by adapting the output of MediaPipe, inferring the hypothenar keypoints through linear interpolation between the wrist and pinky knuckle keypoints as shown in Figure 3. However, this approach assumes that the hand remains completely still and maintains the same position and pose for 15 minutes, which limits its practical applicability.

Moreover, Schemkes [9] demonstrated that consumer-grade smartphone compatible infrared cameras can be used for leprosy detection. Furthermore, hand keypoint detection models are capable of running efficiently on smartphones, enabling real-time analysis. Effective leprosy detection using infrared imaging, however, requires the accurate tracking of specific hand keypoints, including the hypothenar keypoint.

The objective of this study is to improve hypothenar keypoint inference. This improvement could contribute to the development of a robust, automatic solution utilizing available hardware. By reducing reliance on expensive medical equipment, such a solution has the potential to make early leprosy detection more accessible in resource-constrained regions, directly addressing challenges highlighted by the World Health Organization [1].

The primary contributions of this work are: (1) a novel IR and RGB image dataset of 160 manually annotated images with the hypothenar keypoint and 21 standard keypoints (2) another dataset of 800-image hand palm dataset with manually annotated hypothenar keypoint and 21 keypoints annotated by the MediaPipe model; and (3) a comparative evaluation of machine learning models and training data volume for robust hypothenar keypoint inference.

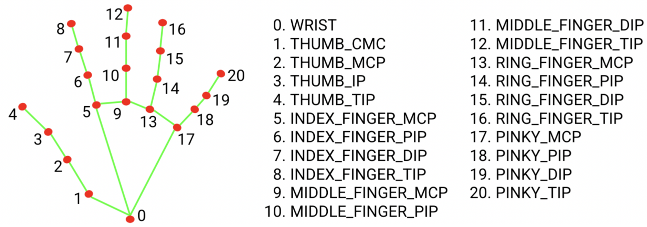


Figure 1: Visualisation of 21 standard keypoints and their names as produced by MediaPipe [17].

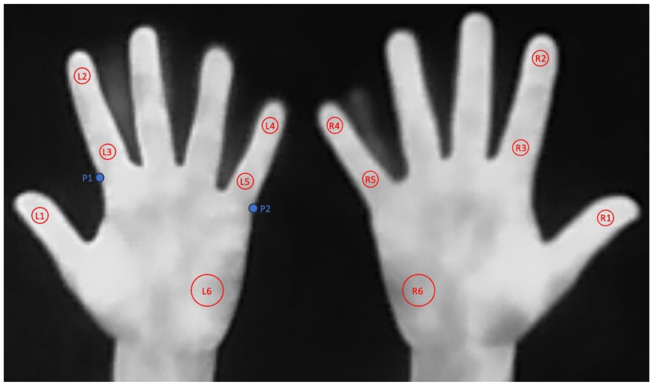


Figure 2: Regions of interest (red circles) for leprosy detection using infrared imaging. Hand breadth indicated as distance between points P1 and P2 (blue points). Adapted from [9].

2 Related Work

2.1 Hand Keypoint Detection Methods

Hand keypoint detection in RGB images is a popular area of research in computer vision, resulting in a variety of approaches [10], [11], [12]. Among them, MediaPipe Hands, OpenPose, and YOLOv11-pose are the most widely used in practice. MediaPipe Hands, developed by Google [8], and YOLOv11-pose, developed by Ultralytics [13], offer implementation tools that simplify their deployment on mobile devices. This is a significant advantage over OpenPose, developed at Carnegie Mellon University [14], which lacks such readily available tooling. However, MediaPipe Hands is arguably the most adopted solution. This is evident in the numerous hand image datasets that utilize MediaPipe-generated keypoint annotations [15], [16]. As such, this paper uses the MediaPipe Hands model.

Infrared Native Hand Keypoint Detection Methods

Compared to hand keypoint detection in RGB images, hand keypoint detection in infrared (IR) is less developed, though research in this area exists. Some studies have adapted body keypoint detection for IR [18], [19], but specific focus on hands is limited. Park et al. [20] detect hand keypoints in IR images for hand pose estimation, focusing on the 21 standard keypoints but excluding the hypothenar keypoint. However, the IR images used by Park et al. resemble near-infrared (NIR) images, which are commonly utilized for night vision and may not be suitable for temperature measurement.

2.2 Regions of Interest Definition

According to Schemkes [9], the keypoints 4, 6, 8, 18, and 20 can be used directly as regions of interest. In a study by Tiago et al. [6], the hypothenar-side region of interest is defined as a large rectangle, while some studies [5], [9] consider it as a point. In this paper, the point is defined in line with Cavalheiro et al. [5]. Moreover, the rectangular definition could be inferred from the hypothenar keypoint and the knuckle keypoints of the pinky and ring finger. Therefore, in this paper, the problem of inferring the rectangular area is reducible to inferring the hypothenar keypoint.

Schemkes [9] explicitly states that the radius of the L6 and R6 regions is 20% of the palm breadth, annotated as P1 and P2 in Figure 2. However, the visual representation in Schemkes' figures suggests a radius closer to 10% of the palm breadth. To err on the side of caution, this paper adopts a radius of 10% of the handbreadth as benchmark.

2.3 Inference Adaptation Methods

The main problem is the lack of annotated datasets with the hypothenar keypoint while there is ample data for other keypoints. The inference of the hypothenar point is essentially a regression task, relying on inherent spatial relationships between existing keypoints and the hypothenar keypoint. In hand gesture detection tasks, the spatial relationships between hand keypoints are assumed to be nonlinear due to stretching and curving of the hands during various hand poses. In contrast, Schemkes [9] studied hands in a flat pose, in which the spatial relationships between keypoints can be assumed to be linear.

Therefore, a linear and nonlinear models were selected for comparison and evaluation of the extent of the linearity assumption. The models chosen for this study include a generalization of Schemkes's interpolation method, k-Nearest Neighbors (k-NN), Random Forest, XGBoost, and Lasso regression.

K-Nearest Neighbors Regression

The k-Nearest Neighbors (k-NN) algorithm was explored as a baseline model. The k-NN model is particularly appealing in this case because of its simplicity. Although k-NN can achieve high accuracy with sufficient training data, it cannot learn the importance of the features [21]. In the context of keypoint matching, k-NN has demonstrated effectiveness in previous studies [22]. The regression is just the interpolation of closest matching keypoint sets. Therefore, k-NN could potentially capture the hypothenar point given a certain hand shape and pose.

Schemkes Interpolation

Schemkes uses linear interpolation of the coordinate values of two keypoints to infer the target hypothenar keypoint[9]. Schemkes uses the pinky finger knuckle joint (PINKY_MCP) and the wrist joint. As seen in Figure 3, the vertical and horizontal distances between keypoints WRIST (0) and PINKY_MCP (17) are multiplied by weights of 1/3 and 2/3 respectively. This can be generalized as α and β, which can be empirically determined as an average. This leads to the definition of Schemkes interpolation, which will use the

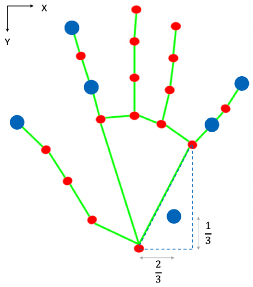


Figure 3: Hypothenar keypoint inference by linear interpolation of WRIST and PINKY_MCP keypoints. Regions of interest highlighted in blue [9].

WRIST and some MCP keypoint to predict a point on the palm. The parameters of this model are the keypoints used for interpolation.

With wrist and knuckle keypoints as u_{wr}, v_{mcp}, the point P is obtained by their weighted interpolation. Therefore, the coordinates of point P are:

$$P = \begin{bmatrix} P_{\text{x}} \\ P_{\text{y}} \end{bmatrix} = \begin{bmatrix} (1-\alpha)x_{\text{wr}} + \alpha x_{\text{mcp}} \\ (1-\beta)y_{\text{wr}} + \beta y_{\text{mcp}} \end{bmatrix}$$

Its effectiveness is dependent on the assumption of relatively flat hand positions, where nonlinear influences are minimized. Despite these limitations, Schemkes interpolation provides a valuable baseline benchmark for assessing the performance of more sophisticated models.

Lasso Linear Regression
Lasso linear regression employs L1 regularization which shrinks unimportant predictors' coefficients to zero [21]. This is beneficial because many x-axis values might be poor predictors for y-axis values. Unlike linear interpolation, which operates within a single axis, lasso regression models the influence between different axes. For example, in Figure 1, the y-coordinate of keypoint 4 could be positively correlated with the x-coordinate of keypoint 1, a relationship L1 regularization can help uncover.

Random Forest Regression
A random forest (RF) is an ensemble learning method that combines the outputs of multiple regression trees to produce a single prediction. For hand keypoint coordinate prediction, Random Forests are robust to outliers due to feature subsampling. For instance, if an important predictor e.g. pinky knuckle y-value contains outliers, the outliers impact is reduced because the feature is not present in all tree despite the importance. Therefore, if only few features are anomalous, the predicted value can still be accurate. Another key advantage of RFs is their ability to yield relatively accurate predictions even with small sample sizes, and increasing the number of trees generally improves accuracy without causing overfitting [23]. However, the random selection of features might be detrimental to the result due to noisy data of finger keypoints.

XGBoost
XGBoost [24], an advanced gradient boosting framework, iteratively builds trees, with each new tree trained to minimize the residual errors of its predecessors. Compared to Random Forest, the adaptive neighborhood property of XGBoost allows flexibility in predictions for hands with some shared properties [25]. This means that if the hypothenar keypoint has high variance when the thumb and pinky keypoints' x-values are close to 0, XGBoost can capture this variance for this particular feature space e.g. increased yaw of the hand - appearing slimmer to the camera. Furthermore, regularization prioritizes the most influential variables, such as palmar instead of fingertip keypoints. Given sufficient data and appropriate hyperparameter tuning, XGBoost is expected to outperform Random Forest.

3 Methodology

This study investigates the comparative accuracy of regression models for the task of hypothenar keypoint prediction. The main challenge is the absence of ground-truth data. Large hand image datasets with standard keypoint annotations are available or can be annotated with hand keypoint detectors such as MediaPipe. Due to the absence of hypothenar keypoint datasets, the models are compared in performing a similar prediction task with large datasets available in order to investigate the data quantity requirements for keypoint prediction tasks.

3.1 Experimental Design

Each dataset inherently carries biases introduced during its creation, such as camera types, instructions given to subjects, or the subjects themselves. The difference between datasets is commonly referred to as ”dataset shift” [26], a factor that must be considered when analyzing results. The reduction in prediction accuracy caused by this difference is known as ”generalization error” [26].

This study is structured around two experiments. The first experiment investigates model performance on a larger, publicly available dataset for a similar keypoint prediction task. This allows for an evaluation of the impact of data size on predictive accuracy and provides insights into the effects of dataset shift on model generalization error between the datasets. The second experiment aims to identify the optimal model for hypothenar keypoint prediction using the datasets created for this study.

3.2 Thumb Keypoint Prediction Experiment

How accurately can a hand keypoint be predicted without data volume constraints? For the task of prediction of a similar keypoint, the THUMB_CMC keypoint was omitted during model training and designated as the prediction target. This experiment allows for an assessment of model performance with increasing training data sizes. The THUMB_CMC keypoint shown in Figure 1 was selected due to its proximity to the hypothenar keypoint.

How does difference in datasets impact hand keypoint prediction accuracy? Model predictions were assessed under two conditions: first, on data sampled from the same dataset as the training data, and second, on data sampled from a different dataset.

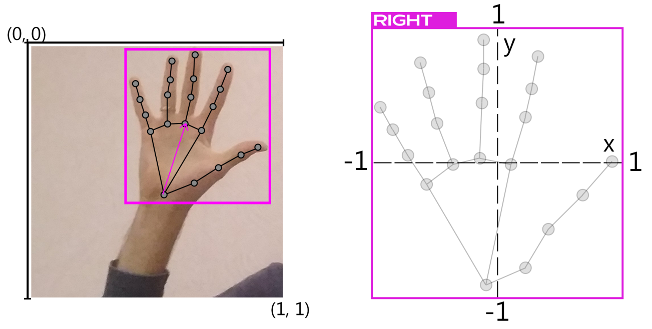
3.3 Hypothenar Keypoint Prediction Experiment

How accurately can the hypothenar keypoint be predicted with limited data volume? For the prediction of the hypothenar keypoint, the standard 21 keypoints were used as input and the hypothenar keypoint was set as the prediction target.

3.4 Data Preprocessing

Keypoint Normalization Pipeline

The keypoint coordinate values in the datasets are normalized relative to the width and height of the image. However, bounding box information is not provided in some datasets. Although hand poses remain consistent, variations in hand rotation and apparent size, due to proximity to the camera, introduce high variance in the keypoints' location, rotation, and scale. As a result, a preprocessing normalization step is required.



(a) Image space

(b) Rotated bounding box space

Figure 4: Keypoint normalization before (a) and after (b).

To address this, a data processing pipeline was developed. Raw keypoint coordinates in image space (as shown in Figure 4a) are used to generate a bounding box. The bounding box center is defined as the new origin, and the keypoint coordinates are scaled relative to the bounding box. The palm of the hand is then aligned cardinally north using the wrist and middle finger knuckle keypoints, a process expressible through linear transformations. This is applied to every set of keypoints, representing a hand. By training models on the normalized keypoints, the predictions can be therefore transformed into useful predictions back into image space to display on an image.

3.5 Error Function

Definition

For predicting hand keypoints, Euclidean distance was chosen as the primary metric due to its simplicity. While the Object Keypoint Similarity (OKS) metric accounts for scale, shape, and segmentation, it requires empirically determined parameters, making it unnecessarily complex for this study's goals. Instead, normalized Euclidean distance in bounding box space provides a straightforward and reliable measure.

The normalized keypoint error function is essentially the sum of distances per keypoint:

$$E_{\text{kpts}} = \frac{1}{N_{\text{kpts}}} \sum_{i=1}^{N_{\text{kpts}}} D_{\text{bb}}(k_i) \tag{1}$$

Where:

- E_{kpts}: Mean Euclidean distance error.
- N_{kpts}: Total number of keypoints evaluated.
- $D_{\text{bb}}(k_i)$: Distance between predicted and ground truth locations for the i-th keypoint, in bounding box space.

The loss function is calculated as the mean of these errors, with M being the number of hands in the dataset. This results in a singular value for evaluating how a model performs on a particular dataset.

$$L = \frac{1}{M} \sum_{i=1}^{M} E_{\text{kpts}}^{(i)} \tag{2}$$

For single-point prediction tasks, $N_{\text{kpts}} = 1$, reducing the loss function to the mean Euclidean distance between the target and predicted point.

4 Experimental Setup

4.1 Data Collection and Annotation

For leprosy diagnosis, the regions of interest need to be visible for taking measurements. Therefore the primary requirement for data is that it contains images of hands in similar positions during leprosy assessment and annotations of the hypothenar keypoint.

Studies [9][5] have shown that it is sufficient to take temperature measurements from the front (palmar) side of the hand, while Tiago et al. [6] additionally took measurements from the back (dorsal) side of the hand. In this study, handedness and face of the hand is neglected and not considered during training. A palm-facing left hand in keypoint representation is considered to be equivalent to back-facing right hand and vice-versa. Thus, three categories of hand poses as defined in the HaGRID [16] dataset are used as reference. The hand poses are shown in Figure 6.

Construction of RGB/IR keypoints dataset

In collaboration with other research team members, we have created a dataset of palmar hand images in RGB and infrared. The hands of 5 subjects were photographed under various conditions and backgrounds. The dataset contains manual annotations of 22 keypoints (21 standard and hypothenar keypoint) illustrated in Figure 5. The RGB and IR camera was placed 70cm above the hands which were laid out open on the table with palms facing up, towards the camera. In both domains, 80 images were taken which totals 160 annotated images for the purpose of this paper. However, the number of subjects is too low despite variance in poses. Thus, more ground-truth data of the hypothenar keypoint is required.

Annotation Quality

The quality of annotations was evaluated using Google's mean palm size normalized absolute error (MNAE) [8], comparing MediaPipe's detections to our manual annotations.

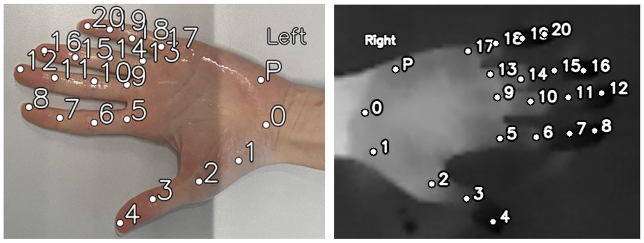


Figure 5: Visualisation of RGB/IR dataset annotations. Standard keypoints are numbered, hypothenar point is annotated with letter "P"

The dataset achieved a MNAE of 1.25, significantly lower than Google's reported 10.09 [27], likely due to the poses in our dataset, which lacks occlusions and complex gestures.

Annotation of Human Palm Images (HPI) Dataset

The HPI [28] dataset contains 800 human palm images. The standard annotatations were generated using MediaPipe and the hypothenar point was manually annotated. It includes 400 male and 400 female palm images, with 200 subjects of each gender providing images of both left and right palms [28].

HaGRID dataset

The HaGRID dataset comprises approximately 30,000 images per gesture category, with keypoint annotations generated using MediaPipe [16]. The dataset features a diverse set of human subjects performing various hand gestures. The selected gesture categories are shown in Figure 6.

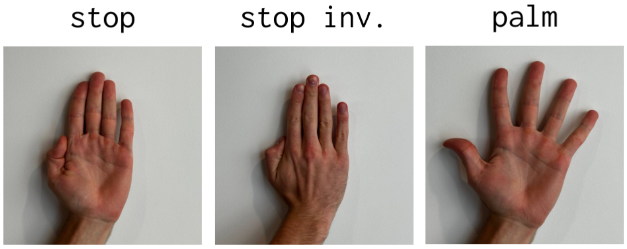


Figure 6: Selected HaGRID gestures. Adapted from [16].

11k Hands Dataset

The 11K Hands dataset provides top-down view images of the dorsal and palmar sides of the hands [29]. All hand gestures in this dataset fall under the categories selected in the HaGRID dataset. This dataset is a subset of an annotated dataset which was annotated using MediaPipe by Dsilva [15]. The size of the dataset is about 11 000 images of 190 subjects' hands.

Average Hand Dimensions

The values for average hand length and hand breadth, 17.48 cm and 8.04 cm respectively, are taken from the results of a study by Khazri [30]. In this study, hand length is defined as the distance between the wrist and the tip of the middle finger, while hand breadth is as shown in Figure 2. Using the Hypothenar ROI sizing defined by Schemkes [9], the resulting real-world radius is 0.8 cm, which serves as a benchmark value.

Tuned Hyperparameters

Negative MSE was selected as the tuning score for simplicity and efficiency during implementation using the *scikit-learn* Python library [31]. The Euclidean distance can be derived from its squared error by taking the square root of twice the squared error value, aligning the objective of maximizing negative squared error with minimizing Euclidean distance. For Random Forest, the number of trees, maximum tree depth, minimum samples required to split a node, minimum samples required at a leaf node, and the proportion of features considered for each split were adjusted. For XGBoost, the parameters tuned included the number of boosting rounds, maximum tree depth, learning rate, minimum child weight, subsample ratio, column subsample ratio, and the minimum loss reduction required for further partitioning. In the k-Nearest Neighbors model, the number of neighbors, weighting scheme, distance metric, and the power parameter for the Minkowski metric were optimized. Finally, for Lasso Regression, the regularization strength parameter was tuned.

4.2 Thumb Keypoint Prediction Experiment

Data Preparation and Partitioning

The HaGRID dataset used its predefined training, validation, and test partitions (Table 1). For the 11k Hands dataset, a test set was created by splitting the set into 80% training and 20% test without shuffling to prevent data leakage, as consecutive entries belong to the same subject. Partition sizes are detailed in Table 1. The combined training set comprises 57,837 samples, and the combined test set comprises 12,863 samples.

Table 1: Hand pose composition of the combined dataset

Subset	Train	Test
Hagrid Palm	18,536	3,974
Hagrid Stop	18,224	3,953
Hagrid Stopinv	17,364	4,007
11k Hands	3,713	929
Total	57,837	12,863

Hypothenar Dataset The RGB/IR and HPI dataset annotations were processed and combined, hereafter referred to as the Hypothenar set. After cleaning, the total size is 927.

4.3 Datasets Used for Evaluation

Two testing sets are used to evaluate models trained on the combined training set, with the THUMB_CMC keypoint as the prediction target. The first evaluation scenario utilizes the large test partition, as outlined in Table 1, hereafter referred to as the "test set." The second evaluation scenario uses the whole Hypothenar set.

Tuning Data

To prevent tuning the parameters to overfit on the same validation set, k-fold cross-validation with $k = 3$ was used with a 50% randomly subsampled training set.

Experimental Procedure
First, all data is ingested and pre-processed as described in Section 3.4. Then, hyperparameters are tuned for each model. The models with tuned hyperparameters are then trained on increasing volumes of training data. The experiment consists of 11 rounds. For each round a training size is determined and the loss as described in 2 of each model is evaluated for both test sets.

Model Settings
Lasso Regression without proximity bias Given the proximity of the THUMB_CMC to the wrist and other thumb keypoints, a standard Lasso regression model may exhibit bias by relying heavily on these features. To address this, an alternative Lasso regression model is evaluated, excluding information from nearby thumb keypoints to simulate a scenario where the target keypoint lacks close-proximity features.

The generalized Schemkes linear interpolation is set to interpolate between the WRIST and INDEX_FINGER_MCP keypoints, displayed in Figures 1.

4.4 Hypothenar Point Prediction Experiment

The experiment for direct hypothenar point prediction largely mirrored the procedure of the thumb keypoint prediction experiment, with key modifications.

This experiment utilized the Hypothenar set, partitioned into an 80/20 split. K-fold cross-validation was used during tuning. Tuning was done with $K_t = 10$ due to low amounts of data. Unlike the thumb keypoint prediction experiment, this experiment leveraged all 21 keypoints as input features to predict the location of the hypothenar point. This setup enables a direct assessment of the feasibility of accurate hypothenar point prediction using the hypothenar set.

5 Results

5.1 Thumb Keypoint Prediction Results

The results are illustrated in Figure 7, showing steady improvements in prediction for XGBoost, Random Forest and KNN with increasing training data volumes. Lasso regression and Schmemkes' interpolation as parametric models, have essentially constant loss throughout.

A summary of the results is shown in Table 2 which contains the loss when training data has 1000 and 57837 datapoints, the resulting reduction in loss and mean generalisation error (MGE) which is the sum of absolute differences between the losses from the Test Set and Hypothenar Set divided by number of evaluation rounds, essentially the mean distance between the Hypothenar Set and Test Set curves in shown Figure 7.

The best prediction models are XGBoost and Lasso regression in that respective order. XGBoost is the only model which performs worse on the Hypothenar Set with higher training data volumes while other non-parametric models improved. The Thumbless Lasso regression performs worse than regular Lasso regression with absolute difference of 0.0172. At all training volumes, all models outperform the generalized Schemkes' linear interpolation of two points.

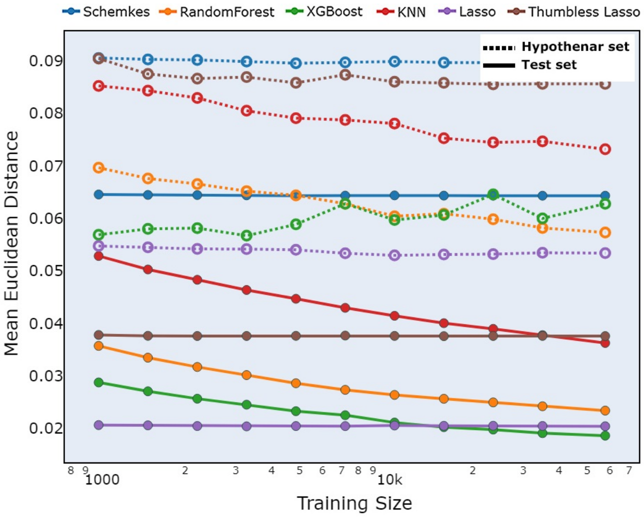


Figure 7: Comparison of models' THUMB_CMC prediction losses with increasing training data size on Test and Hypothenar Set.

5.2 Hypothenar Keypoint Prediction Results

The result for Schemkes [9] interpolation using the original ratios, as shown in Figure 3, was 0.234. However, this result was omitted from the plots for clarity. A clear elbow is observed in the training curves in Figure 8, occurring at a training size of 222.

As summarized in Table 3, all models consistently outperformed the generalized linear interpolation by Schemkes [9] for Hypothenar keypoint prediction. The loss reduction was calculated from the elbow point to the maximum training data size. Among the models, XGBoost achieved the best performance, followed closely by Lasso regression. The resulting training curves for all models are plotted in Figure 8.

Lasso regression uniquely exhibits a valley in its training curve, with a global minimum loss of 0.064 at a training size of 518. Additionally, Lasso regression achieves the lowest loss at smaller training sizes. Across all models, the training curves are steep at low training set sizes and plateau beyond the elbow point. This trend is particularly evident for Lasso regression, which demonstrates a modest loss reduction of 0.37% between the elbow point and the maximum training volume.

Feature Importances
Lasso regression model's intercepts for the x and y coordinates were -0.08 and -0.66, respectively. The strongest predictors of the Hypothenar keypoint's x and y coordinates were the x-coordinates of the PINKY_MCP and PINKY_TIP keypoints, with coefficients of 0.82 and 0.05, respectively. The Schemkes' interpolation model achieved the lowest loss when using the WRIST and PINKY_MCP keypoints. The most important features for XGBoost were the x-coordinates of THUMB_MCP, RING_FINGER_MCP and PINKY_FINGER_MCP in the respective order.

5.3 Residual Error Analysis

The 2D residual distributions for hypothenar prediction are visualized in Figure 9. The residuals for most models exhibit

Table 2: Performance metrics for thumb keypoint prediciton on Test and Hypothenar Sets at low and high training data volume

Model	**Test** $[10^{-2}]$		**Hypothenar** $[10^{-2}]$		**Loss Reduction** $[\%]$		**MGE** $[10^{-2}]$
	Loss_{1k}	Loss_{57k}	Loss_{1k}	Loss_{57k}	**Test**	**Hypoth.**	
Schemkes	6.46	6.43	9.06	8.96	0.37	1.05	2.55
RandomForest	3.57	2.34	6.97	5.73	34.56	17.74	3.47
XGBoost	2.87	1.86	5.69	6.28	35.28	-10.41	3.72
KNN	5.28	3.63	8.53	7.32	31.38	14.15	3.52
Lasso	2.06	2.04	5.47	5.34	1.08	2.48	3.33
Lasso_no_thumb	3.78	3.76	9.05	8.56	0.57	5.35	4.91

Table 3: Performance metrics for Hypothenar keypoint prediction.

Model	Loss_{74} $[10^{-2}]$	$\text{Loss}_{222}[10^{-2}]$	$\text{Loss}_{741}[10^{-2}]$	Loss $\text{Reduction}_{222\text{-}741}$ (%)
Schemkes	11.07	11.149	11.17	-0.18
RandomForest	10.04	7.57	6.61	12.74
XGBoost	9.81	6.96	6.24	10.33
KNN	10.86	8.66	7.11	17.97
Lasso	8.66	6.54	6.52	0.37

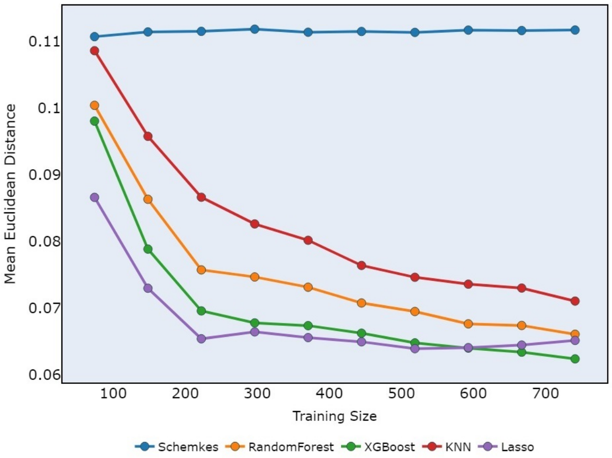


Figure 8: Comparison of models' Hypothenar keypoint prediction losses with increasing training data size.

a normal distribution, validating the use of bivariate normal modeling for confidence radius estimation. The exception is Schemkes Interpolation, which shows a non-normal residual pattern. Based on the residual distributions, a 95% confidence ellipses from covariances are derived. A circular approximation is used by taking the mean of width and height of ellipses as circle diameter.

5.4 Real World Estimates

The mean distance between the corresponding keypoints in the Hypothenar set is 1.88 corresponding to 17.48 cm. Since the keypoints normalization step first scales and then orients the hands keypoints, the axes of the residual plots do not necessarily align with hand length and breadth. As a consequence, an assumption of isotropic scaling by the larger value (hand length) is used to calculate an approximate radius of the 95% confidence circle.

The resulting 95% confidence ellipses and approximated circle radii are featured in Table 4. The best prediction model is XGBoost with a corresponding real-world radius of 1.19cm, indicating that 95% of predictions lie within 1.19cm of ground-truth, higher than 0.8cm benchmark value.

The dataset shift manifests as loss delta of 0.0372, corresponding to 0.34cm and 0.3cm for XGBoost and Lasso regression respectively.

Table 4: Comparison of 95% confidence ellipse parameters. Asterisk (*) denotes real-world estimate.

Model	width	height	angle	radius	radius* [cm]
Schemkes	0.51	0.33	160	0.21	1.94
R. Forest	0.28	0.27	-177	0.14	1.27
XGBoost	0.26	0.25	146	0.13	1.19
KNN	0.31	0.27	171	0.15	1.36
Lasso	0.27	0.26	-154	0.13	1.23

Sample prediction visualizations are shown in Figure 10 across three cases: variable predictions, successful model agreement, and unsuccessful predictions. It is important to note that the circle centers are placed at the model predictions for the sake of illustrating the predictions, the circles do not represent the prediction area. The actual 95% confidence circle would be placed center on the ground-truth to indicate the area where 95% of the predictions would lie. As highlighted in Figure 10c, all circles do not contain the ground-truth Hypothenar keypoint, indicating an anomalous case. Figure 10b shows a success case of predictions across all models. Despite overall worst accuracy, there are cases where Schemkes interpolation performs the best amongst the models as shown in Figure 10a where the other models, although still within circle bounds, predict towards the edge of the hand.

6 Discussion

This study demonstrates the feasibility of inferring the hypothenar keypoint despite limited data. XGBoost and Lasso

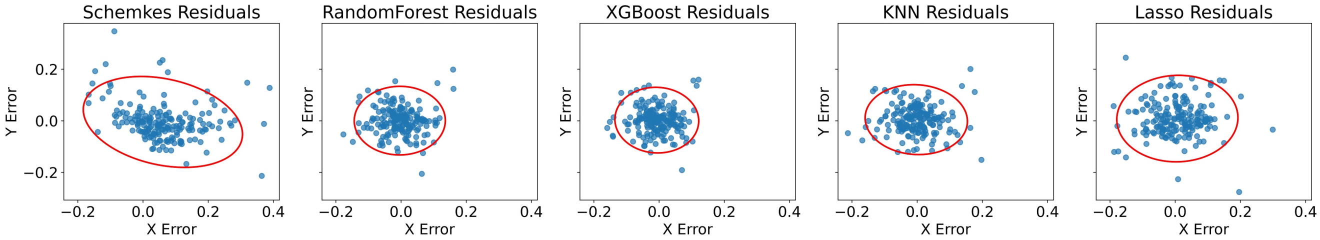
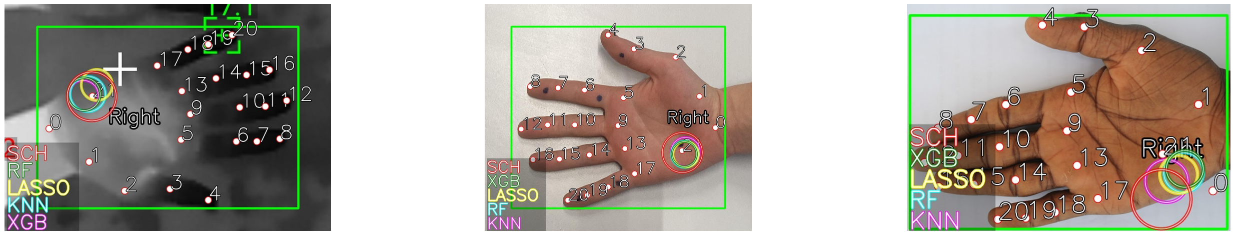


Figure 9: Hypothenar keypoint prediction residual errors with 95% confidence ellipses highlighted.



(a) Model disagreement.

(b) Model agreement.

(c) Low-quality input.

Figure 10: Comparison of Hypothenar Keypoint (21) Prediction with 95% Confidence Circles

regression, applied to MediaPipe outputs, proved most effective for inferring the hypothenar region. Lasso's performance highlights the linear spatial relationships between keypoints, primarily using the y-value of e.g. PINKY_MCP. For XGBoost, x-values were most important, likely due to the inclusion of both hands and faces in the data.

Consistent with Schemkes [9], a generalized interpolation model using average ratios outperformed interpolation with original ratios. The best interpolation model used WRIST and PINKY_MCP, aligning empirically with Schemkes' choice [9].

Dataset shift impacted thumb keypoint prediction accuracy for XGBoost and Lasso by 0.34 cm and 0.3 cm, respectively. The thumb experiment suggests a potential 35% loss reduction with more data for XGBoost, though increased Hypothenar Set loss indicates overfitting. While not directly applicable, the observed MGE could serve as a lower bound for generalization error on new data. XGBoost achieved the best result of 1.19cm but might be prone to overfitting.

6.1 Limitations

Normalization in pre-processing limits real-world conclusions. Ideally, loss should have been calculated after transforming keypoints back to the original space, and an alternative loss function like Google's palm-normalized MNAE could have been used for comparison. However, these choices do not negatively impact prediction accuracy.

MediaPipe annotations of the HPI dataset were sometimes poor, as seen in WRIST keypoint detection in Figure 10c. Manual annotations for infrared images and the hypothenar keypoint were not validated with MediaPipe. Lower contrast in some infrared images increased annotation difficulty. Since this dataset served as ground truth, it could negatively impact the results.

7 Conclusions and Future Work

This study investigated the feasibility of adapting existing hand landmark detection models and datasets to predict keypoints essential for leprosy diagnosis. The results demonstrate the capability of achieving sufficiently high accuracy in predicting the hypothenar keypoint, even with limited training data. Among the models evaluated, Lasso regression proved effective in capturing general relationships with smaller data volumes, while XGBoost achieved the highest accuracy at larger scales.

The contribution of this study lies in improving a component of a larger automatic early leprosy detection solution. If methods for detecting hand keypoints in infrared images, such as adapting existing tools like MediaPipe, achieve sufficient accuracy, the inference of the hypothenar keypoint enables tracking of regions not natively supported by such tools. This would facilitate automatic temperature measurements using smartphones and infrared cameras, offering a practical solution for aiding leprosy diagnosis in resource-limited settings.

7.1 Recommendations for Future Research

Future research could focus on several areas. First, an investigation into data imputation and correction using anomaly detection on missing or incorrectly detected keypoints by MediaPipe such as the wrist keypoint as observed in this study. Second, investigation of methods to obtain and detect hand keypoints in infrared images. Third, future research should address the limitations of this study by incorporating a more diverse dataset that includes hands with missing fingers or deformities for patients with leprosy relapse.

8 Responsible Research

This research has several ethical considerations. Firstly, it relies heavily on the fairness of the MediaPipe model. While the MediaPipe model card addresses fairness considerations [27], Google advises against its use in life-critical applications. This is acceptable in this context, as the study focuses on preliminary assessment as a potential indicator for diagnosis, not as a definitive diagnostic tool.

However, a crucial limitation is the lack of diversity in the datasets. Hands with missing fingers or deformities, which are potential manifestations of leprosy, were not included. This exclusion raises concerns about the accessibility and generalizability of the proposed method.

Regarding transparency and reproducibility, the code will be added to TU Delft repository containing all exploratory and data visualization steps. Furthermore, the creation of the dataset gathered consent from the subjects and followed ethical practices. The datasets used in this are under licenses that permit their adaptation and use in non-commercial settings.

Given the time limitation, the literature review and the research process might not be exhaustive or comprehensive enough to influence decisions in development of healthcare tools. Nonetheless, this study addresses a research gap of a global healthcare problem and can serve to provide a foundation for future work in improving healthcare tools.

Lastly, large-language models (LLMs) were used during the writing of this paper. Gemini 1.5 Pro Flash was used for summarization and length reduction of paragraphs and aided the configuration of code required for plotting data and visualisation. All design decisions and technical consideration were made without the use of LLMs. Sample prompts include ”Rewrite this paragraph to reduce redundant definition and descriptions of [...]” or ”Given this plotly code:[...], increase xticks font to 24, increase the legend font size...”.

References

[1] W. H. Organization, *Towards zero leprosy: global leprosy (Hansen's disease) strategy 2021–2030.* World Health Organization, 2021.

[2] J. Bhandari, M. Awais, B. A. Robbins, and V. Gupta, *Leprosy*, J. Bhandari, M. Awais, B. A. Robbins, and V. Gupta, Eds. Treasure Island (FL): StatPearls Publishing, Jan 2025.

[3] E. Wilder-Smith, A. Wilder-Smith, and M. Egger, “Peripheral autonomic nerve dysfunction in asymptomatic leprosy contacts,” *Journal of the Neurological Sciences*, vol. 150, no. 1, pp. 33–38, Sep. 1997. [Online]. Available: https://doi.org/10.1016/s0022-510x(97)05363-x

[4] N. C. Abbot, J. S. Beck, P. D. Samson, C. R. Butlin, P. J. Bennett, and J. M. Grange, “Cold fingers in leprosy,” *International Journal of Leprosy and Other Mycobacterial Diseases*, vol. 60, no. 4, pp. 580–6, Dec. 1992.

[5] A. L. Cavalheiro, D. T. d. Costa, A. L. F. d. Menezes, J. M. Pereira, and E. M. d. Carvalho, “Thermographic analysis and autonomic response in the hands of patients with leprosy,” *Anais Brasileiros de Dermatologia*, vol. 91, no. 3, pp. 274–83, May 2016.

[6] L. Tiago, D. Santos, D. Antunes, L. Tiago, and I. Goulart, “Assessment of neuropathic pain in leprosy patients with relapse or treatment failure by infrared thermography: A cross-sectional study,” *PLoS Neglected Tropical Diseases*, vol. 15, no. 9, p. e0009794, September 2021. [Online]. Available: https://doi.org/10.1371/journal.pntd.0009794

[7] L.-R. Müller, J. Petersen, A. Yamlahi, P. Wise, T. J. Adler, A. Seitel, K.-F. Kowalewski, B. Müller, H. Kenngott, F. Nickel, and L. Maier-Hein, “Robust hand tracking for surgical telestration,” *International journal of computer assisted radiology and surgery*, vol. 17, no. 8, pp. 1477–1486, 2022.

[8] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, “Mediapipe hands: On-device real-time hand tracking,” 2020. [Online]. Available: https://arxiv.org/abs/2006.10214

[9] I. Schemkes, “Semi-automatic temperature analysis based on real-time hand landmark tracking in infrared videos,” Master's Thesis, Delft University of Technology, July 2024, thesis is confidential and cannot be made public until July 10, 2026. [Online]. Available: http://repository.tudelft.nl/

[10] Y. Li, X. Wang, W. Liu, and B. Feng, “Pose anchor: A single-stage hand keypoint detection network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2104–2113, 2020.

[11] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” in *CVPR*, 2017.

[12] H. Dutta, M. K. Bhuyan, R. Karsh, S. Alfarhood, and M. S. Safran, “Multiscale attention-based hand keypoint detection,” *IEEE Transactions on Instrumentation and Measurement*, 2024.

[13] R. Khanam and M. Hussain, “Yolov11: An overview of the key architectural enhancements,” 2024. [Online]. Available: https://arxiv.org/abs/2410.17725

[14] Z. Cao, Z. Cao, G. Hidalgo, G. Hidalgo, T. Simon, T. Simon, S.-E. Wei, S.-E. Wei, Y. Sheikh, and Y. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” *arXiv: Computer Vision and Pattern Recognition*, 2018.

[15] R. Dsilva, “Hand Keypoint Dataset [26K],” 2024. [Online]. Available: https://www.kaggle.com/datasets/riondsilva21/hand-keypoint-dataset-26k

[16] A. Kapitanov, K. Kvanchiani, A. Nagaev, R. Kraynov, and A. Makhliarchuk, “Hagrid – hand gesture recognition image dataset,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 4572–4581.

[17] MediaPipe, “Hand landmarks detection guide,” https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker, 2024.

[18] G. Vdoviak and T. Sledevič, “Enhancing keypoint detection in thermal images: Optimizing loss function and

real-time processing with yolov8n-pose," in *2024 IEEE 11th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 2024, pp. 1–5.

[19] Z. Zhu, W. Dong, X. Gao, A. Peng, and Y. Luo, "Towards human keypoint detection in infrared images," in *Neural Information Processing*, M. Tanveer, S. Agarwal, S. Ozawa, A. Ekbal, and A. Jatowt, Eds. Singapore: Springer Nature Singapore, 2023, pp. 528–539.

[20] G. Park, T.-K. Kim, and W. Woo, "3d hand pose estimation with a single infrared camera via domain transfer learning," *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 588–599, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229309718

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[22] V. Lepetit, P. Lagger, and P. Fua, "Randomized trees for real-time keypoint recognition," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 775–781 vol. 2.

[23] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016. [Online]. Available: https://doi.org/10.1007/s11749-016-0481-7

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[25] D. Nielsen, "Tree boosting with xgboost-why does xgboost win" every" machine learning competition?" Master's thesis, NTNU, 2016.

[26] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, *Dataset Shift in Machine Learning*, ser. Neural Information Processing series. MIT Press, 2022. [Online]. Available: https://books.google.nl/books?id=MBZuEAAAQBAJ

[27] Google, "Model card: Hand tracking with fairness indicators," https://storage.googleapis.com/mediapipe-assets/Model%20Card%20Hand%20Tracking%20(Lite_Full)%20with%20Fairness%20Oct%202021.pdf, Oct. 2021, accessed [Date you accessed it].

[28] F. Amujo, "Human palm images," Kaggle, Aug. 2023, version 1. [Online]. Available: https://www.kaggle.com/datasets/feyiamujo/human-palm-images

[29] M. Afifi, "11k hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20 835–20 854, 2019.

[30] H. Khazri, Z. Mustapha, S. Shimmi, M. T. Hossain Parash, and A. B. M. Hossain, "Hand anthropometry: Baseline data of the major ethnic groups in sabah," *Borneo Journal of Medical Sciences (BJMS)*, vol. 17, 01 2023.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

A Appendix

A.1 Data Processing Definition

A hand is represented by its keypoint set. Let $P = \{p_i\}_{i=1}^{n}$ be a set of n points on the hand, where each point is represented as:

$$p_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad \text{with } 0 \leq x_i \leq 1 \text{ and } 0 \leq y_i \leq 1. \tag{3}$$

The width (w_{bb}) and height (h_{bb}) is padded by a constant C to make it more consistent with datasets that already contain bounding box data. The bounding box is defined by its centre and dimensions as follows:

$$\begin{aligned} w_{bb} &= \max_i\{x_i\} - \min_i\{x_i\} + C_x \\ h_{bb} &= \max_i\{y_i\} - \min_i\{y_i\} + C_y \\ x_c &= \min_i\{x_i\} + \frac{w_{bb}}{2} \\ y_c &= \min_i\{y_i\} + \frac{h_{bb}}{2} \end{aligned}$$

Using the bounding box, the scaling matrix S and translation vector t are defined. Thus, translated and scaled point p_i' is obtained by:

$$p_i' = Sp_i + t = \begin{bmatrix} \frac{2}{w_{bb}} & 0 \\ 0 & \frac{2}{h_{bb}} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} -\frac{2c_x}{w_{bb}} \\ -\frac{2c_y}{h_{bb}} \end{bmatrix} = \begin{bmatrix} x_i' \\ y_i' \end{bmatrix} \tag{4}$$

The vector from the translated and scaled wrist to the translated and scaled middle finger knuckle is defined as:

$$v_1 = p_m' - p_w'$$

Let v_2 represent the basis vector of the positive y-axis. The angle θ between v_1 and v_2 is then calculated as:

$$v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \theta = \arccos\left(\frac{v_1 \cdot v_2}{\|v_1\|}\right),$$

The angle θ defines the rotation matrix R which results in the final transformed point p_i'' (translated, scaled, and rotated) as follows:

$$p_i'' = Rp_i' = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x_i' \\ y_i' \end{bmatrix} = \begin{bmatrix} x_i'' \\ y_i'' \end{bmatrix} \tag{5}$$

$$\text{with } -1 \lessapprox x_i'' \lessapprox 1 \text{ and } -1 \lessapprox y_i'' \lessapprox 1.$$

The result is a set of normalized keypoints $P'' = \{p_i''\}_{i=1}^{n}$. This transformation is applied to each hand as part of preprocessing.