

## Detailed Design

### Architecture

The architecture that best suits our project is Client-Server.

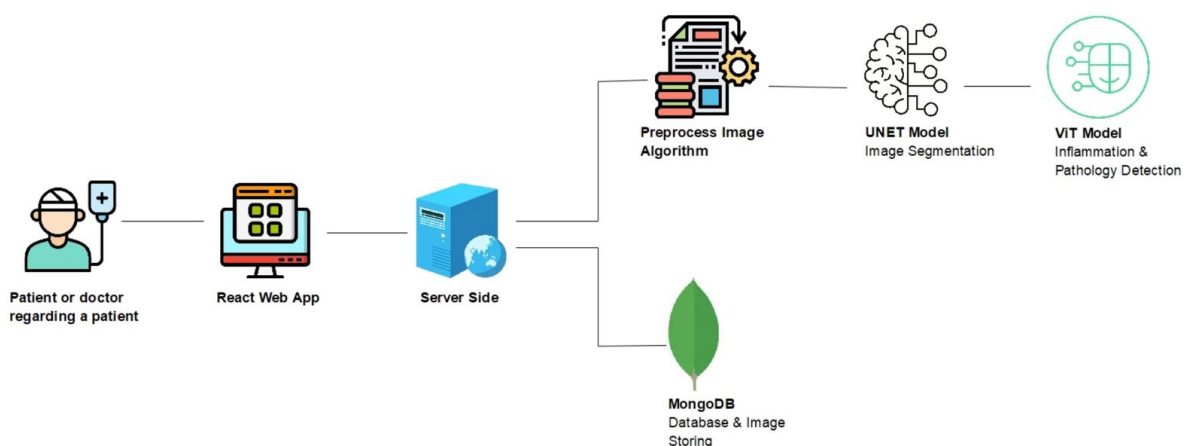
Client-Server Architecture in our project:

- Server:
  - ❖ Manages data and images in a database in MongoDB.
  - ❖ Handles the processing of the images sent from the client.
  - ❖ Send data between the database and the processing model.
- Client:
  - ❖ Uploads images for processing and downloads reports from server.
  - ❖ The client sends his requests through a web app.

### Data Storage

- The user's data is stored in JSON.
- Images and trained ML models will be stored in files.
- Image sets' data will be stored in JSON as well.

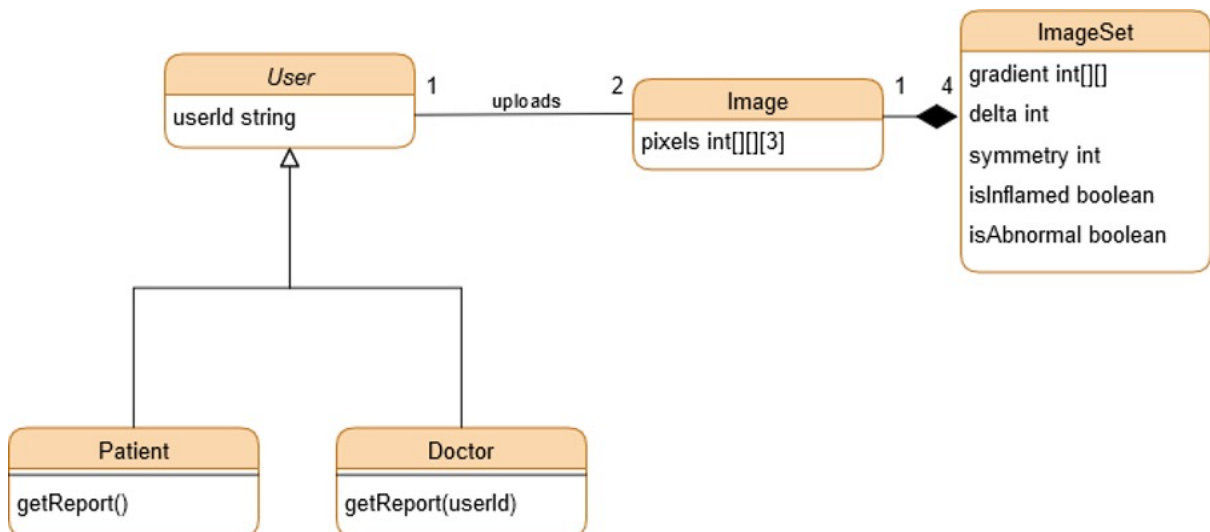
### Graphic Description



## Data Description

- User – a JSON file that has the following fields:
  - ❖ userID
  - ❖ whether the user is patient or doctor
- Image sets' data – will be saved as JSON with the following fields:
  - ❖ Delta
  - ❖ Gradient
  - ❖ Symmetry
  - ❖ Any other data that we will decide in the future is relevant for the model.

## Graphic Description



## API specification

- Upload optical and thermal images of hands.
- Download reports based on past images sent to the website.
- Download reports based on `userId`, if the user is a doctor.
- Display prediction results of the ML model.
- Upload the images to the database for training the model.

## **Programming languages and tools**

We have selected Python for making statistics analysis and for the vast libraries for machine learning. We chose to build our web application using React because of its low learning curve.

## **Algorithm Description**

### ***Non-Machine Learning:***

#### **Hand Registering Algorithm:**

**Algorithm:** *To register the hands in the thermal image using the optical image, the algorithm aligns both images through feature-based matching. First, ORB (Oriented FAST and Rotated BRIEF) feature matching is used to detect and match key points between the thermal and optical images. Once matched, a homography transformation aligns the thermal image to the optical image's perspective. This ensures that the thermal image is accurately mapped onto the optical image, preserving anatomical details for further processing.*

**Time Complexity:** The time complexity of ORB feature matching is  $O(n + M^2)$ , where  $n$  is the number of pixels in the image and  $M$  is the number of keypoints detected, because the complexity of the feature detection is  $O(n)$  and the complexity of brute force feature matching is  $O(M^2)$ . The complexity of the homography transformation is  $O(KM)$ ,  $M$  still being the number of keypoints detected and  $K$  is the number of iterations the estimation computes, however  $K$  will always be between 100 to 200 and most likely will be hardcoded and that's why it won't increase the complexity. In conclusion, the time complexity will be  $O(n + M^2)$ , depending either linearly on the number of pixels in the image or exponentially on the number of keypoints detected.

### ***Machine Learning:***

#### **1. Convolutional Neural Network – UNET:**

**Algorithm:** *UNet is a CNN architecture designed for image segmentation, especially in medical imaging. It has a U-shaped structure with an encoder that extracts*

*features, a bottleneck for deep representations, and a decoder that restores spatial details. Skip connections help retain localization accuracy, making UNet effective for pixel-wise segmentation, even with limited data.*

**Time Complexity:** Since UNet has both a contracting path (downsampling) and an expanding path (upsampling), the overall complexity remains in the order of  $O(n)$  per layer, leading to  $O(D*n)$  for the full network, where  $n$  is the number of pixels in the image and  $D$  is the depth of the network.

## 2. **Transformer – Vision Transformer:**

**Algorithm:** *The Vision Transformer (ViT) is a deep learning architecture designed for image analysis using self-attention mechanisms instead of traditional convolutional layers. It splits an image into fixed-size patches, flattens them, and embeds them with positional encodings before passing them through a Transformer encoder, originally used in NLP. This allows ViT to model long-range dependencies and global relationships more effectively than CNNs. While ViT requires large datasets for training, it achieves state-of-the-art performance in image classification, segmentation, and object detection, often outperforming CNNs in capturing complex patterns.*

**Time Complexity:** The time complexity of a Vision Transformer (ViT) is primarily determined by the self-attention mechanism, which operates on image patches. The dominant computational cost comes from multi-head self-attention (MHSA), which has a complexity of  $O(M^2 * d)$ , where  $M$  is the number of patches and  $d$  is the embedding dimension. Unlike CNNs, where complexity scales linearly with image size, ViT's self-attention scales quadratically with the number of patches.

### Graphic Description

