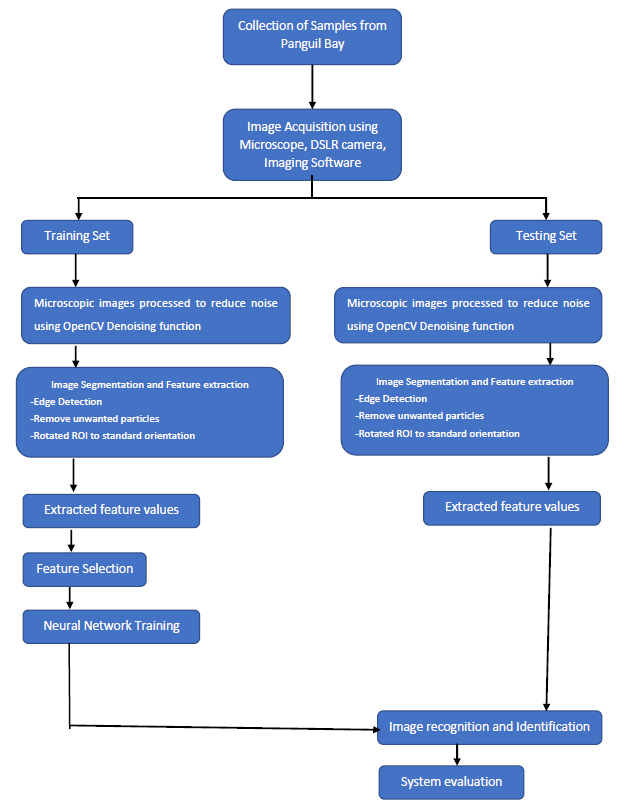
**Chapter 4: Methodology**



**Figure 4.1** Process Flowchart

**4.1 Sample Collection and Image Acquisition**

Five genera of marine copepods commonly encountered in mangrove waters will be examined: Acartia (A. spinicauda), Bestiolina (B. similis), Oithona (O. aruensis, O. dissimilis and O. simplex), Parvocalanus (P. crassirostris) and Tortanus (T. barbatus and T. forcipatus). Copepods will be sampled from four stations from the upper estuary in the Panguil bay to near shore waters of Marandin Lala Lanao Del Norte. Horizontal plankton tows (0.5-1 m depth) using paired 45 cm-diameter bongo nets (180 μm) will be made and collected plankton will be preserved in buffered 10% formaldehyde. In the laboratory, collected copepods will then be sieved through stacked Endecott sieves of 1,000 μm, 500 μm, 250 μm and 125 μm mesh sizes, and the sieved fractions will be preserved in 80% alcohol in individual vials for a long-term preservation.

**4.2 Image acquisition**

Specimens of copepod will be randomly pipetted onto a microscope slide from the preserved samples and each identified to species level under a compound microscope (Olympus BH2). To enable the dorsal aspect of the identified copepod to be imaged, often the copepod body had to be rotated. Body rotation could be easily achieved by first placing two short nylon fishing lines (0.36 mm diameter) on either side of the specimen and gently moving a cover slip placed over them by using the tip of the index finger. The desired view of the copepod body will be acquired by an Digital camera connected to a computer installed with an imaging software (Olympus cellSens Standard ver. 1.12) for real-time viewing, capturing and storing of the images. The built-in function in cellSens called Extended Focus Imaging (EFI) will be used to create a single plane image with sharp, in-focus details and high contrast. The EFI function recorded the image data as the sample was gradually focused through from top to bottom to obtain single dorsal image of the copepod with all body parts. Besides, the contrast and brightness of the images were set to the best before they were captured using cellSens software. The resolution of the captured images was standardised (2448 × 1920 pixels) and all the images were saved in uncompressed Tagged Image File Format (TIFF) by renaming them according to the date when the images were captured.

**4.3 Image storage**

A simple database will be made where the image will be organized according to taxa and verified by plankton expert. Thirty images for each species will be stored as a training set and twenty will be for testing set.

**4.4 Image Processing**

**4.4.1 Image Pre-processing**

Images will be converted to binary using Otsu’s binarization method with proper threshold. Images will then be converted to 2D grayscale image using OpenCV cvtColor() function. Then noise will be removed using morphologyEx() function.

**4.4.2 Image Segmentation**

Image segmentation will be done using the marker-based watershed segmentation algorithm by OpenCV where the regions with one color, background region, or regions with another color, or unsure regions will be labelled/marked with 0 beforehand. After doing so, apply the watershed algorithm then the markers will be updated and the boundaries will have a value of -1.

The image will then be converted to binary image with appropriate threshold and the borders will be cleared. The holes that occurred during the conversion from grayscale to binary will be filled using imfill() in OpenCV. Particles that are less than 50000 pixels will be excluded to ensure that only copepods will be segmented in the image.

The orientation represented by the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the region of interest (ROI) will be obtained using Contour properties function in OpenCV. Image rotation will be done using the getRotationMatrix2D function so that the ROI will have an orientation of 90 degrees.

**4.4.3 Feature Extraction**

The ROI of the copepod will be cropped by getting the coordinates of the boundary of copepods.

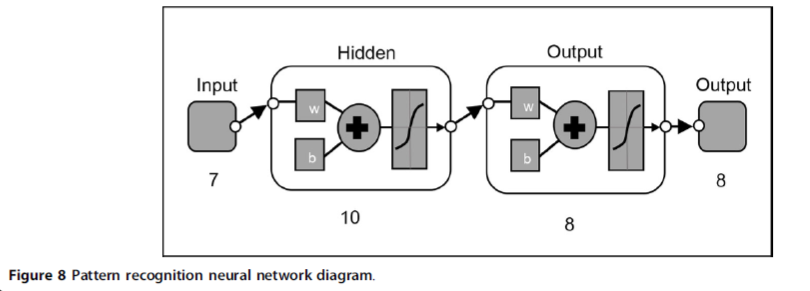
Features will be extracted from the shape descriptors represented by the binary images of the ROI using Contour properties function in OpenCV. The measurement to be taken are area, convex area, eccentricity, major axis length, minor axis length, perimeter, solidity, equivdiameter (sqrt(4\*area/pi)), extent and orientation.

**4.7 Feature Selection**

To avoid overfitting in the Neural Network training and to increase performance, not all the 11 extracted features will be used. The extracted features will be evaluated to make sure that only significant features will be selected to classify the copepods into their respective taxa. Forward stepwise discriminant analysis (FSDA) was used to aid the selection of the most useful features (StatSoft Inc.). In order to visualize how well a selected feature clustered the specimens in the training set into the eight classes (species), 2D and 3D scatter plots will be graphed with different combinations of features as the axes.

**4.9 Artificial Neural Network Training and Performance Evaluation**

An Artificial Neural Network (ANN) will be used as the pattern recognition tool to classify the extracted features values into the eight classes (species). The architecture of the ANN is a two-layer feed-forward network with sigmoid hidden (ten nodes) and output (eight nodes) neurons and the network will be trained with scaled conjugate gradient backpropagation. A total of 240 sample images will be used in the training set with 30 images from each class. The input data for the input nodes of the network will have seven selected features of each specimen from the training set, whereas the target data defined eight desired output classes. The 240 samples will be divided into three sets, the training set (168 samples, or 70% of samples), validation set (36 samples, 15%) and testing set (36 samples, 15%). The data from the training set will be used for network training; the validation set for measuring network generalization and terminating training before overfitting; and the testing set for independent measure of network performance during and after training. The performance of the network training will be evaluated using Mean Square Error (MSE) and confusion matrices. The training stopped when the MSE of the samples in the validation set started to increase indicating that the network generalization stopped improving. The network will be trained several times to get the trained network with best performance. Another 160 independent samples (20 samples for each species) will be used for system performance evaluation. The trained network will be simulated using the testing data as input and the output will be then compared to the predicted data and recorded in a confusion matrix.



**Figure 4.3** Pattern recognition neural network diagram