

# UNIVERSITI TEKNOLOGI MALAYSIA

## MODULE 5 RESEARCH METHODOLOGY



# Main Body of the thesis

~~○ Introduction~~

~~○ Literature Review~~

○ Methodology

# Research Methodology

## Why you need to write Methodology chapter?

- The methodology chapter explains **WHAT** you did and **HOW** you did the research you conduct.



# Research Methodology

## What is the purpose of a Methodology?

- Explains the research flow.
- Present the proposed method that leads the research novelty.



# Methodology

## Recommended sub-topics

- Research Framework
- Problem Formulation
- Dataset
- The Proposed Solution (Algorithm/flowchart/step etc.)
- Performance Measurement

# Research Framework/Flow

- Explain all phases of research implementation

# Research Flow

## Example 1

The necessity for conducting numerous assessments can contribute to increased operating expenses for medical practitioners. As a result of the greater operating costs, ophthalmologists' knowledge and ability in managing the tests must still be taken into account. Perhaps a greater consistency of clinical evaluations has resulted from the introduction of quantitative grading scales and the insertion of images to enhance the scales. It has been suggested that automated grading systems that estimate redness based on physical attributes be used in order to avoid the subjectivity involved in grading because the variability between evaluations can be fairly large. Researchers were inspired to develop computational methods to quantify pterygium difficulties in order to solve these issues. Figure 3.1 shows the flow of proposed eye redness grade classification framework.

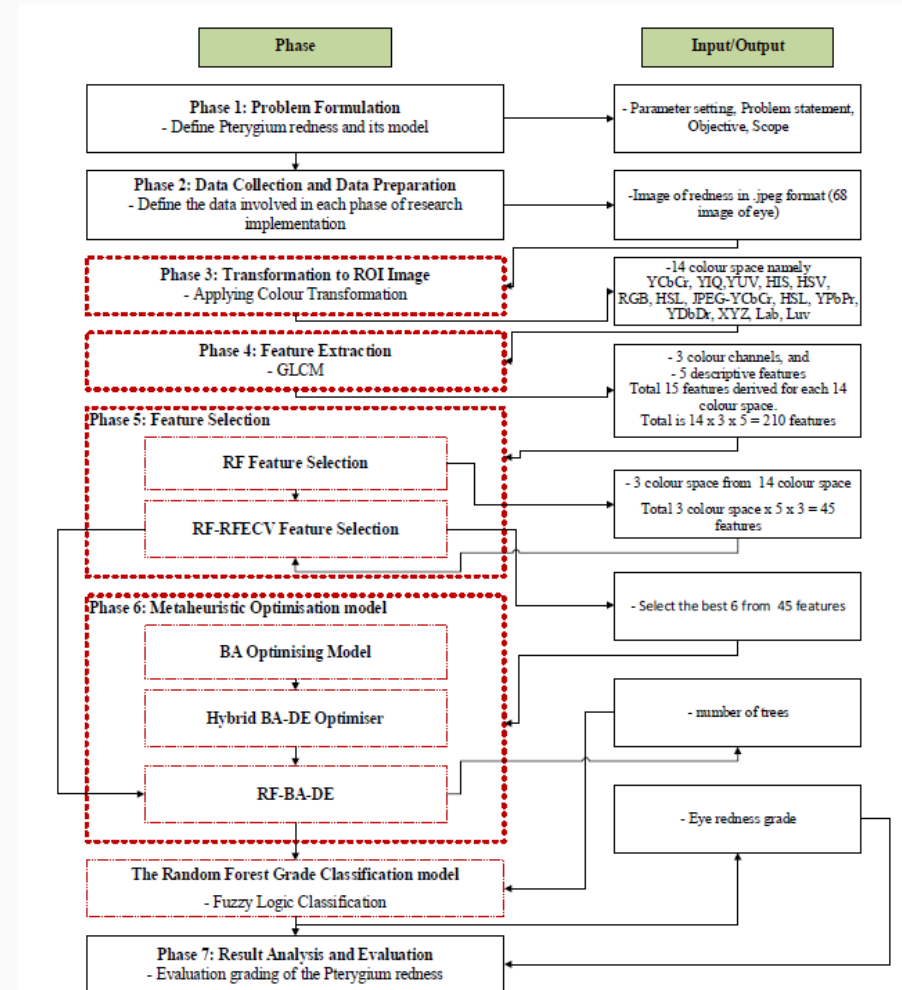


Figure 3.1 The flow of proposed eye redness grade classification framework



# Research Flow

## Example 2

In the filter phase, Ensemble Filter will be used to disregard redundant and irrelevant features in the HAR datasets. The relevant features that give high classification accuracy with an RF classifier are identified to form a new subset named the Ensemble Filter dataset. The Ensemble Filter is applied in order to remove irrelevant and features and reduces the computational load for the HSABC algorithm and RF classifier. The wrapper selection phase and the classification phase will be discussed next. The general flow of this research is shown in Figure 3.1 below.

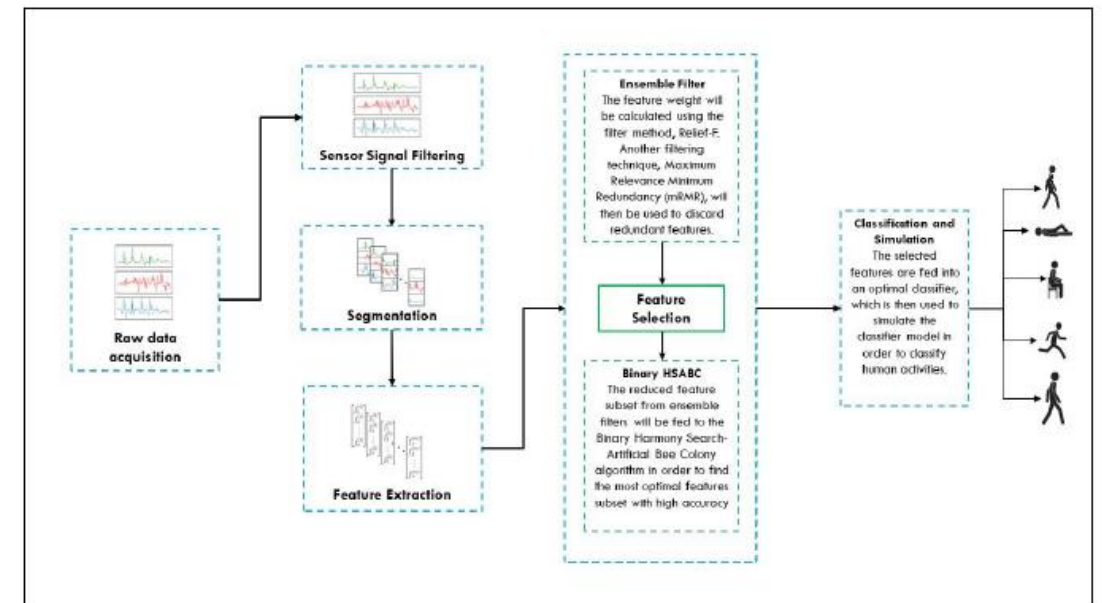


Figure 3.1 Research Flow



# Research Flow

## Example 3

### 3.2 Research Framework

The research work has been framed into three major phases; these three phases are shown in Figure 3.1. The first phase is analysis phases; in this phase the extensive literature review on the hybrid global solar radiation (GSoR) forecasting model using deep learning (DL) is carried out which consequently leads to problem formulation. The problem formulation is based on the consideration of external factors which influences GSoR; depth of evaluation, evaluation metrics, and comparison in different locations; the need to integrate many data sources for accuracy improvement; and the limitations of using LSTM and MLP DL techniques. In GSoR forecasting, the accuracy of prediction is one of the major issues because due to external atmospheric variables such as cloud and amount of aerosols, the radiation is affected which in turn exerts negative effect on the output of any solar generating plant. The proposed HYB-CNN-LSTM-MLP algorithm has ability to make reliable forecast in consideration of the many external atmospheric factors.

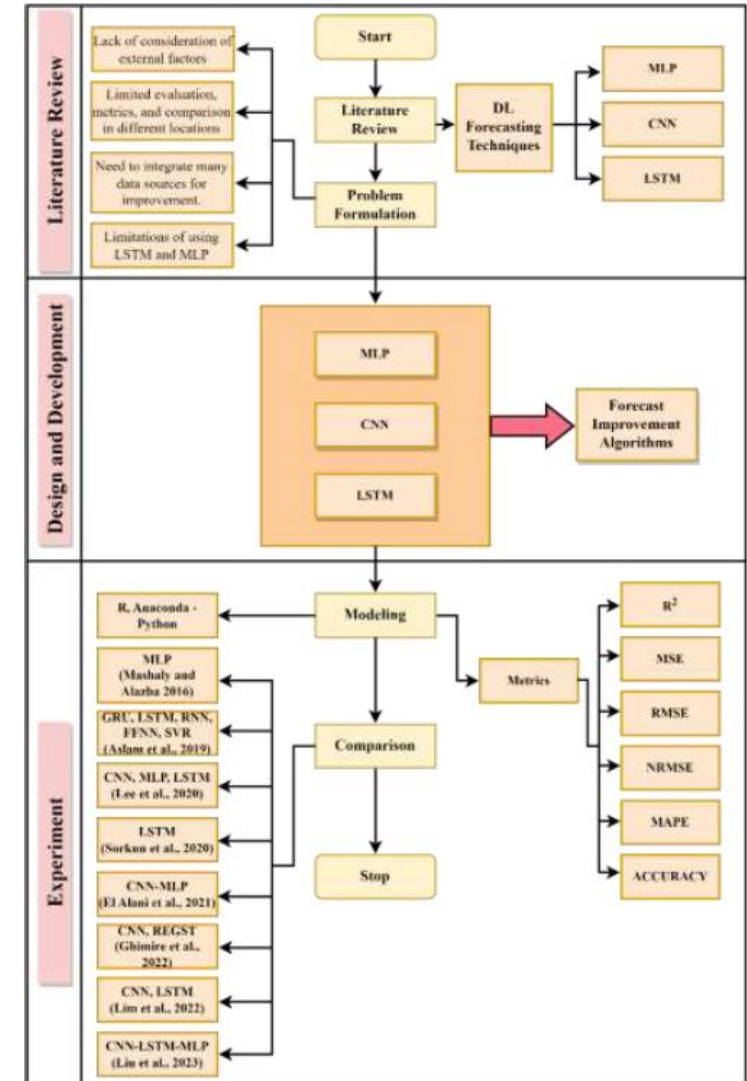


Figure 3.1 Flow chart of the research framework

# Research Flow

## Example 4

### 3.3 Overview of Research Framework

The research work is delineated into three primary stages, outlined in Figure 3.1. In the initial stage, termed the analysis stage, the focus lies on summarizing existing microarray feature selection methods, identifying their limitations, and formulating research questions. These limitations encompass challenges associated with direct classification of microarray datasets, constraints within current feature selection methodologies, hurdles in applying heuristic algorithms to microarray feature selection, and shortcomings of existing classifiers when applied to microarray data.

Moving to the second stage, a novel microarray feature selection method grounded in the AFS algorithm is proposed. This method aims to enhance efficiency in feature selection and elevate classification accuracy in cancer classification experiments. Key steps in the design of this method include refining the AFS algorithm

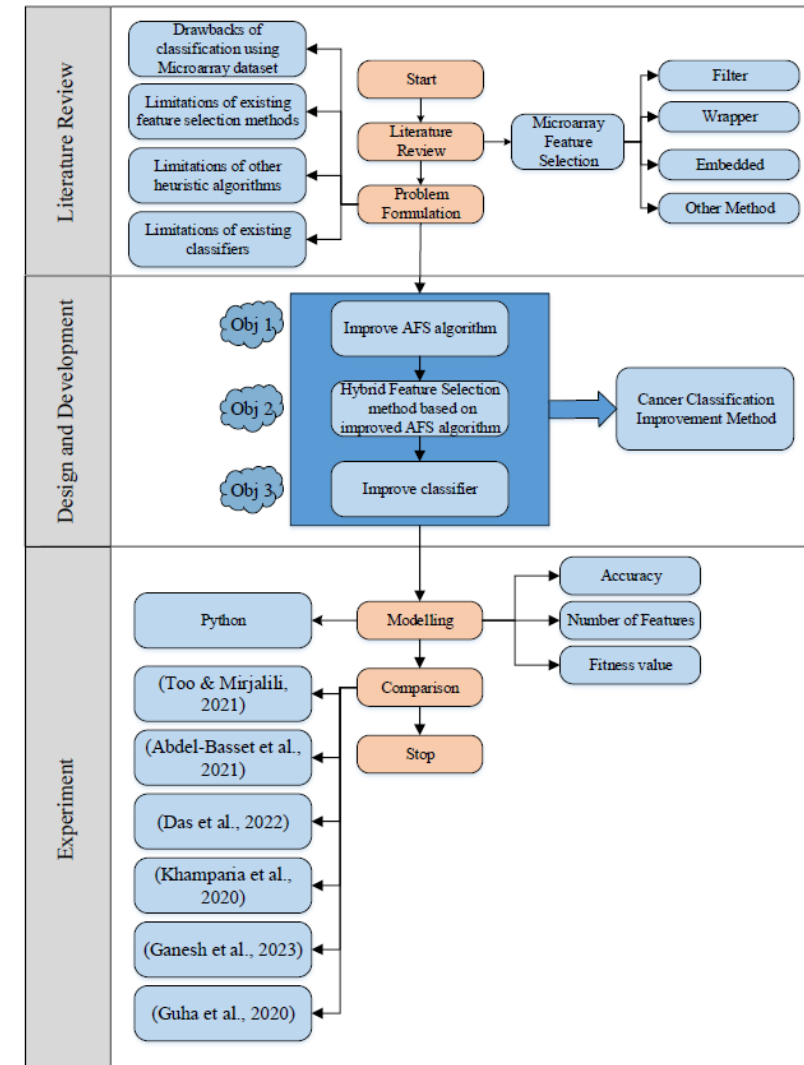


Figure 3.1 Flow chart of the research framework

# Problem Formulation

- Explain the connection between Research Questions, Research Objectives and Proposed Solution

# Problem Formulation

## Example 1

Table 3.1 Research Objectives and proposed solutions

No.	Research Questions	Research Objectives	Proposed Solutions
1	How effectively extract features from Pterygium eye redness images using multiple colour spaces and texture features, and how does this impact the accuracy of eye redness classification grade?	To systematically derive multiple colour spaces, colour channels, and features from texture features and explore their integration within the Gray Level Co-Matrices (GLCM) model through the application of descriptive statistics.	The development of comparative study with statistical analysis on the features extracted using descriptive statistics and GLCM model. Analyse the distribution of each feature and how they relate to each other and to the ocular condition being diagnosed.
2	Can the enhanced RF-RFECV feature selection model reduce the dimensionality of derived colour channels without sacrificing accuracy in machine learning models trained on Pterygium eye redness images?	To enhance RF by hybridizing with RFECV as RF-RFECV feature selection model in selecting best feature from the derived colour channel.	Conduct a study to extract features from derived colour channels using the enhanced RF-RFECV feature selection model and train machine learning models using these features. Compare the accuracy of these models to those trained using the full set of features and those trained using standard RF model.
3	How to improve the accuracy of the classification of Pterygium eye redness grade using a hybrid RF-BA-DE metaheuristic model, and how does this compare to other state-of-the-art classification models?	To improve the proposed RF-RFECV model by hybridizing with BA and DE as RF-BA-DE metaheuristic classification model of eye redness grade.	Conduct a study to compare the accuracy of the hybrid model to these traditional models. Evaluate the models using the benchmarks results and performance measurement metric.

# Problem Formulation

## Example 2

Table 3.2 Summary of research questions

Research Question	Research Objectives	Proposed Solutions
How to develop a robust AFS algorithm for high-dimensional problems taking into account factors such as step size and initialization and verified in different high-dimensional optimization problem areas?	Develop a novel AFS algorithm to solve the unique challenges posed by high-dimensional data sets in the biomedical field.	Introduce heuristic information or integrate domain knowledge to refine the movement direction of fish schools, making the algorithm more suitable for microarray data feature selection tasks.
How to design a feature selection method based on objective 1, and use it to solve microarray data sets feature selection?	A new hybrid method is designed based on the improved AFS algorithm and combined with other feature selection methods.	Combining filter and wrapper feature selection method, a double-level feature selection framework is constructed.
How to propose a novel classifier to perform classification tasks on microarray datasets to achieve high performance?	Design a new KNN classifier improvement scheme to improve its classification accuracy and generalization ability in microarray data classification.	Different weights are assigned according to the importance or similarity of different samples, thereby improving the performance of the KNN classifier in sample classification tasks.

# Dataset

- Explain the data source
- Dataset description

# Problem Formulation

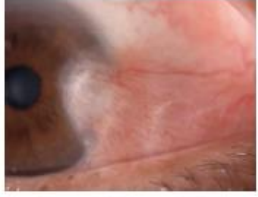
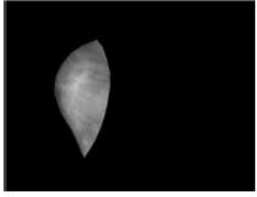
## Example 1

### 3.3 Phase 2: Data Collection and Data Preparation

This research adopts purposive sampling as method of recruitment. Consent form was given to all qualified participants. All participants were screened in order to ensure they fulfil the inclusion criteria needed. All participants involved in this research were on voluntary basis. Data is collected from historical data from IIUM Eye Specialist. Minimum number of 60 participants required to achieve the level of significance (using power of 90% and p-value of 0.05). Hence, a minimum of 60 participants were set in order to achieve the targeted level of significance. Specifically for this thesis, it was conducted by the recommendations of the tenets of the Declaration of Helsinki and received approval from the International Islamic University Malaysia (IIUM) Research Ethical Committee (IREC) (IIUM/310/G13/4/4-125). All participants obtained written and informed consent before any procedures were performed. The datasets were gathered at IIUM, and in particular, there are 68 participants involved. These data are in the format of .mat file. Due to critical importance of maintaining data integrity, confidentiality, and privacy. Access to these datasets will be limited to the authorized personnel involved in the approved research project. Appendix A shows the letter of approval and certification of datasets of the eye image and its ROI image that represents three categories namely atrophy, intermediate and fleshy.

The 68 images utilised in this investigation are displayed in Appendix A. It shows the ROI.jpeg image and redness grade along with the original eye redness.jpeg photographs. One example of an eye image, together with its ID, ROI image, and redness grade, are shown in Table 3.2.

Table 3.2 Eye redness image and its relevant information

Image ID	Original .jpeg image (slit-lamp biomicroscopic SLB)	ROI .jpeg image	Pterygium redness Grade Scale
1			1.5



# Problem Formulation

## Example 2

### 3.3 Datasets

In this research, two publicly available physical activities accelerometer sensor datasets are utilized: SBHAR and USC-HAD. A more elaborated detail for each data set is explained in the next following section. Table 3.1 briefly described the general information of the datasets. Both datasets in this research employed accelerometers and gyroscope sensors, which denotes as  $(A_x, A_y, A_z)$  and  $(G_x, G_y, G_z)$  where A and G are referring to Accelerometer and Gyroscope, while  $x,y,z$  referring to the three-dimensional (3D) axes of both sensors.

Table 3.1 SBHAR and USC-HAD datasets

Dataset	Wearable sensors	Sensor Position	No of Activity	No. of Subject
SBHAR	Accelerometers ( $A_x, A_y, A_z$ )	Waist	6	30
USC-HAD	Gyroscope ( $G_x, G_y, G_z$ )		10	14

The location of the sensor is attached at a single body position, which is the waist, as illustrated in Figure 3.2. There is a total of six activities for the SBHAR dataset and ten activities for the USC-HAD dataset employed in this research, with the number of subjects 30 and 14, respectively.

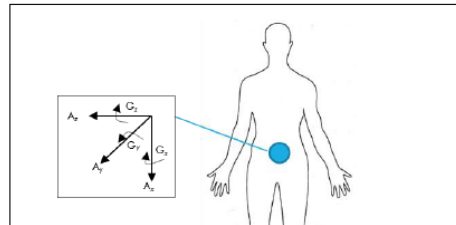


Table 3.4 Activities description

SBHAR Dataset	USC-HAD Dataset
<b>Three Dynamic Activities</b> [A1] Walking– the subject is walking normally [A2] Walking Upstairs– the subject is walking upstairs [A3] Walking Downstairs– the subject is walking downstairs  <b>Three Static Activities</b> [A4] Sitting– subject in a sitting position [A5] Standing– subject in a standing position [A6] Lying Down– subject in a lying down position	<b>Seven Dynamic Activities</b> [A7] – Walking Forward – the subject is walking forward in a straight line [A8] – Walking Left – the subject is walking counterclockwise in a full circle [A9] – Walking Right – Subject is walking clockwise in a full circle [A10] – Walking Upstairs - the subject is walking upstairs [A11] – Walking Downstairs - the subject is walking downstairs [A12] – Jumping – the subject is jumping up and down [A13] – Running – the subject is running in the forward line  <b>Three Static Activities</b> [A14] – Standing - subject in a standing position [A15] – Sitting - subject in a sitting position [A16] – Sleeping – the subject is sleeping or in a lying down position

# Problem Formulation

## Example 3

The data set in the following Table will be used as the test data for this experiment:

Table 3.1 Data set for testing

No.	Datasets	Data URL
1	Diabetes Data Set	<a href="http://archive.ics.uci.edu/ml/datasets/Diabetes">http://archive.ics.uci.edu/ml/datasets/Diabetes</a>
2	Pima Indians Diabetes Data Set	<a href="https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes">https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes</a>
3	Thyroid Disease Data Set	<a href="http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease">http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease</a>
4	Breast Cancer Data Set	<a href="http://archive.ics.uci.edu/ml/datasets/Breast+Cancer">http://archive.ics.uci.edu/ml/datasets/Breast+Cancer</a>
5	Breast Cancer Coimbra Data Set	<a href="http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra">http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra</a>
6	Breast Cancer Wisconsin (Original) Data Set	<a href="http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)">http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)</a>
	Breast Cancer	

# Problem Formulation

## Example 3

### 3.3.1 Data Sources and Data Collection Techniques

#### 3.3.1.1 Study Sites

The study site comprises sixteen (16) locations from Queensland, Australia (Figure 3.3 and Figure 3.4). Australia has been selected because of its potential for solar (Deo and Şahin, 2017), being characterised by high insolation, low cloud coverage, and low rainfall. The sixteen study sites are: Glendower Station, Gregory Springs Station, Toomba Stud, Low Holm Station, Hillgrove Station, Mingela Post Office, Bruslee, Trafalgar Station, Holmleigh, Ulcanbah, Jochmus, Tiree, Woodbine Station, Hillview Station, Katandra, and Cameron Downs. These locations are grouped into four groups (Figure 3.4) based on proximity to the cross-validation sites. In each group, three locations will be used for model training and the fourth for cross-validation.

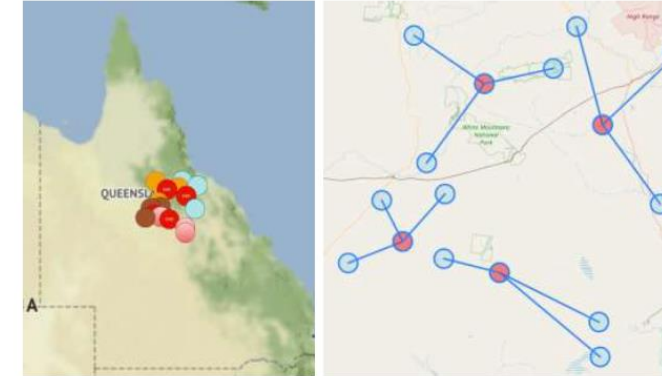


Figure 3.3 Study locations on Queensland's (Australia) terrain map. Figure 3.4 Grouped study locations with the cross-validation sites in red.

The distances between the training sites and cross-validations are calculated and shown in Table 3.2. These distances in kilometres (km) depicts how far is each training location from its corresponding cross-validation site. The grouping of the sites was done according to their nearness to the cross-validation locations. In addition, the corresponding latitude, longitude, and altitude (in metres 'm') of every location is presented (Table 3.2). The altitude represents the elevation of each location above sea level.

Table 3.2 Study sites

ID	Training and Cross-validation Sites	Location (Lat, Lon)	Altitude (m)
0001	Glendower Station (08.21km)	20.7280, 144.4853	420

# The Proposed Solution

- The **fundamental (general) idea** of the modified/improved/enhanced methods, techniques, framework, equations etc.
- The discussion should support with diagram/algorithm/flowchart/step etc.

# The Proposed Solution

## Example 1

### 3.6.3 Hybrid method of microarray feature selection

The hybrid method of feature selection combines the strengths of filter and wrapper methods to achieve more efficient feature selection. Initially, the filter method is typically employed in the pre-processing stage to identify a subset of features highly correlated with the target variable by evaluating and ranking the features. Subsequently, the wrapper method incorporates feedback information from the classifier, assessing feature quality based on classifier performance, and further refining the optimal feature subset. A schematic diagram of microarray feature selection using this method can be depicted as Figure 3.4.

Hybrid method feature selection seamlessly integrates the filter and wrapper methods, utilizing the efficiency and simplicity of the filter method for initial feature screening. It then employs the accuracy and robustness of the wrapper method for in-

advantages of both methods, enhancing feature selection efficiency and performance. It is applicable to feature selection challenges across various fields and complexities.

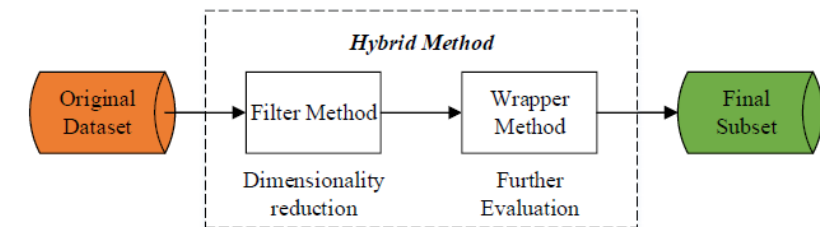


Figure 3.4 Hybrid Method for Microarray Datasets

# The Proposed Solution

## Example 2

The flow chart of the improved HS is as follows, and the main part of the improvement is circled in red (See Figure 3.2). The pseudo code of the red circled part is as follows:

**Algorithm:**

If ( $r < PAR$ ):  $x_{new} = x_{new} + r * bw$   $r \in (0,1)$

Figure 3.3 shows the curve of  $PAR$  with the number of iterations. As the number of iterations increases,  $PAR$  gradually increases from the initial minimum along the curve. When the number of iterations is the largest,  $PAR$  is the maximum, and the change of  $PAR$  gradually tends to be flat, preventing the disturbance and destruction of the better solution vector at the later stage of the iteration. Figure 3.4 shows the curve of  $bw$  with the number of iterations. As the number of iterations increases,  $bw$  gradually decreases from the maximum value according to the curve. When the number of iterations is the largest,  $bw$  is approximately equal to the minimum, and gradually tends to be flat, which can ensure a fine search for areas with better performance in the solution space at the later stage of the iteration.

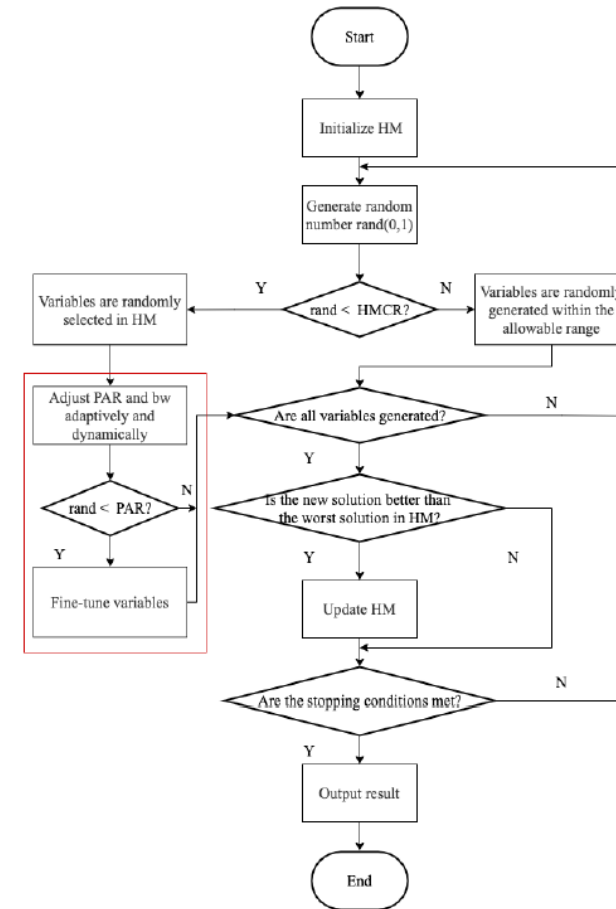


Figure 3.2 Improved harmony search algorithm

# Performance Measurement

- Explain the method used to validate the result



# Performance Measurement

## Example 1

### 3.9 Performance Measurement

In this phase, the result of grading scales has been validated using the statistical analysis. To get the best model of prediction values in pterygium eye redness grading scales, higher value of Coefficient of Determination ( $R^2$  Score) and lower Root Mean Squared Error (RMSE), Maximum Residual Error (maximum error) and Mean Absolute Percentage Error (MAPE) are considered. The  $R^2$  Score, RMSE, maximum error, and MAPE value are calculated as stated in formula in Equation (3.1), (3.2), (3.3) and (3.4):

$$R^2 = 1 - (\text{Residual sum of squares}) / (\text{Total sum of squares}) \quad (3.1)$$

Where:

Residual sum of squares (RSS) is the sum of the squared differences between the predicted values and the actual values.

Total sum of squares (TSS) is the sum of the squared differences between the actual values and the mean of the actual values.

$$\text{RMSE} = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}} \quad (3.2)$$

$$\text{Max Error}(y, \hat{y}) = \max(|y_i - \hat{y}_i|) \quad (3.3)$$

$$\text{MAPE} = \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{N} \right| \times \frac{100}{n} \quad (3.4)$$

# Performance Measurement

## Example 2

For each performance measure used, 0 indicates the worst, while 1 refers to the best results. For MCC, a correlation of 1 indicates perfect agreement, a correlation of 0 indicates little better than random agreement, and a negative correlation of -1 indicates absolute disagreement between prediction and observation. The results of the classifier can be evaluated by using a confusion matrix. The confusion matrix for classification problems is shown in Figure 3.15.

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Figure 3.15 Confusion Matrix

To evaluate the quality of classification, the number of correctly recognized class (TP), the number of correctly recognized examples that do not belong to the class (TN), and the number of examples that were neither correctly assigned class (FP) nor recognized as the class (FN) will be computed. There are several measurement standards that have been defined for the confusion matrix. Table 3.9 below shows the standard measurement for Confusion Matrix.

# Performance Measurement

## Example 3

### 3.4 Performance Evaluation

As stated in the scope of this research, although the best performance metrics in forecasting is MAPE, this study also included other indicators such as MAE, MSE,  $R^2$ , and RMSE. These indicators are used in regression models. Note that the implementation of the MLP, LSTM, and CNN are all in regression form. Suppose  $GSoR_A$  and  $GSoR_F$  represents the actual and forecasted global solar radiation respectively, then the metrics can be defined as follows.

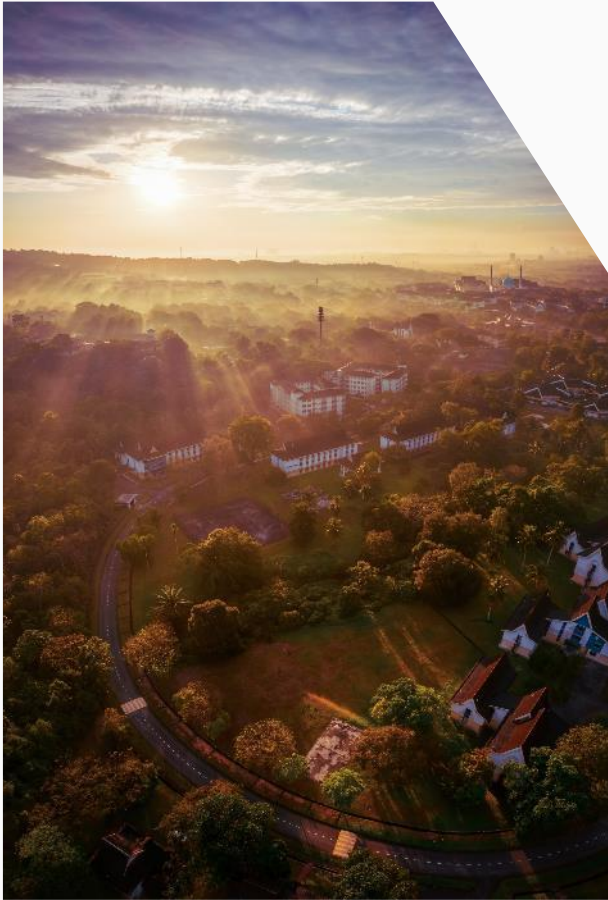
The mean absolute error (MAE) is the simplest error metric. It can be defined contextually, as the average of the sum of absolute difference between the actual and forecasted GSoR. Basically, MAE reveals how wrong the forecast was. It does not measure underperformance or overperformance of the model. Mathematically, it is expressed as in equation (3.2):

$$MAE = \frac{1}{n} \sum |GSoR_A - GSoR_F| \quad (3.2)$$

where  $n$  is the number of observations.

The Mean Square Error (MSE) is like MAE but differ in behaviour. MSE squares the difference between the actual and forecasted values before summing them. It is expressed in equation 3.3.


$$MSE = \frac{1}{n} \sum (GSoR_A - GSoR_F)^2 \quad (3.3)$$



## CONCLUSION

- Methodology chapter is important to answer WHAT you did and HOW you did the research you conduct.
- Illustrates the proposed method that leads the research novelty.

# EXERCISE

- Find one research paper on **ScienceDirect** that the paper title is related to your research topic  ScienceDirect
- Read the Methodology part.
- Identify the idea of the proposed solution (diagram/algorithm/flowchart/step etc.) of the study.