6

OPTIMAL FEATURE SUBSET SELECTION AFTER DENOISING

The previous chapter focused on the DBA-PSO denoising method, as the pre- processing step for original River ice image samples. GA and DEFS based feature selection method is implemented to select best features from the original feature set for effective classification of River ice types. But, the previously used GA and DEFS based feature selection methods do not focus on selecting the optimal feature subset from the original feature set extracted from feature extraction methods, which in turn affect the classification results for River ice types.

In order to select optimal subsets of features, this research work focused on Computational Intelligence methods in order to overcome the limitations of the existing GA based optimization methods. The optimal feature subset is selected to increase the matching accuracy based on the performance of the classifiers (Kadir et al., 2011; Tzionas et al., 2005). Reducing the dimensions of the feature space not only reduces the computational complexity, but also increases the performance of the classifiers.

Moreover, in feature extraction step, features are extracted using First Order and Second Order Statistical methods. Original feature set consists of 34 features from first order and second order statistical features. From the results, it is observed that most of the features selected are GLCM features and provide higher classification accuracy from the features selected by DEFS feature selection method. So in order to cover the pixel intensities in all directions, the features are extracted at 0^0, 45^0, 90^0 and 135^0 rather than considering only 0^0. This chapter proposes an efficient optimal feature subset selection algorithm which improves in selecting optimal feature subset and classification results of River ice types. The importance of optimal feature subset selection approach for River ice classification is thoroughly investigated in this chapter.

6.1. Thorough Analysis of GA and DEFS based Feature Selection Methods

Genetic Algorithm (GA) based approach is a feature selection method used to select a feature subset that can describe the classification performance by using the machine learning classifier (Sinan et al., 2010). Genetic Algorithms are not always the best choice, because, in certain scenarios, the convergence time taken by the GA is observed to be higher.

- Certain optimization problems cannot be solved by means of Genetic Algorithm. This occurs due to poorly known fitness functions which generate bad chromosome blocks in spite of the fact that only good chromosome blocks cross-over, and
- Moreover, there is no absolute assurance that a genetic algorithm will find global optimum. It happens very often when the populations have a lot of subjects.

Differential Evolution Feature Selection (DEFS) is another feature selection approach used in this work. DEFS has the following unique characteristics which make it suitable for feature selection approach.

- Differential evolution is very efficient because of its simplicity. It has only three input parameters controlling the search process, namely, the size of population N_p, the mutation parameter F and the crossover parameter C_r, and
- It is quite easy to determine the true global minimum, regardless of the initial parameter values, fast convergence.

But, DEFS feature selection method also has certain limitations which are as follows:-

- Far slower than required because slow gradient based learning algorithms are used,
- All parameters such as weight must be tuned iteratively, and
- Several issues like local minima, improper learning rate, etc.

To overcome these problems in this research work, an Extreme Learning Machine (ELM) is used to calculate the weight values of the features in the DEFS. ELM is a simple tuning-free three-step algorithm:-

- The learning speed of ELM is extremely fast,
- Smallest norm of weights,
- ELM tends to reach the solutions straightforward without such trivial issues like local minima, improper learning rate and over-fitting.

The following sections clearly discuss the working of the ELM algorithm in tuning the weight parameter of DEFS for attaining the optimal features.

6.2. Research Contribution in Feature Selection

Most of the existing River ice image classification methods classify the image samples into different ice types, but still it requires an optimal feature subset selection method to solve the curse of dimensionality problems. Irrelevant, redundant or unnecessary, input features may lead to greater computational cost for predictions. So, to overcome this difficulty, feature selection methods are used. This chapter has proposed optimal feature subset selection methods from the initially pre-processed step, which is carried out using DBA-PSO denoising method. Then, the same process carried out in the previous chapters is followed. It involves the segmentation process using Gabor filter and FCM clustering.

The four feature categories were first analyzed for their efficiency in ice type identification through ice class types from River ice images. The experiments revealed that most of the GLCM features are selected from feature selection methods and show higher accuracy when compared with other feature categories. Motivated by this fact and in search of further improvement possibilities towards accuracy, the GLCM features were extracted in 0^0, 45^0, 90^0 and 135^0. Seventeen GLCM features at four angles are extracted, making a total of 68 feature set.

The major contribution of this work is the optimal feature selection process using Extreme Learning Machine based Differential Evolution Feature Selection (ELM-DEFS) and Sequential Extreme Learning Machine (SELM) based Differential Evolution Feature Selection (SELM-DEFS). The proposed work for feature subset selection, from River ice image samples for classification of River ice types, consists of the following steps:

- Image denoising using Decision Based Adaptive Median filter with Particle Swarm Optimization (DBA-PSO),
- Image segmentation using FCM clustering,
- Feature extraction using GLCM method with four angles,
- Optimal Feature Subset Selection using Extreme Learning Machine based Differential Evolution Feature Selection (ELM-DEFS) and Sequential Extreme Learning Machine (SELM) based Differential Evolution Feature Selection (SELM-DEFS), and
- Feature selection based Classification using Probabilistic Neural Network (PNN) classifier.

The major aim of this work is to perform the optimal feature subset selection methods for better classification of River ice image samples. The major problem of DEFS algorithm is the random generation of weight value, and this problem is solved by determining the optimized weight value through ELM. Since the traditional Neural Network learning algorithms have the limitations of slower training and trap into local minimum and over-fitting, this work uses ELM for optimizing the weight values of DEFS approach. ELM has randomly generated input layer and hidden layer connected with weights and hidden layer neuron threshold. In ELM process, the output layer weights obtained by the least square method, in which continuous testing is carried out, to adaptively set the hidden layer node number and randomly assigned for the input weights and hidden layer bias, wherein the whole learning process completed through one mathematical change without iteration. The whole learning process needs only an iteration to complete. Thus, compared with other traditional Neural Network training methods, ELM has the advantages of faster learning speed and better generalization performance, which makes it a good choice for this research work. Moreover, ELM can easily handle the problems like local minima, improper learning rates and over-fitting. The procedure of the Extreme Learning Machine to calculate the weight values of DEFS is specified in the following section.

6.3. Overview of Extreme Learning Machine (ELM)

Neural Networks have been extensively used in many fields due to their ability to approximate complex non-linear mappings directly from the input sample; and to provide models for a large class of natural and artificial phenomena that are difficult to handle using classical parametric techniques. The success of machine learning and artificial intelligence relies on the co-existence of three necessary conditions: powerful computing environments, rich and/or large data, and efficient learning techniques (algorithms). ELM is an emerging learning technique which provides efficient unified solutions to generalized feed-forward networks including but not limited to (both single- and multi-hidden-layer) neural networks, radial basis function (RBF) networks and kernel learning (Chen, S., Cowan, C. F. N. and Grant, P.M. 1991).

ELM theories (Sole and Tsoeu, 2011) show that hidden neurons are important but can be randomly generated and independent of applications, and that ELMs have both universal approximation and classification capabilities; they also build a direct link between multiple theories (specifically, ridge regression, optimization, neural network generalization performance, linear system stability and matrix theory). Consequently, ELMs which can be biologically inspired, offer significant advantages such as fast learning speed, ease of implementation and minimal human intervention. They thus have strong potential as a viable alternative technique for large scale computing and machine learning.

From past few decades, it has been observed that Artificial Neural Networks (ANN) play a major role in image classification and pattern recognition applications. It is because of their generalization and conditioning requirement of minimal training points and faster convergence time. ANNs are found to perform better, and result in faster output in comparison with that of the conventional classifiers. Selection time incurred due to pre-processing speed delay is the limitations and to increase classification accuracy, more training data is utilized in comparison with that of testing data found in conventional classifiers. ANNs are to be addressed with improving the training performance and better classification accuracy noted in neural network architecture. The limitations of conventional classifier are overcome by using ELM, which handles the training for single hidden layer feed forward neural networks.

- **Conventional Extreme Learning Machine**

ELM is a single hidden layer feed forward neural network. The weights for the input layer, hidden layer and biases are randomly assigned without any training process. Moore-Penrose inverse and norm least square solution are used for calculating the output weights which reduces the training time of the ELM network (Huang G. B et al., 2012). ELM is the best match for larger training samples. This ELM network is compared with that of the conventional neural network using the classification rate for River ice images and to classify ice types. Figure 6.1, shows the basic ELM architecture.

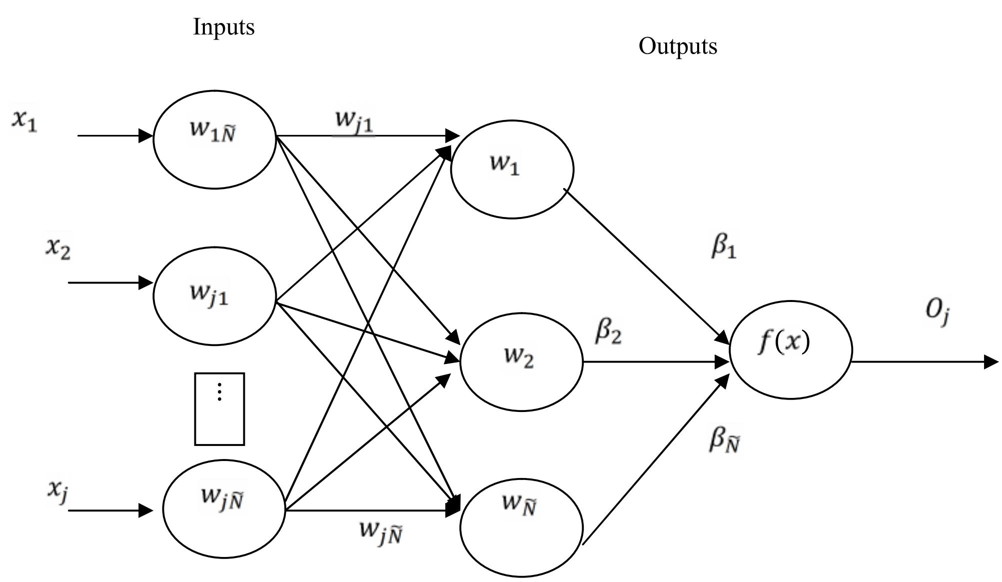

Figure 6.1: ELM Architecture

Conventional ELM classifier algorithm is given as:

Given a training set, $N = \{ (x_i), t_i) | x_i \epsilon R_n, t_i \epsilon R_m, i = 1,..,N \}$, kernel function *f(x)*, and hidden neuron $\tilde{N}$.

Step1: Select activation function and number of hidden neurons Ñ for the given problem.

Step 2: Assign arbitrary input weight w_i and bias b_{i1}, $i = 1, \ldots, H$

Step 3: Calculate output matrix H at the hidden layer

$$\sum_{i=1}^{\tilde{N}} \beta_i f(w_i . x_j + b_i) = t_j, j = 1, \ldots, N \quad (6.1)$$

$$H = f \cdot (w \oplus x + b) \quad (6.2)$$

Step 4: Calculate the output weight β as:

$$\beta' = H^{\dagger} T \quad (6.3)$$

where $H^{\dagger}$ is the Moore-Penrose generalized pseudo-inverse of hidden layer output matrix.

Initial steps like pre-processing, segmentation and feature extraction are carried out in the same way as discussed in Chapter 4 and Chapter 5, and then the proposed optimal feature subset selection method is carried out. Finally, the selected optimal features are given as input to the PNN classifier for classification. The following section clearly discusses about the proposed ELM based DEFS methods.

6.4. Proposed ELM based DEFS (ELM-DEFS) Method for Optimal Feature Subset Selection

Proposed ELM-DEFS method combines the concept of ELM for optimizing the weights in DEFS feature selection method. In DEFS, the population matrix will contain either the trial vector, u_0 or the original vector x_0, depending on one of them will achieve better fitness which is nothing but lower error rate and it is calculated by using ELM network. The input weights and bias weights are used to increase the generalization performance and the conditioning of the ELM network (K.Kothavari, 2014). ELM-DEFS enables in selecting the best feature set with higher accuracy.

Steps involved in the proposed methodology are as follows:

Algorithm 1: ELM-DEFS

Step 1: *The original feature set is taken as input.*

Step 2: *The positions with a set of input weights is initialized with the biases and hidden nodes:* $[W_{11}, W_{12}, \dots, W_{1n}, \dots, W_{21}, W_{22}, \dots, W_{2n}, \dots, b_1, b_2, \dots, b_H]$, *randomly initialized range of [-1, 1] on D dimensions in the search space.*

Step 3: *For all population member vector* $\boldsymbol{x_i}$, *three different randomly selected vectors are merged to create a mutant vector,* $\boldsymbol{v_{i,g}}$.

Step 4: *With the newly created mutant vector* $\boldsymbol{v_{i,g}}$ *, the trial vector* $\boldsymbol{u_{i,g}}$ *is generated by crossover* $\boldsymbol{x_i}$ *with* $\boldsymbol{v_{i,g}}$ *, by using equation (6.4)*

$$u_{ijg} = \begin{cases} v_{j,i,g} \; if \; rand(0,1) \le C_r \\ x_{j,i,g} \; otherwise \end{cases} \qquad (6.4)$$

where $v_{j,i,g}$ *is the* j^{th} *dimension from the* i^{th} *trial vector along the current population g. The crossover probability* C_r □ $[0,1]$, *is a user defined value that controls the fraction of parameter values that are copied from the mutant.*

Step 5: *To find the improved points in the optimal region, the scale factor is dynamically changed by using rand function, as given in equation (6.5).*

$$F = \frac{C_1 \times rand}{max\,(x_{j,r_1,g}, x_{j,r_2,g})} \qquad (6.5)$$

where c_1 *is a constant smaller than 1.*

Step 6: *Additionally, a system constant with specification is implemented to find a better fitness value as in equation (6.6).*

$$x_{j,i,g} = \begin{cases} NF \; if \; x_{j,i,g} > NF \\ 1 \; if \; x_{j,i,g} < 1 \end{cases} \qquad (6.6)$$

where NF is the total number of features.

Step 7: *For each member in the given population, the respective output final weights from ELM network is computed by using the equation (6.3).*

Step 8: *The fitness function is evaluated by Mean Square Error (MSE) as given below:*

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_k^i - d_k^i\right)^2 \qquad (6.7)$$

where the term y_k and d_k are the errors of actual output and target output of the k^{th} output neuron of i^{th} sample.

Step 9: *To find the unique feature set, the distribution factor of feature f_j within the current generation g, is referred as $FD_{j,g}$ which is calculated using below equation:*

$$FD_{j,g} = a_1 \times \left(\frac{PD_j}{PD_j + ND_j}\right) + \frac{NF - DNF}{NF} \times \left(1 - \frac{(PD_j + ND_j}{max\,(PD_j + ND_j)} \right) \qquad (6.8)$$

where PD_j is the number of times that feature f_j has been used in the good subsets. ND_j is the number of times that feature f_j has been used in the less competitive subsets. NF is the total number of features, a_1 is a suitably chosen positive constant that reflects the importance of features in PD, and DNF is the desired number of features to be selected.

Step 10: *The relative difference is obtained according to the equation (6.11).*

Divide the estimated distribution factors for the current and the next iterations by the maximum value, i.e.,

$$FD_g = FD_g / max(FD_g) \qquad (6.9)$$

and

$$FD_{g+1} = FD_{g+1} / max(FD_{g+1}) \qquad (6.10)$$

Compute the relative difference according to the following equation:

$$T = (FD_{g+1} - FD_g) \times FD_{g+1} + FD_g \qquad (6.11)$$

Step 11: *For all the population member in vector x_i, if the trail vector $f(u_i)$ is less than or equal to current population vector $f(x_i)$, then the current population vector x_i is initialled to trail vector u_i*

Step 12: *Repeat step 3 to step 11 until best fitness value is obtained or till the maximum iterations are reached.*

The best feature subset is selected at the end.

Feature subset selected from ELM-DEFS after denoising and the features extracted from GLCM method with 4 angles is clearly shown in Table 6.1.

Table 6.1: Feature Subset Selected from ELM-DEFS Method for GLCM Features Extracted in Four Angles (0^0, 45^0, 90^0 and 135^0)

Method	Best Features
ELM-DEFS	Autocorrelation (0°), Entropy (45°), Difference variance(90°), Information measures of correlation2 (90°), and Inverse difference moment normalized (135°)

The performance evaluation of ELM-DEFS feature subset selection method with GLCM extracted features at 0^0, 45^0, 90^0 and 135^0 is shown in Table 6.2. The results show that the performance of PNN classifier with best ELM-DEFS feature subset achieves good results when compared with DEFS and GA methods.

Table 6.2: Comparison of ELM-DEFS Feature Subset Selection Approach with GA and DEFS Methods for GLCM Features Extracted in Four Angles through PNN Classifier

Performance Evaluation Metrics	GA	DEFS	ELM-DEFS
Features	**5**	**5**	**5**
Accuracy (in %)	69.76	81.44	**98.39**
Sensitivity	0.82	0.67	**0.9600**
Specificity	0.67	0.89	**0.9798**
Precision	0.44	0.76	**0.9231**
F_measure	0.67	0.71	**0.9412**
Gmean	0.62	0.77	**0.9698**
Time	0.0029	0.0017	**2.122063**

The proposed ELM-DEFS based optimal feature subset selection approach for selection of optimal features from River ice image samples is clearly discussed in the above mentioned section. However, in this work, ELM-DEFS algorithm also has certain limitations:

(1) ELM directly calculates the least squares solution. The users cannot fine-tune according to the characteristics of the data sets. It is also poorly controllability,

(2) ELM algorithm is based on the empirical risk minimization, without considering the structural risk and this may lead to over-fitting problems, and

(3) When there are outliers in the data sets, the performance of model will be greatly affected, with poor robustness.

These above mentioned problems in the ELM-DEFS feature selection is solved by proposing a SELM for DEFS method, which will improve in selecting best optimal feature subset and classification accuracy of the River ice samples. The proposed SELM based DEFS method is discussed in the following section.

6.5. Proposed Sequential Extreme Learning Machine (SELM) based DEFS (SELM-DEFS) for optimal feature subset selection

Sequential Extreme learning machine (SELM), (G. B. Huang et al., 2005; Liang NY et al., 2006) stems from batch ELM. It is common that the training observations are sequentially inputted to the learning algorithm; that is, the observations will be tested one-by-one or chunk-by-chunk. It is a versatile sequential learning algorithm in the following sense.

1) The training observations are sequentially (one-by-one or chunk-by-chunk with varying or fixed chunk length) presented to the learning algorithm,

2) At any time, only a single or chunk of observations (instead of the entire past data) are seen and learned,

3) A single or a chunk of training observations is discarded as soon as the learning procedure for that particular (single or chunk of) observation(s) is completed, and

4) The learning algorithm has no prior knowledge as to how many training observations will be presented.

- **SELM encompasses the following principles:-**

1) The generalized inverse of the hidden layer output matrix in batch ELM is proven to be estimated by the left pseudo inverse of the matrix $H^{\dagger} = (H^{T}H)^{-1}H^{T}$, such that the unique solution for the output unit parameters is defined by $\beta^{\square} = (H^{T}H)^{-1}H^{T}T$, which is referred as the least-squares solution to the linear system of equations $H\beta = T$, and

2) The derivation of the one-by-one formula of recursive least-squares (RLS) algorithm results in SELM in the learning mode of one-by-one. This kind of learning mode is suitable for the sequential applications where the training samples arrive one at a time.

- **According to the principle 1,**

Given N distinct input-output samples $\{(X_i, t_i)|X_i \in \mathbb{R}^d, t_i \square \mathbb{R}^q\}_{i=1}^{N}$ and $\epsilon > 0$. For n<N $\mathcal{N}_{G,n}(x)$ realized by a SLFN with n hidden units using infinitely differentiable activation function $G: \mathbb{R}^d \rightarrow \mathbb{R}$ .Let the hidden unit parameters $\{(a_i, b_i)\}_{i=1}^{n}$ be randomly chosen from $\mathbb{R}^d \times \mathbb{R}$ according to some continuous probability distribution. Then the output unit parameters are uniquely determined by $\beta^{\square} = (H^{T}H)^{-1}H^{T}T$ such that $\mathcal{N}_{G,n}(x)$ approximates the input-output correspondences with $\sum_{k=1}^{N} \left|\left|\mathcal{N}_{G,n}(x_k) - t_k\right|\right|^2 < \epsilon$

From Principle it has been shown that the output unit parameters are uniquely defined as the least-squares solution to a linear system of equations. The recursive formula of the solution is referred to as recursive least-squares (RLS). Depending on the update formula of RLS, the proposed SELM can operate in the learning mode of one-by-one. Given a series of N distinct input output samples $\{(x_k, t_k)\}_{k=1}^{N}$ and assume one sample is presented at a time. It may denote the k^{th} input sample. For convenience further define a series of matrices $\{(X_j, H_j, T_j)\}_{j=0}^{m}$ as follows:-

For j=0

$$X_0 = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_{n_0}^T \end{bmatrix}_{n_0 \times d}, H_0 = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{n_0} \end{bmatrix}_{n_0 \times n} T_0 = \begin{bmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_{n_0}^T \end{bmatrix}_{n_0 \times q} \quad (6.12)$$

For j≠0

$$X_j = \begin{bmatrix} X_{j-1} \\ X_{j-2} \\ \vdots \\ X_{n_0+j}^T \end{bmatrix}, H_j = \begin{bmatrix} h_{j-1} \\ h_{j-2} \\ \vdots \\ h_{n_0+j}^T \end{bmatrix} T_j = \begin{bmatrix} t_{j-1} \\ t_{j-2} \\ \vdots \\ t_{n_0+j}^T \end{bmatrix} \quad (6.13)$$

where $n_0 \geq n$ and $N = n_0 + m$ .When $j = m$,$X = X_m$,$H = H_m$ and $T = T_M$. Given that the model of a SLFN with n hidden units with an infinitely differentiable activation function $G: \mathbb{R}^d \rightarrow \mathbb{R}$, it represents a set of input–output mapping functions defined by:

$$\mathcal{N}_{G,n}(x_k) = \sum_{t=1}^{n} \beta_i \,.\, G(x, a_i, b_i) \quad (6.14)$$

Let the hidden unit parameters $\{(a_i, b_i)\}_{i=1}^{n}$ be randomly chosen from $\mathbb{R}^d \times \mathbb{R}$ according to some continuous probability distribution, and then fixed throughout the learning procedure. Only the output unit parameters $\{\beta_i\}_{i=1}^{n}$ are iteratively updated according to the one-by-one formula of RLS algorithm. For j = 0, an initial set of n_0 distinct input-output samples is accumulated. The initial output unit parameters are estimated by solving the linear system of equations:

$$H_0\beta = T_0 \quad (6.15)$$

where $H_0 \in \mathbb{R}^{n_0 \times n}$,$n_0 \geq n$ and $\text{rank}(H_0) = n$

According to principle 1 the solution for the initial output unit parameters are uniquely determined by

$$\beta^{(0)} = K_0^{-1} H_0 T_0 \quad (6.16)$$

$$\text{where } K_0 = H_0^T H_0 \quad (6.17)$$

Suppose the $(n_0 + 1)^{1st}$ input–output sample is presented. Then the output unit parameters are estimated by solving the current linear system of the equations:

$$\beta^{(1)} = K_1^{-1} H_1^T T_1 \quad (6.18)$$

$$= K_1^{-1}(H_0^T T_0 + h_{n_0+1}^T t_{n_0+1}^T)$$

where

$$K_1 = \begin{vmatrix} H_0 \\ h_{n_0+1} \end{bmatrix}^T \begin{bmatrix} H_0 \\ h_{n_0+1} \end{bmatrix} \quad (6.19)$$

$$H_0^T H_0 + h_{n_0+1}^T h_{n_0+1}$$

$$= K_0 + h_{n_0+1}^T h_{n_0+1} \quad (6.20)$$

For $K_0 K_0^{-1} = I_{n_0}$ may be written as

$$\beta^{(1)} = K_1^{-1}(K_0 K_0^{-1} H_0^T T_0 + h_{n_0+1}^T t_{n_0+1}^T) \quad (6.21)$$

$$= K_1^{-1}(K_0 \beta^{(0)} + h_{n_0+1}^T t_{n_0+1}^T)$$

From equation (6.20), $K_0 = K_1 - (h_{n_0+1}^T t_{n_0+1}^T)$ and substitute it into the equation (6.21) and obtain,

$$\beta^{(1)} = K_1^{-1}(K_0 \beta^{(0)} - h_{n_0+1}^T t_{n_0+1}^T \beta^{(0)} + h_{n_0+1}^T t_{n_0+1}^T) \quad (6.22)$$

$$= \beta^{(0)} + K_1^{-1} h_{n_0+1}^T (t_{n_0+1}^T - h_{n_0+1} \beta^{(0)})$$

where

$$K_1 = K_0 + h_{n_0+1}^T h_{n_0+1}$$

Observe that the solution $\beta^{(1)}$ is modified by adding to $\beta^{(0)}$ a correction term $K_1^{-1} h_{n_0+1}^T (t_{n_0+1}^T - h_{n_0+1}\beta^{(0)})$. The above argument can be generalized in a straightforward way. For instance when the $(n_0 + j)^{th}$ input-output sample is presented the solution $\beta^{(j)}$ is updated by:

$$\beta^{(j)} = \beta^{(j-1)} + K_j^{-1} h_{n_0+j}^T \left(t_{n_0+j}^T - h_{n_0+j}\,\beta^{(j-1)}\right) \tag{6.23}$$

$$K_j = K_{j+1} + h_{n_0+j}^T\, h_{n_0+j}$$

Further the recursive algorithm for updating the K_j^{-1} using the Sherman-Morrison formula (Egidi N, 2006)

$$K_j^{-1} = K_{j-1}^{-1} - \frac{K_{j-1}^{-1} h_{n_0+j}^T h_{n_0+j} K_{j-1}^{-1}}{1 + h_{n_0+j} K_{j-1}^{-1} h_{n_0+j}^T} \tag{6.24}$$

Let $P_j = K_j^{-1}$ then the update equations

$$P_j = P_{j-1} - \frac{P_{j-1} h_{n_0+j}^T h_{n_0+j} P_{j-1}}{1 + h_{n_0+j} P_{j-1} h_{n_0+j}^T} \tag{6.25}$$

and

$$\beta^{(j)} = \beta^{(j-1)} + P_j h_{n_0+j}^T (t_{n_0+j}^T - h_{n_0+j}\beta^{(j-1)}) \tag{6.26}$$

These update equations comprise the one-by-one formulation of the RLS algorithm, which is applied for iteratively updating the output unit parameters in SELM. More important, when $j = m$, the final linear system is given by:

$$H_m \beta = T_m \tag{6.27}$$

$$\beta^{} = \beta^{(m)} \tag{6.28}$$

For $H = H_m$ and $T = T_m$

Finally the output unit parameters are estimated by SELM network.

The flowchart illustration of the proposed SELM-DEFS method for optimal feature subset selection is shown in Figure 6.2.

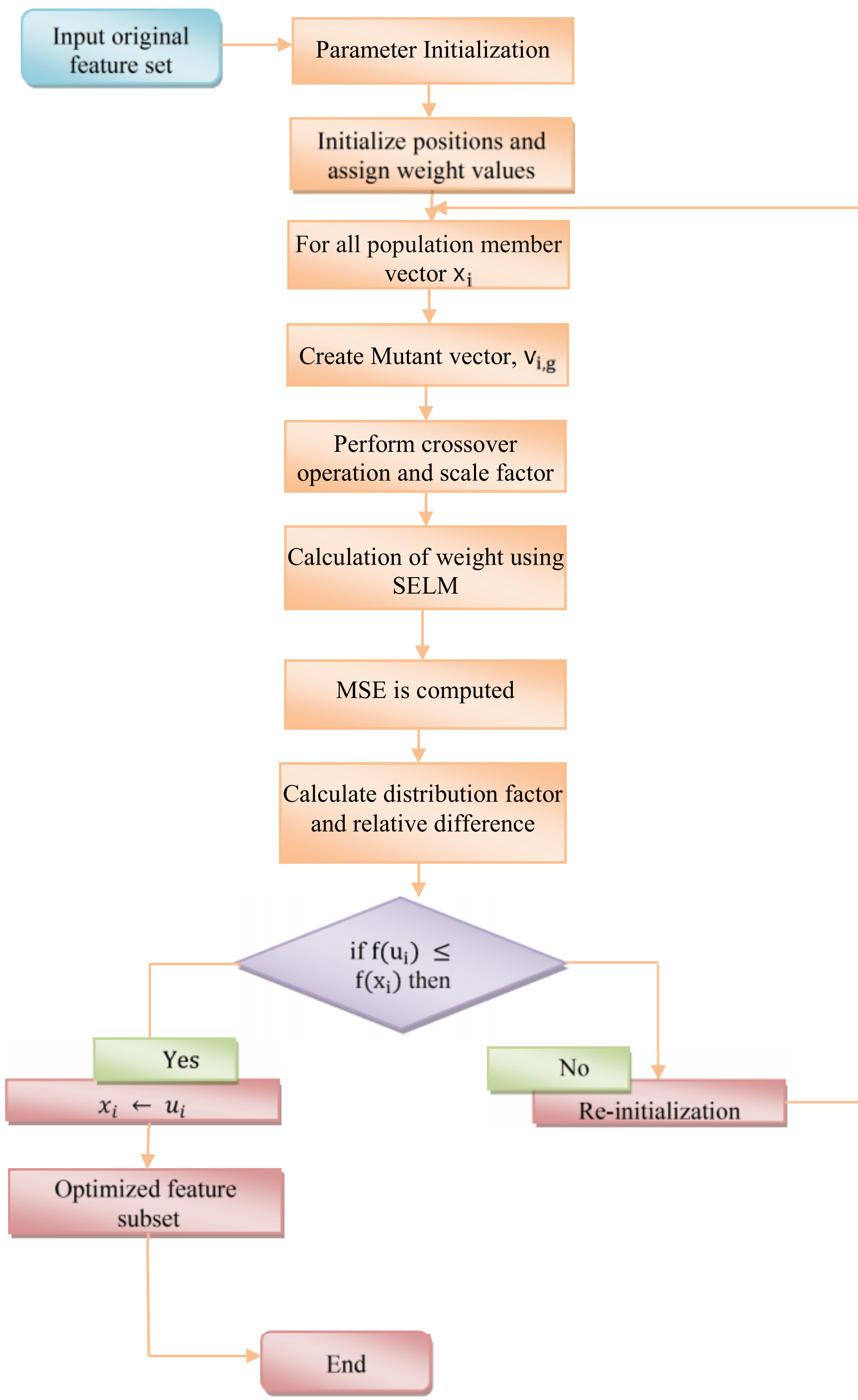

Figure 6.2: Flow Chart representation of SELM-DEFS Feature Subset Selection Method

Algorithm 2: SELM-DEFS

Step 1: *The original feature set is taken as input.*

Step 2: *The positions with a set of input weights is initialized with the biases and hidden nodes:* $[W_{11}, W_{12}, \dots, W_{1n}, \dots, W_{21}, W_{22}, \dots, W_{2n}, \dots, b_1, b_2, \dots, b_H]$, *randomly initialized range of [-1, 1] on D dimensions in the search space.*

Step 3: *Output matrix H at the hidden layer is calculated using equation 6.18 - 6.22*

Step 4: *Output weight β, is obtained through equation 6.23 – 6.26*

Step 5: *For all population member vector* $\boldsymbol{x_i}$, *three different randomly selected vectors are merged to create a mutant vector,* $\boldsymbol{v_{i,g}}$.

Step 6: *With the newly created mutant vector* $\boldsymbol{v_{i,g}}$ *, the trial vector* $\boldsymbol{u_{i,g}}$ *is generated by crossover* $\boldsymbol{x_i}$ *with* $\boldsymbol{v_{i,g}}$ *, by using equation (6.4)*

$$u_{ijg} = \begin{cases} v_{j,i,g} \ if \ rand(0,1) \leq C_r \\ x_{j,i,g} \ otherwise \end{cases} \qquad (6.4)$$

where $v_{j,i,g}$ *is the* j^{th} *dimension from the* i^{th} *trial vector along the current population g. The crossover probability* C_r □ $[0,1]$, *is a user defined value that controls the fraction of parameter values that are copied from the mutant.*

Step 7: *To find the improved points in the optimal region, the scale factor is dynamically changed by using rand function, as given in equation (6.5).*

$$F = \frac{C_1 \times rand}{max\ (x_{j,r_1,g}, x_{j,r_2,g})} \qquad (6.5)$$

where c_1 *is a constant smaller than 1.*

Step 8: *Additionally, a system constant with specification is implemented to find a better fitness value as in equation (6.6).*

$$x_{j,i,g} = \begin{cases} NF \ \ if \ x_{j,i,g} > NF \\ 1 \ if \ x_{j,i,g} < 1 \end{cases} \qquad (6.6)$$

where NF is the total number of features.

__Step 9:__ For each member in the group, the respective output final weights in SELM network is computed by using the equation (6.28),

__Step 10:__ The fitness function is evaluated by Mean Square Error (MSE) as given below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_k^i - d_k^i\right)^2 \qquad (6.7)$$

where the term y_k and d_k are the errors of actual output and target output of the k^{th} output neuron of i^{th} sample.

__Step 11:__ To find the unique feature set, the distribution factor of feature f_j within the current generation g, is referred as $FD_{j,g}$ which is calculated using below equation:

$$FD_{j,g} = a_1 \times \left(\frac{PD_j}{PD_j + ND_j}\right) + \frac{NF - DNF}{NF} \times \left(1 - \frac{(PD_j + ND_j}{max\,(PD_j + ND_j)}\right) \qquad (6.8)$$

where $PD_j$is the number of times that feature $f_j$has been used in the good subsets. ND_j is the number of times that feature f_j has been used in the less competitive subsets. NF is the total number of features, $a_1$is a suitably chosen positive constant that reflects the importance of features in PD, and DNF is the desired number of features to be selected.

__Step 12:__ The relative difference is obtained according to the equation (6.11).

Divide the estimated distribution factors for the current and the next iterations by the maximum value, i.e.,

$$FD_g = FD_g / \,max\left(FD_g\right) \qquad (6.9)$$

and

$$FD_{g+1} = FD_{g+1} / \,max\left(FD_{g+1}\right) \qquad (6.10)$$

Compute the relative difference according to the following equation:

$$T = \left(FD_{g+1} - FD_g\right) \times FD_{g+1} + FD_g \qquad (6.11)$$

Step 13: *For all the population member in vector* x_i *, if the trail vector* $f(u_i)$ *is less than or equal to current population vector* $f(x_i)$*, then the current population vector* x_i *is initialled to trail vector* u_i

Step 14: *Repeat step 5 to step 13 until best fitness value is obtained or till the maximum iterations are reached.*

Thus, based on the SELM-DEFS, optimal features are selected from the River ice image samples. In SELM, the parameters of hidden nodes (the input weights and biases of additive nodes and impact factors nodes) are randomly selected and the output weights are analytically determined based on the sequential data. The algorithm uses the ideas of ELM for sequential learning which has been shown to be extremely fast with generalization. Apart from selecting the number of hidden nodes, no other control parameters have to be manually chosen. SELM with DEFS (SELM-DEFS) are combined in the proposed method, which selects four most relevant features from the feature set with maximum accuracy. The optimal selected features are given as input to the PNN classifier to classify unknown feature vectors into predefined classes. With the selected features, the PNN is then trained to serve as a classifier for discriminating seven different types of River ice classes. The optimal feature subset selected from the proposed SELM-DEFS method for GLCM features after denoising using DBA-PSO filter are shown in Table 6.3. It is clearly observed from the table that, the best features attained from SELM-DEFS method is different from ELM-DEFS method.

TABLE 6.3: Optimal Feature Subset Selected from SELM-DEFS Method for GLCM Features extracted in Four Angles (0^0, 45^0, 90^0 and 135^0)

Techniques	Best features
SELM-DEFS	Correlation (0°), Contrast (0°), Sum Average (45°), Information measures of correlation2 (135°)

The performance evaluation of optimal feature subset selection method SELM-DEFS method with GLCM extracted features at 0^0, 45^0, 90^0 and 135^0 is shown in Table 6.4. The results show that the performance of PNN classifier with optimal

SELM-DEFS feature subset selection method achieves significant results when compared with ELM-DEFS, DEFS and GA methods.

Table 6.4: Comparison of SELM-DEFS Feature Subset Selection Approach with GA, DEFS and ELM-DEFS Methods for GLCM Features Extracted in Four Angles through PNN Classifier

Performance Evaluation Metrics	GA	DEFS	ELM-DEFS	SELM-DEFS
Features	**5**	**5**	**5**	**4**
Accuracy (in %)	69.76	81.44	98.39	**99.19**
Sensitivity	0.82	0.67	0.9600	**0.9831**
Specificity	0.67	0.89	0.9798	**0.9973**
Precision	0.44	0.76	0.9231	**0.9517**
F_measure	0.67	0.71	0.9412	**0.9601**
Gmean	0.62	0.77	0.9698	**0.9792**
Time	0.0029	0.0017	2.122063	**3.162239**

6.6 Computational Complexity Measures

The proposed SELM-DEFS feature subset selection method is observed to have lesser computational complexity when compared with ELM-DEFS method. The computational complexity of the proposed and existing method have been measured based on f-information-measure. The proposed feature selection method has lesser computational complexity with respect to number of features from River ice images and number of samples or objects of the original data set. The computational complexity of selected features is determined through $O(nfsk)$, where 'nfs' represents the number of features selected from algorithm and k is the total number of image samples in the data set. Table 6.5 shows the computational complexity results of the features selection methods wherein NF represents number of features and DNSF represents desired number of selected features. It is clearly observed that the proposed approach shows lesser computational complexity comparatively as NSF is lesser for SELM-DEFS when compared with ELM-DEFS approach.

Table 6.5: Complexity measures for ELM-DEFS and SELM-DEFS methods

Methods	NF	DNSF	Complexity Measures
ELM-DEFS	68	5	244800
SELM-DEFS	**68**	**4**	**195840**

The computational measures of the proposed SELM-DEFS method based on the implementation aspect, plays a vital role in the performance analysis. Halstead's Metric is one of the metrics used to measure the complexity of the design program. This metric focuses on the number of operators and operands used in the program in order to evaluate the computational complexity (Ref. Section 7.1.3 in Chapter 7 for performance measures) which is shown in Table 6.6.

Table 6.6: Computational Measures of Halstead Metrics for Proposed ELM-DEFS and SELM-DEFS Method

Computational Measures	ELM-DEFS	SELM-DEFS
Program vocabulary	60	**90**
Program Length	219	**357**
Estimated time	1.2454	**3.1236**
Volume	896.6615	**1.6064e+003**
Difficulty	25	**35**
Effort	2.2417e+004	**5.6225e+004**
Time	1.2454	**3.1236**

Table 6.6 shows the comparision of complexity measures for the proposed ELM-DEFS and SELM-DEFS methods. Measures like program Vocabulary, Length, Estimated time, Volume, Difficulty and Effort are evaluated to validate the efficient

computational complexity of the proposed SELM-DEFS method. In proposed SELM-DEFS method, the computational measures are higher which is negligible when compared with complexity measures.

6.7 Summary

The key issue in the development of pattern recognition of River ice type classification is the formation of feature extraction analogy, feature subset selection and the classifier. Novel feature subset selection algorithm, based on a combination of Extreme Learning Machine and Differential Evolution Feature Selection optimization techniques, is discussed in this chapter. An efficient DBA-PSO filtering approach is proposed in this work which is used for denoising the corrupted pixels by impulsive noises in gray-level images. In the proposed filter, the weights are optimized by using PSO which minimize the mean square error rate. Optimal feature subset selection is performed using Extreme Learning Machine based Differential Evolution Feature Selection (ELM-DEFS), Sequential Extreme Learning Machine (SELM) based Differential Evolution Feature Selection (SELM-DEFS) is proposed. In the proposed methodology, the input weights and bias weights of the SELM network is used to optimize the fitness function of the DEFS method. In this research work, SELM-DEFS method is proposed, to select optimal feature subset from River ice image samples, to reduce the feature subset, which in turn reduces the dimensionality and computational cost of the classifier. From the results, it is observed that the computational measures are higher for the proposed SELM-DEFS method which is negligible when compared with complexity measures.