Intelligent approach for large-scale data mining

Khaled M. Fouad*

Information Systems Department,
Faculty of Computers and Artificial Intelligence,
Benha University,
Benha, Egypt
Email: kmfi@fci.bu.edu.eg
*Corresponding author

Doaa L. El-Bably

Scientific Computing Department,
Faculty of Computers and Artificial Intelligence,
Benha University,
Benha, Egypt
Email: doaa.elbably@fci.bu.edu.eg

Abstract: Large-scale data mining has become a very difficult issue using traditional methods because the data complexity is very high. In the proposed approach, an integration of three methods; Optimised Principal Component Analysis (OPCA), Optimised Enhanced Extreme Learning Machine (OEELM), and stratified sampling, called OPCA-EELM2SS, is presented to provide intelligent and enhanced large-scale data mining. OPCA provides a good representation of large-scale data sets by using the Stratified Sample (SS). By using OEELM, the optimal number of Hidden Nodes (HNs) in ELM is exploited to build a single hidden layer feedforward neural network (SLFN). The proposed approach is tested by using nineteen benchmark data sets. The experimental results demonstrate the effectiveness of the proposed approach by performing different experiments for classical PCA and Independent Component Analysis (ICA), which are integrated with the enhanced ELM using different evaluation criteria. For more reliability, the proposed approach is compared with many previous methods.

Keywords: principal component analysis; extreme learning machine; particle swarm optimisation; large-scale data mining.

Reference to this paper should be made as follows: Fouad, K.M. and El-Bably, D.L. (2020) 'Intelligent approach for large-scale data mining', *Int. J. Computer Applications in Technology*, Vol. 63, Nos. 1/2, pp.93–113.

Biographical notes: Khaled M. Fouad received his BSc degree in 1995, MSc degree in 2003 and PhD degree in 2012, Department of Systems and Computers Engineering, Faculty of Engineering. Working now as An Associate Professor in Information Systems Department, Faculty of Computers and Informatics, Benha University, Egypt, his current research interests focus on intelligent systems, text mining, data mining, cloud computing, big data analytics, semantic web, and expert systems.

Doaa L. El-Bably is an MSc Student and received her BSc degree in Computers and Informatics in 2012, Faculty of Computers and Informatics, Department of Scientific Computing, Benha University, successfully passed the IBM Academic Certificate exam and earned the title "Big Data Specialist with IBM Big Insights V2.1", Nov-2015. She is working now as a Demonstrator in Computers and Informatics, Benha University, Egypt. Her research interests are in big data, dimensionality reduction, and neural networks.

1 Introduction

Recently, data has grown on a large scale in several fields for both structured and unstructured or semi-structured or multi-structured data that is so large that it is complicated to analyse using traditional methods (Yadav et al., 2013). Therefore, data preparation includes all types of processes performed on data to prepare it for another processing procedure; data pre-processing turns the data into a form that will be more effective and easily processed for the purpose of the user. The main challenge for information science and data mining (Fouad, 2018) to extract essential information from a large

amount of high-dimensional data and to meet the requirement of Big Data`s World (Lohr, 2008). Today's rapidly growing quantity of data leads to a need to analyse and process this data. Representation of the high-dimensional data is inevitable, so dimension reduction is a necessary pre-processing step for many machine-learning techniques. Therefore, dimension reduction is an important method to address the dimensionality problem by removing the redundant or irrelevant information that was performed by many kinds of research (Huo and Smith, 2008; Sarveniazi, 2014).

The Principal Component Analysis (PCA) is the most vastly used multivariate method using statistical methods. It is generally utilised to decrease the dimensionality of high-dimensional data in order to examine its underlying covariance structure of a collection of variables. While singular value decomposition computes Principal Components (PCs) that are the favourite method for numerical data accuracy and to provide a simple means for identification of the PCs for the standard PCA. This paper describes several optimisation models related to PCA by determining the optimal number of the PCs, which are necessary to represent the high-dimensional data.

Until now, Extreme Learning Machine (ELM) has more attention for classification and regression tasks (Feng et al., 2015), because ELM is extremely fast learning model which treats many real-world classification problems with good accuracy rate and has more capability of time processing that is ensured by many researchers (Cao et al., 2016; Li, et al., 2016). It is based on Single-hidden Layer Feedforward Neural Network (SLFN) that generates the hidden layer parameters randomly without tuning or local minima. However, there are many approaches that have been proposed in recent years to progress the performance of the standard ELM in different directions (Huang, 2013; 2014; Cao et al., 2014; Iosifidis et al., 2015; Iosifidis, 2015; Zhang et al., 2015). In our main core of selecting the optimal number of hidden nodes, Huang et al. (2006b) proposed an approach to handle the problem of choosing the proper number of hidden nodes by using several techniques. The Incremental Extreme Learning Machine (I-ELM) which increases the number of hidden nodes until it reaches a certain error (Feng, 2009; Huang et al., 2008), Castaño et al. (2013) used the information retrieved from PCA on training data to estimate the number of hidden nodes, and Memetic-ELM to get the optimal network parameters according to each task (Zhang et al., 2016).

It is difficult to work with massive data using the most relational databases and statistics desktop packs simulation management systems to store, analysis these volumes of data, where it requires new processing models which have the better storage, making decisions, and are capablee of analysing with Big Data technology, so that our main contribution is to create the data value by increasing the processing capacity of the data (Snijders et al., 2012). Therefore, this study provides a new way for processing and analysing Big Data based on PCA-EELM (Elbably and Fouad, 2018) inspired by both the PCA and enhancement ELM, which was developed by implementing new activation functions for the standard EELM, then EELM classifier is improved by using a dimensionality reduction phase based on PCA. PCA-EELM is ensured to work effectively in (binary and multi) classification problems through small and large data sets.

The proposed approaches were applied on the hybrid technique (PCA-EELM) (Elbably and Fouad, 2018) to optimise its main parameters. The first one was named as OPCA-OEELM which considers optimisation model with the main objective function that is maximising the predictive accuracy based on the two important constrains. The first constraint is selecting the optimal number of PCs which obtain effective representation for high-dimensional data through PCA, which transforms the original features to the principal components then uses the proposed approach to select the optimal components which map to the optimal features. The second constraint focuses on the number of hidden nodes for enhanced extreme learning machine.

The second approach was named OPCA-EELM2SS. OPCA-EELM2SS is proposed to apply particle swarm optimisation (PSO) technique on stratified samples collected from the big data to obtain the optimal number of principal components. PSO is a common heuristic algorithm, which has significant interest from many researchers in several research areas and has been effectively applied to different optimisation real-world problems (Foody, 2002) over the years, such as data classification problems.

The performance of the proposed approaches is evaluated using different evaluation criteria and compared with many of the previous works for selecting the optimal number of hidden nodes of ELM and for dimensionality reduction by feature selection techniques. Table 5 (p. 100) presents short description for all comparable previous works with the two proposed approaches, OPCA-OEELM for medical data sets and OPCA-EELM2SS for big data sets.

This paper presents sections as follows: Section 2 displays a short review of all previous works, Section 3 is the background of the used algorithms such as PCA-EELM, PSO and short review on big data processing, Section 4 shows the details of proposed approaches, (OPCA-OEELM) and its effectiveness on medical data sets; (OPCA-EELM2SS) and its role in overcoming big data processing problem. The experimental results of OPCA-OEELM on 14 benchmark data sets and OPCA-EELM2SS on five big data sets are explained in Section 5. Finally, in Section 6, this paper is concluded and some future works are outlined.

2 Related work

2.1 PCA-based approaches

The dimension reduction, which depends on the PCA, aims at reducing the dimensionality of the data by creating a set of derived variables, which are linear combinations of the original variables. Kernel-based PCA (Rizvi et al., 2016), has been developed for achieving efficient lower-dimensional scheduling in variables in Linear Parameter-Varying (LPV) models by reducing the number of scheduling variables to minimise the complexity of computation for LPV controller (design and implementation). Kernel-based PCA is desirable to get LPV models of interest in a logical form. Its purpose is to extract data components efficiently because of its capability

to perform the extraction in a high-dimensional space. The method can solve the problem of optimisation and achieve an affine representation in relation to variables of reduced scheduling.

Sharma and Saroha (2015) integrated the principal component analysis with feature ranking because feature evaluation aims to select the suitable subset of features from the original features, but the authors found that this algorithm is inefficient and impractical for very high dimensionality data sets. To address this issue, the output of PCA, which is a set of reduced for uncorrelated features, is applied to feature ranking and evaluation algorithms to improve the computation time as compared to using feature evaluation and ranking for all the features. The proposed method has been tested and applied to the breast cancer data set.

In the cloud computing environments, the development of dimension reduction methods can support pre-processing of the data and effective storage; therefore, Wu et al. (2015) implemented the hyperspectral PCA method on Spark platform, and the MapReduce model is used, taking full features of the high throughput access and high achievement capabilities of distributed computing in cloud computing environments.

Wang et al. (2016) improved PCA to overcomes the structuring of the observations in the PC space using the linear additive property for normal distribution by proposed the ND-PCA, which can exploit the variance information in the original data. It can obtain analytical results rather than approximate results. Also, it has the ability to handle data of normal distribution form and other additive distributions and PCA of wind speed time series proposed by Heckenbergerova et al. (2014) for concluding future ramp estimation from series of power forecast. In this proposed method, it can be accurate forecasting of wind power and Numerical Weather Prediction (NWP) model is not required for producing wind forecasts.

Raihana et al. (2016) presented sparse PCA based on the inverse power method to carry out the sparsity of PCA because of the principal classical components (PCs) can be complicated to explicated because of the linear combinations of PCA, so that sparse PCA was proposed to address this complication problem to be suitable for reducing the dimensions of complex data. It is desirable to feature extraction for big data since the accuracy rate is larger than input data to any classifier.

2.2 Medical data set classification

Classification of the medical data relies on the performance of the Extreme Learning Machine (ELM) classifier, which was proposed by Huang et al. (2006a) to handle the training for (SLFN) single-hidden layer feedforward neural networks. Matias et al. (2014) and Chyzhyk et al. (2014) ensured that ELM is most appropriate for considerable training samples and also the influence of the number of hidden nodes using variant ratios of the number of features for training and testing data was demonstrated.

Feature selection and classification techniques have shown that exploiting machine learning in high-performance applications aims at supporting scientific research that is based on the medical field. Hassan and Subasi (2016) revealed that the use of a classification algorithm that was based on feature selection and the Leaner Programming Boosting (LPBoost) made the processes of monitoring epilepsy seizures and patient management simple. Furthermore, Hassan et al. (2016) applied a distinguished ensemble learning, named as bootstrap aggregating, to discover epileptic seizures. Kirar and Agrawal (2017) aimed at distinguishing brain signals from electroencephalogram (EEG), by using a machine learning approach. Hassan and Haque (2015) used wireless capsule endoscopy videos to carry out a real-time model that aims at exposing bleeding in the small intestine by extracting a large volume of images that are classified by using Support Vector Machine (SVM) to discover gastrointestinal hemorrhage. On the other hand, the study of many genes is applied as the first step for all gene eliciting data sets. Therefore, Shah and Kussaik (2007) inspected that it is costly to gather genetic data. They proved that not all genes educed are considerable. Therefore, they focused on selecting the most convenient genes from the enormous genes' data set. Therefore, the inessential and redundant genes are removed; then, complexity is reduced. Otherwise, Mohamad et al. (2009) considered a wrapper approach, which, after the feature selection process, delivered features to be provided as input to the next step, which is a classification method. Moreover, Alba et al. (2007) demonstrated that the wrapper method undertakes a proposed algorithm to deliver the accuracy of the classification method. Many of these algorithms achieved good experimental results when they are compared with others in terms of exactly so accuracy. However, there is still required more works to be perfect while comparing the feature selection and classification algorithms with respect to their achievement when applied to a cancer data set. Thus, time computation is a considerable parameter in the comparative study between these methods. Porkodi and Suganya (2015) exploited *K*-NN (*K*-Nearest Neighbours) and neural network classifier to achieve the highest accuracy rate for colon cancer classification. However, they explained that the optimisation techniques could be integrated during the classification process. Many algorithms have been applied for the selection and classification of cancer genes by Al-Rajab and Lu (2014), such as Genetic Algorithm (GA), Particle Swarm Optimisation (PSO).

2.3 Big data reduction

Big data can be represented by an advanced model "6Vs" (Gani et al., 2015), which includes the volume, velocity, value, variety, variability, and veracity while acquiring the data. Many variables in big data sets lead to the imprecate of dimensionality problem, which needs indefinite computational resources. Therefore, the data reduction is the most considerable phase of big data analytics. The data reduction operations are exploited to enhance the quality of big data. The data reduction operations involve a wide range of methods that are utilised for many aims, such as noise reduction, big data gathering from IoTs-based sensory data sources, and data streams of internet-based social media, which provide large amounts of unstructured and unused information.

Salmon et al. (2014) applied noise reduction techniques to eliminate noise and irrelevant data based on noise non-local PCA; extracting features: unstructured data streams in big data

paradigms require a significant effort, and therefore, feature extraction methods are utilised to obtain the useful and structured data from the original big data sets.

Many statistical methods are used to determine domain features from large amounts of data based on the nature and the type of data (Grzegorowski and Stawicki, 2015); such as reducing dimensions, which means that big data sets usually contain thousands of dimensions (i.e., attributes/features in data tables). Therefore, analysing this huge data set can be an issue. Dimension reduction techniques are used to produce highly pertinent data sets for big data analysis (Zhai et al., 2014) and for tackling missing values due to many missing values despite the creation of uniformly structured big data sets. However, it causes a minimisation of the quality of knowledge patterns.

3 Background

3.1 Hybrid approach of principle component analysis and enhanced ELM (PCA-EELM)

A conceptual view of the hybrid approach PCA-EELM (Elbably and Fouad, 2018) is inspired by both the PCA and the enhancement ELM, which was developed by a lot of modification in the activation function of the standard EELM. Then, PCA-EELM improves EELM classifier by using a dimensionality reduction phase based on PCA, which is a powerful statistical technique, to identify the principal components in high-dimensional spaces by reducing the dimensions.

EELM algorithm implements all basic computation functions of the standard ELM, in addition to the following effective functions: softmax function, softsig sigmoid function, and hyperbolic tangent, however, this paper will focus on using EELM with softmax as activation function (Tang, 2013) for all classification. Because the output of the function is interpretable as posterior probabilities, so it used for representing categorical distribution that is useful for multi-class classification and multinomial logistic regression (Bishop and Christopher, 2006).

The PCA-EELM has ensured to work effectively in (binary and multi) classification problems through small and large data sets by applying perfection of the single hidden layer feedforward neural network (SLFN) in classification tasks, therefore the proposed approaches in this study are based on (PCA-EELM).

3.2 Particle swarm optimisation: an overview

PSO is a simple and efficient technique because it depends on swarm intelligence. Therefore, PSO has the lowest time computation rather than genetic algorithms.

Each potential solution is represented as a particle swarm. Each particle has a position vector the search space, which is given by

$$\vec{x_i} = \left(x_{i1}, x_{i2}, ..., x_{iD}\right) \tag{1}$$

where D is the dimensionality of the search space and each particle moves to get the optimal solutions, therefore, it has a velocity vector that represented as

$$\vec{v_i} = \left(v_{i1}, \mathrm{v}_{i2}, ..., \mathrm{v}_{iD}\right) \tag{2}$$

The personal best recorded as (*pbest*), which is the best position of the particle in all previous iterations, and it is updated only when the new position of the particle at the current iteration (k) yields a better function value rather than at the previous iteration (k–1). And the global best named as (*gbest*) is the position that has produced the best function value of all positions, and it is shared with all particles.

During the search, the position and velocity of each particle are updated based on the following equations:

$$x_i^{(k+1)} = x_i^{(k)} + v_i^{(k+1)} \tag{3}$$

$$\begin{aligned} v_i^{(k+1)} = {} & \omega * x_i^{(k)} + c_1 * r_1 * \left(p_i\left(k\right) - x_i^{(k)}\right) \\ & + c_1 * r_1 * \left(g\left(k\right) - x_i^{(k)}\right) \end{aligned} \tag{4}$$

$k + 1$ indicates the next iteration number, k indicates the current iteration number, v_i indicates the velocity vector of particle i, x_i de indicates notes the position vector of particle i, ω is the static inertia weight between 0 and 1, c_1 and c_2 are the constant acceleration coefficients, r_1 and r_2 are random values distributed in [0, 1], $p_i\left(k\right)$ is the personal best of particle i, and g (k) is the global best of the swarm. Based on the local best positions, the global best position, and the updating criteria of all particles can be easily reached to the target.

3.3 Big data processing

The complexity of big data systems was shown in three forms (Jin et al., 2015); data complexity, computational complexity, and system complexity. Data complexity which emerges due to multiple formats of big data that elevates the issue of multi-dimensions and the complexity of inter-dimensional and intra-dimensional relationships. For example, the semantic relationship between different values of the same attribute (i.e., the noise level in the particular areas of the city, raises the inter-dimensional complexity. Likewise, the linked relationship among different attributes (i.e., age, gender, and health records) increases the intra-dimensional complexity. In any big data system, the increasing level of data complexity is directly proportional to the increasing level in computational complexity and while the implementation the algorithms and methods can handle the extremely large data volumes, the level of system complexity are increased because of the computational requirements of big data systems.

Data preparation includes all types of processes performed on data to prepare it for another processing procedure; data pre-processing turns the data into a form that will be more effective and easily processed for the purpose of the user.

4 Proposed approaches

4.1 The proposed OPCA-OEELM

The OPCA-OEELM approach presented in this paper for dimensionality reduction and classification is based on the PCA and enhanced ELM (EELM). The main target of the method is reducing the dimensional spaces of data without losing important information. On the other hand, the goal is improving the accuracy rate of the EELM classifier. The main defect of PCA requires the number of PCs to represent the high-dimensional data so that the proposed approach handles this defect by using PSO to select the optimal number of PCs to transform the data from high-dimensional space to low-dimensional space. EELM also has a drawback that it has free parameters, such as the number of Hidden Nodes (HNs), initial weight, and bias that need to be defined by the user. Since the quality of ELM models depends on a proper setting of these parameters, the main issue for this paper trying to apply EELM is how to set one of these parameter values to ensure perfect generalisation performance of the training data set.

OPCA-OEELM optimises two important parameters PCs for PCA and HNs for EELM using two phases. Firstly, the principal component analysis was applied to overcome the curse of dimensionality with choosing the proper number of PCs by PSO. Secondly, EELM training starts with the optimal number of HNs to maximise the classification performance using the evolution function of PSO.

The overview of the OPCA-OEELM approach is shown in Figure 1, which operates in two main steps. In the first step, particle swarm optimisation is run on the original data sets to obtain the optimal number of PCs and HNs. In the second step, the optimised values for the number of PCs and HNs are used as input to PCA-EELM to reduce the space dimensions by eliminating dimensions that are linear combinations of others, and the EELM classifier is applied with the minimum computation time.

Figure 1 Flowchart of OPCA-OEELM

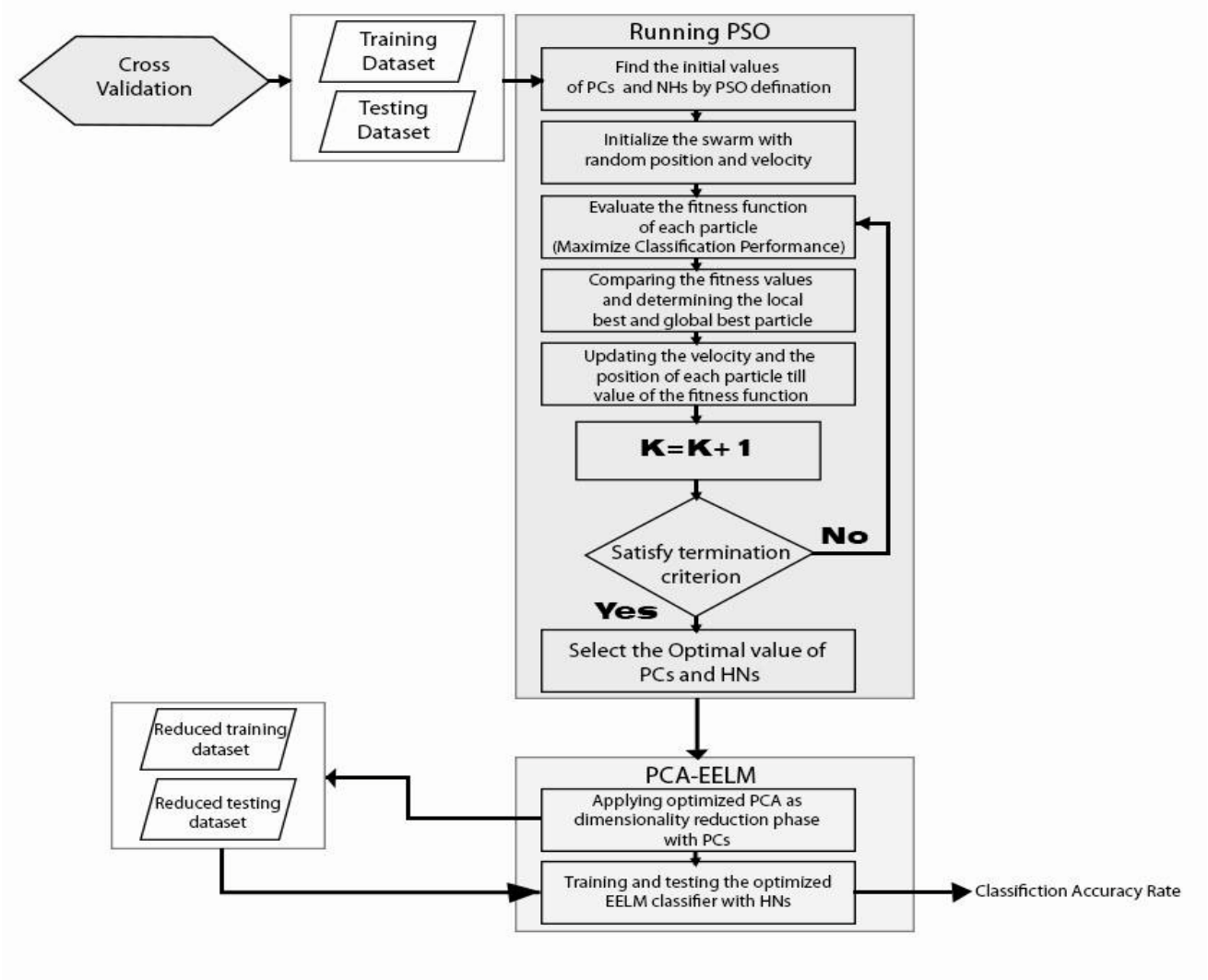


Table 1 Pseudo-code of OPCA-OEELM

Pseudo-code of OPCA-OEELM
Input: Data set and activation function (*softmax*)
Output: The positions of particles and classification accuracy rate

Process:

*step*1: dividing the data set using 10-fold cross-validation to training and testing data sets

*step*2: find the initial value of PCs and HNs from PSO definition

*step*3: initialise the particles with random values to the position (x) and velocity (v)

*step*4: while maximum iterations are not finished do

evaluate the fitness function by maximising the accuracy classification that calculated from the following equation: -

$$\text{Accuracy(ACC)} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

*step*5: determine *pbest* and *gbest* by comparing fitness value to each particle

step6: updating the position (x) and velocity (v) according to fitness value using equations (3) and (4).

*step*7: satisfy termination criterion (maximum iteration=100)

if (satisfied)

go to step8

else

go to step4

*step*8: execute dimensionality reduction using optimised PCA with PCs

*step*9: apply the optimised EELM classifier on the reduced data with HNs

*step*10: Return the classification accuracy rate

The first optimisation version of the hybrid approach of PCA-EELM is summarised in the pseudo-code of OPCA-OEELM in Table 1.

4.2 The proposed OPCA-EELM2SS

The first proposed approach (OPCA-OEELM) provided the perfect classification accuracy as compared with the previous works because the used medical data sets have small numbers of features and records. When OPCA-OEELM is applied to big data sets, it takes large computation time for the optimisation phase. Therefore, OPCA-EELM2SS was proposed to tackle the big data processing tasks within the minimum computation time by applying the optimisation phase on stratified samples (samples that include the same categorical distribution of from the original data set) to get the proper number of PCs then run PCA-EELM within the optimised value as shown in OPCA-EELM2SS algorithm design at Figure 2.

Figure 2 shows that OPCA-EELM2SS goes through three main steps; the first step is stratified sampling process which is to get the optimal sample collected from each category with the same distribution of the original data sets, the second step is running the optimisation technique (PSO) on the stratified sample to obtain the proper number of PCs then pass it as input with the original data sets to the standard PCA-EELM, which is the third step to get the optimal value of the accuracy rate.

The second optimisation version of the hybrid approach of PCA-EELM is summarised in pseudo-code of OPCA-EELM2SS in Table 2.

Table 2 Pseudo-code of OPCA-EELM2SS

Pseudo-code of OPCA-EELM2SS
Input: Original data set and activation function (softmax)
Output: the positions of particles and classification accuracy rate

Process:

*step*1: apply the stratified sampling process to obtain a stratified sample with the same categorical distribution of the original data

*step*2: dividing the stratified sample to S.training and S.testing data sets

*step*3: find the initial value of PCs from PSO definition

*step*4: initialise the particles with random values to the position (x) and velocity (v)

*step*5: while maximum iterations are not finished do

evaluate the fitness function by maximising the accuracy classification that calculated from the following equation:

Table 2 Pseudo-code of OPCA-EELM2SS (continued)

$$\text{Accuracy(ACC)} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

*step*6: determine *pbest* and *gbest* by comparing fitness value to each particle

*step*7: updating the position (x) and velocity (v) according to fitness value by

$$x_i^{(k+1)} = x_i^{(k)} + v_i^{(k+1)}$$

$$v_i^{(k+1)} = \omega * x_i^{(k)} + c_1 * r_1 * \left(p_i(k) - x_i^{(k)}\right) + c_1 * r_1 * \left(g(k) - x_i^{(k)}\right)$$

k denotes the iteration number, ω is the static inertia weight between 0 and 1, c_1 and c_2 are the constant acceleration coefficients, r_1 and r_1 are random values distributed in [0, 1]

*step*8: satisfy termination criterion (maximum iteration = 100)

if (satisfied) go to step9

else go to step5

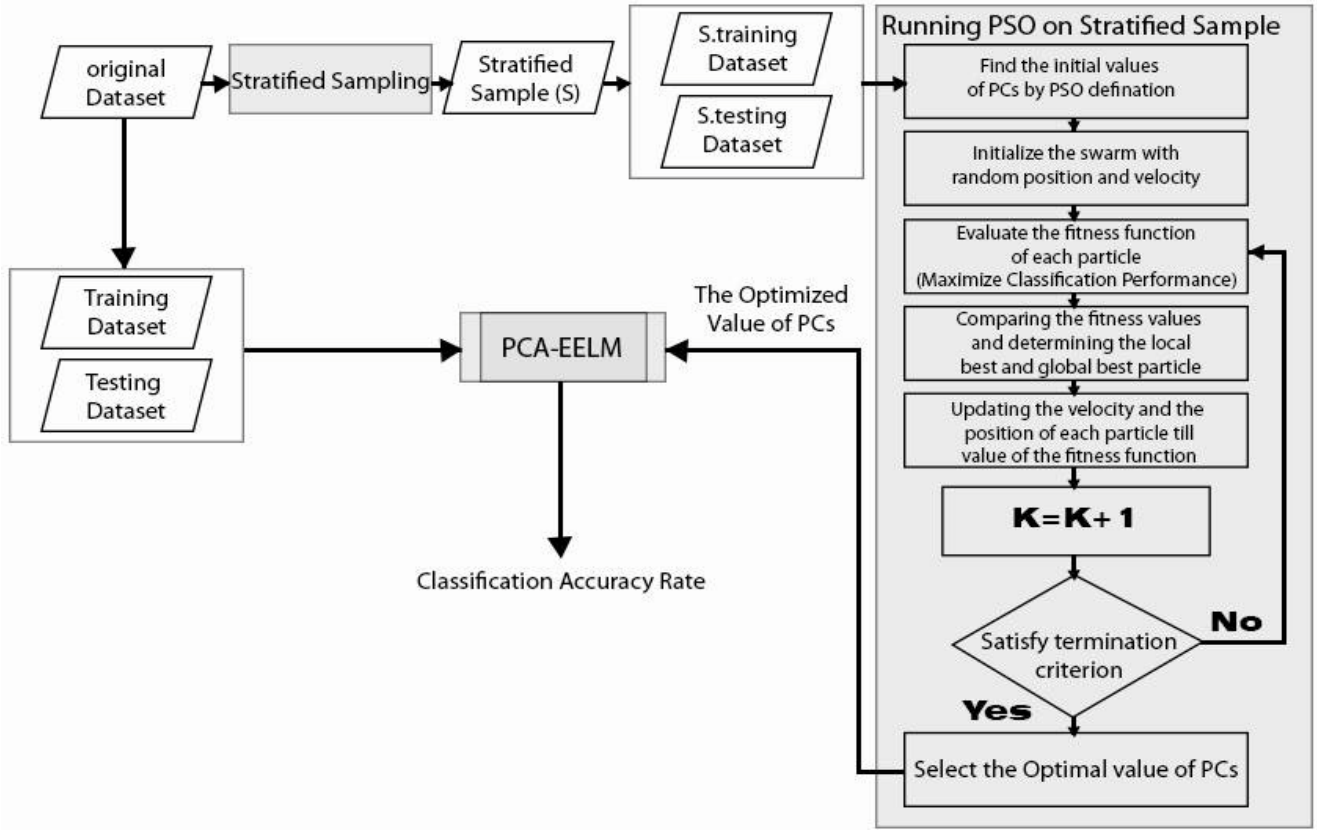
*step*9: pass the optimised number of PCs to the standard PCA-EELM

*step*10: divide the original data set to training and testing data sets

*step*11: Run PCA-EELM within the optimised value of PSc

*step*12: Return the classification accuracy rate

Figure 2 Flowchart of OPCA-EELM2SS



5 Experimental results

The proposed approaches were carried out in R (3.3.2) on computer specifications of an Intel(R) Core (TM) i7-8550 1.99 GHz CPU and 16.00 GB RAM.

5.1 Data sets

The achievement of the proposed approaches is proved by performing the experiments on 19 benchmark classification problems from UCI Machine Learning Repository (UCI Repository, 2017). The training and testing data sets, but are generated by ten-fold cross-validation at small data sets but, for big data sets the training and testing data are partitioned (3n/4) for training data and (n/4) for testing data, where n is the total number of instances. As well, the instances which include missing values have been ignored before execution of the proposed approaches. These data are widely used in many fields, such as medical diagnosis, physical sciences, business, games, and computer sciences.

The detailed description of all small benchmark data sets used in the experiments is listed in Table 3, and the detailed description of big data sets is listed in Table 4. Also, Tables 3 and 4 include the number of instances and the number of attributes. The achievement of the proposed approaches is verified by performing all the experiments on (binary-multi) classification problems.

5.2 Results and discussion

The achievement of the proposed approaches is proved on nineteen benchmark classification problems in many fields (UCI Repository). Ten-fold cross-validation is employed for partitioning the original data to training and testing data in small data sets (3n/4) for training data and (n/4) for testing data in big data sets that were used while learning and testing EELM classifier to improve the predictive accuracy.

The accuracy, sensitivity, and specificity are the common statistical measures used to evaluate the classification performance of the proposed approaches (OPCA-OEELM and OPCA-EELM2SS):

$$\text{Accuracy(ACC)} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5}$$

$$\text{Sensitivity or True Positive Rate } TPR = \frac{TP}{(TP + FN)} \tag{6}$$

$$\text{Sensitivity or True Negative Rate } TNR = \frac{TN}{(TN + FP)} \tag{7}$$

$$\text{F1 score} = 2 * \frac{\text{precision} + \text{recall}}{\text{precision} * \text{recall}} \tag{8}$$

Matthews correlation coefficient

$$(\text{MCC}) = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \tag{9}$$

where TP (true positive) refers to the correctly classified positive case. TN (true negative) refers to the correctly classified negative case. FP (false positive) refers to the incorrectly classified negative case. FN (false negative) refers to incorrectly classified positive cases.

The first proposed approach OPCA-OEELM is compared with many previous algorithms across medical data sets, and the second proposed OPCA-EELM2SS, which based on stratified sampling for big data processing, is applied across five real-world classification problems and is compared with other techniques as shown in Table 5.

Table 3 Description of the small datasets

Data sets	# *Observation*	# *Attributes*	*Source*
Hepatitis	155	19	UCI Repository
Heart	270	13	
Vote	435	16	
German	1000	25	
Yeast	1484	10	
Ecoli	336	7	
Haberman	306	3	
Ionosphere	351	34	
Post-Op	90	20	
Pima Indians Diabetes	768	9	
Wisconsin Breast-Cancer	699	9	
Blood	748	4	
Australian Credit	690	14	
Cleveland Heart Disease	296	13	

Table 4 Description of the big data sets

Data sets	# *Observation*	# *Attributes*	*Size*	*Source*
DIABETIC DATA	101,766	50	18.2MB	UCI Repository
DOTA	102,943	117	23.7MB	
COVERT YPE	581,012	54	71.6MB	
HEPMASS	3,500,000	29	2.4GB	
SUSY	5,000,000	18	2.2GB	

Table 5 Description of all the comparable previous works with the two proposed approaches

The proposed approach	*The previous work*	*Description*
OPCA-OEELM	PCA-ELM (Castaño et al., 2013)	PCA-ELM is using the information that delivered from principal components analysis to fit the hidden nodes of ELM.
	PCA-EELM (Elbably and Fouad, 2018)	PCA-EELM is constituted from PCA as a linear data reduction for dimension reduction with a static determination for the number of principal component analysis by removing irrelevant attributes to speed up the classification method and to reduce the complexity of computation, and EELM is achieved by modifying the activation function of single hidden layer feedforward neural network (SLFN) perfect distribution of categories.

Table 5 Description of all the comparable previous works with the two proposed approaches (continued)

The proposed approach	*The previous work*	*Description*
	PSO + ELM (Subbulakshmi and Deepa, 2015)	The conventional PSO is integrated with the ELM for optimising the input weights in ELM neural network to grow the generalisation performance.
	SRLPSO + ELM (Subbulakshmi and Deepa, 2015)	This self-regulated learning PSO is integrated with ELM for optimising the input weights in ELM neural networks to increase the generalisation performance, which attempts to replace the global position with both the best and worst position by the MSE on the validation set.
	SRM-ELM (Huang and Lai, 2012)	Structured risk minimisation which is proposed to obtain the optimal number of hidden nodes by PSO with the SRM principle that includes the empirical risk and VC confidence to prevent the overfitting problem.
	ELM (Huang and Lai, 2012).	The original extreme learning machine with ten-fold cross-validation.
	$N = \sqrt{a+b+c}$ (Huang and Lai, 2012).	Is the cut and try work to select the number of hidden nodes.
	Optimised LVQ (10 × CV) (Goodman et al., 2002)	There are two different forms of Linear Vector Quantisation (LVQ) "big LVQ" and "optimised LVQ" in which the data represent averages over three runs of 10-way cross-validation use in an LVQ classifier after a reasonable attempt at determining the proper number of output vectors for the classification problem.
	Big LVQ (10 × CV) (Goodman et al., 2002)	
	AIRS (10 × CV) (Goodman et al., 2002)	Artificial Immune Recognition System (AIRS) which is developed as a classifier that depends on the principles of resource-limited artificial immune systems
	Supervised fuzzy clustering (10 × CV) (Abonyi and Szeifert, 2003)	The fuzzy classifier is an extension of the quadratic Bayes classifier that exploits a mixture of models for estimating the conditional class densities, in which each rule can represent more classes with different probabilities.
	Fuzzy-AIS-knn (10 × CV) (Sahan et al., 2007)	Fuzzy-AIS-knn is a new hybrid method of machine learning by integrating a fuzzy-Artificial Immune System (AIS) with the classical k-nearest neighbour algorithm.
	F-score + support vector machine (Akay, 2009)	This method is based on a SVM, which is integrated with feature selection to diagnose breast cancer.
	Association rule + neural network (Karabatak and Ince, 2009)	AR+NN method provides an automatic diagnosis system for detecting breast cancer based on Association Rules (AR) and Neural Networks (NN). In this method, AR is implemented for reducing the dimension of breast cancer data, and NN is used to obtain an intelligent classification.
	Artificial metaplasticity neural network (Marcano-Cedeño et al., 2011)	This approach is considered as an improvement in training the neural network for pattern classification. The algorithm is constituted by the biological metaplasticity property of neurons and Shannon's information theory. While the training phase, the Artificial metaplasticity Multilayer Perceptron (AMMLP) technique gives priority to updating the weights for the less frequent activations over the more frequent ones.
	Mean selection method (Jaganathan and Kuppuchamy, 2013)	Feature selection minimises the computational cost by removing irrelevant features. This study presents the measurement of feature relevance based on fuzzy entropy using three FS strategies which are devised to get the valuable subset of relevant features (Mean, Half selection method and Neural network selection).
	Half selection method (Jaganathan and Kuppuchamy, 2013)	
	Neural network for threshold selection (Jaganathan and Kuppuchamy, 2013)	
	PCA-ANFIS (10 × FC) (Polat and Gunes¸ 2007)	In this study, the diabetes disease is detected using PCA and Adaptive Neuro-Fuzzy Inference System (ANFIS) to improve the diagnostic accuracy of diabetes. PCA-ANFIS has two phases. In the first phase, the dimension reduction is applied using principal component analysis. In the second phase, the diagnosis of diabetes disease is proceeding via an adaptive neuro-fuzzy inference classifier.
	LS-SVM (10 × FC) (Polat et al., 2008)	The aim of this method is the diagnosis of diabetes disease by using Generalised Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM). Also, a new cascade learning system is proposed using Generalised Discriminant Analysis and Least Square Support Vector Machine. This method consists of two stages. The first stage, The GDA to discriminant features as a pre-processing process. In the second stage, LS-SVM is used for the classification of diabetes data set.
	GDA-LS-SVM (10 × FC) (Polat et al., 2008)	

Table 5 Description of all the comparable previous works with the two proposed approaches (continued)

The proposed approach	*The previous work*	*Description*
	MLNN with LM (10 × FC) (Temurtas et al., 2009)	This study includes a multilayer Neural Network (MLNN) structure, which was trained by the Levenberg-Marquardt (LM) algorithm and a Probabilistic Neural Network (PNN) structure. Diabetes diagnosis by a suitable interpretation of the diabetes data is a necessary classification problem to overcome all risks of diabetes.
	PNN (10 × FC) (Temurtas et al., 2009)	
	LDA-MWSVM (Calisir and Dogantekin, 2011)	LDA-MWSVM is introduced based on Linear Discriminant Analysis (LDA) and Morlet Wavelet Support Vector Machine (MWSVM). The structure LDA-MWSVM for the diagnosis of diabetes is included in the feature extraction and feature reduction stage by using the (LDA) method and the classification stage by using the (MWSVM) classifier stage.
	Evolutionary sigmoidal unit neural network (ESUNN) (Hervás-Martínez et al., 2008)	This method depends on a special class of feed-forward neural networks (product-unit neural networks). Product-units are used the multiplicative nodes instead of the additive nodes to obtain the possible strong interactions between variables then the evolutionary algorithm is applied to determine the basic structure of the product-unit model and to estimate the coefficients of the model by the softmax transformation as the decision rule and the cross-entropy error function because of its probabilistic interpretation.
	Evolutionary product unit neural network (EPUNN) (Hervás-Martínez et al., 2008)	
	Multi logistics regression + EPUNN (Hervás-Martínez et al., 2008)	A multi-logistic regression approach depends on the integration of linear and product-unit models, where the product-unit nonlinear functions are constructed with the product of the inputs raised to arbitrary powers. The estimation of the coefficients of the model is carried out in two steps. In the first step, the number of product-unit basis functions and the exponents' vector is selected by means of an evolutionary neural network algorithm. In the second step, a standard maximum likelihood optimisation method determines the rest of the coefficients in the new space given by the initial variables and the product-unit basis functions previously estimated.
	C4.5 (Cheung, 2001)	The standard C4.5 decision tree and naïve Bayes are performed to compare the classification results with Bayesian Network with Naïve Dependence (BNND). However, if there are many dependencies are found, then the prediction performance of the network is significantly reduced by (BNNF).
	Naive Bayes (Cheung, 2001)	
	BNND (Cheung, 2001)	
	BNNF (Cheung, 2001)	
	AIRS (Polat et al., 2005)	A classification method to diagnosis heart disease by the supervised artificial immune system (AIRS), which is based on the principles of resource-limited AIR.
	Hybrid neural network (Kahramanli and Allahverdi, 2008)	The hybrid neural network is presented for classification of data of a medical database using a Fuzzy and crisp value neural network that includes an Artificial Neural Network (ANN) and a Fuzzy Neural Network (FNN).
	Neural networks ensemble (Das et al., 2009)	A neural network ensemble method is developed by integrating the posterior probabilities or the predicted values from multiple predecessor models.
OPCA-EELM2SS	PCA-EELM (Elbably and Fouad, 2018)	PCA-EELM is described in the second row in the same Table 5.
	ICA-EELM	The standard independent component analysis is integrated as a dimensionality reduction technique with the enhanced ELM to compares the effectiveness of the PCA for big data sets.

The reason for the optimisation of PCA-EELM is the complementary strength of integration between PCA and EELM that achieves the best performance by assessing it against many algorithms (Elbably and Fouad, 2018; Castaño et al., 2013) in a lot of previous work. Therefore, OPCA-OEELM was implemented to optimise the number of principal components, which are necessary to PCA as dimension reduction phase; in addition, the classification performance is improved by the optimal number of hidden nodes as shown in Table 6, which ensures the effectiveness of the proposed approach OPCA-OEELM by comparing it with two previous versions of ELM.

When the proposed approach is compared with the conventional classifiers from previous work (PCA-ELM and PCA-EELM), it is found that OPCA-OEELM achieved better results in terms of the number of hidden nodes and accuracy of classification across all data sets.

Figure 3 illustrates clearly the value of choosing the number of hidden nodes by the proposed approach. The optimal number of principal components (features), which is used for improving classification performance, is justified in Figure 4. These figures obviously discover the benefit of selecting the hybrid technique (PCA-EELM).

Table 6 Comparison results of the proposed approach with two main versions of ELM as a previous works

Data set	*Previous work (PCA-ELM)*		*Previous work (PCA-EELM)*		*Proposed Approach (OPCA-OEELM)*	
	NHN	*Accuracy*	*NHN*	*Accuracy*	*NHN*	*Accuracy*
Hepatitis	12	79.487	10	89.30	8	91.61
Heart	8	77.941	10	83.08	4	93.78
Vote	11	92.885	10	94.36	9	98.29
German	28	76.000	10	69.96	16	93.59
Yeast	7	51.482	10	89.53	9	90.16
E.coli	5	87.882	10	98.90	6	98.99
Haberman	3	76.315	10	97.23	7	96.00
Ionosphere	17	86.363	10	98.14	12	99.11
Post-Op	9	81.818	10	95.94	6	96.51
Pima Indians Diabetes	6	72.875	10	99.18	9	99.72
Mean	10.6	78.304	10	91.562	8.6	96.02

Figure 3 Comparison results of the number of hidden nodes for PCA-ELM, PCA-EELM, and OPCA-OEELM

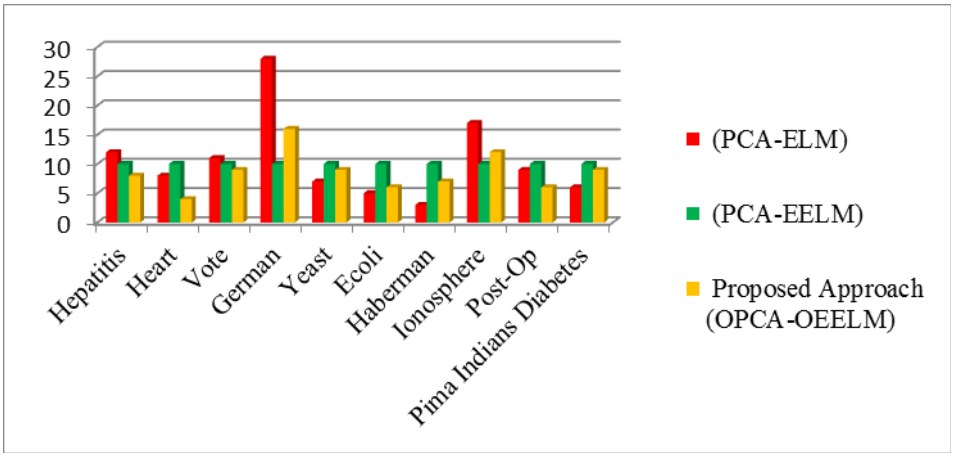
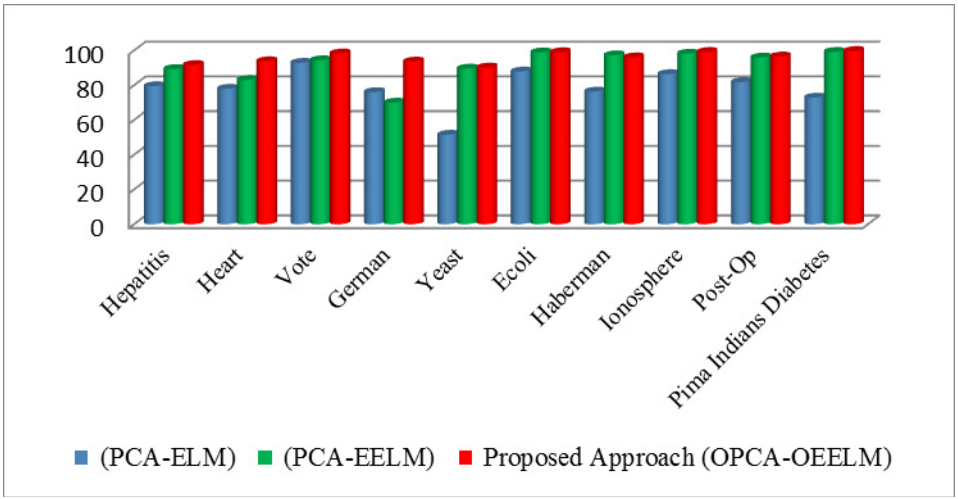


Figure 4 Classification accuracy parameters with PCA-ELM, PCA-EELM, and OPCA-OEELM



While running the proposed OPCA-OEELM, many graphs were considered to show the capability of OPCA-OEELM to search for the optimal solution in two dimensions; one is for PCA; the other one is EELM as illustrated in Figures 5 and 6.

5.2.1 Effectiveness of OPCA-OEELM for medical data set classification

Artificial Neural Networks (ANNs) are widely utilised to treat the real-world classification problems in medical applications, which are prime data mining problems. ANN algorithms have good performance and poor computation time. Therefore, disease diagnosis, which is managed by machine learning methods, is based on ANNs.

The proposed OPCA-OEELM was tested on five medical data sets of the UCI Repository for handling data classification, By applying the proposed OPCA-OEELM algorithm for the considered data sets, experimental results indicate that the proposed model is able to achieve better classification accuracy rate compared to the previous algorithms. Hence, the classification performance of the proposed approach using the common statistical measures (sensitivity and specificity) which were developed to get the performance of classification techniques.

Table 7 shows the results for several feature selection algorithms on the breast cancer data set. It is observed that the highest classification accuracy rate is achieved by the proposed approach (OPCA-OEELM), therefore, Figure 7 displays the classification accuracy parameters for PSO-ELM and SRLPSO-ELM classifier and the proposed approach OPCA-OEELM in terms of accuracy, sensitivity, and specificity.

Figure 5 The proposed approach at iteration 1

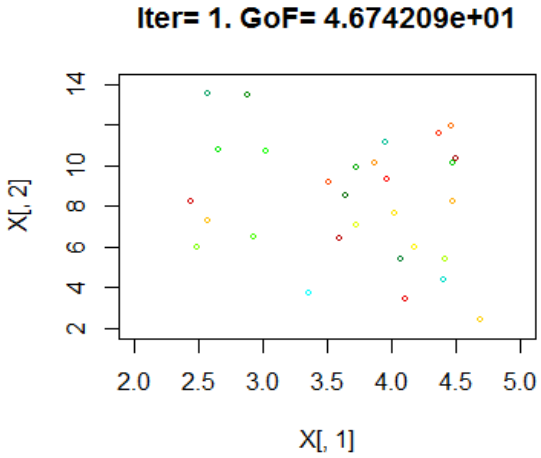


Figure 6 The proposed approach at iteration 100

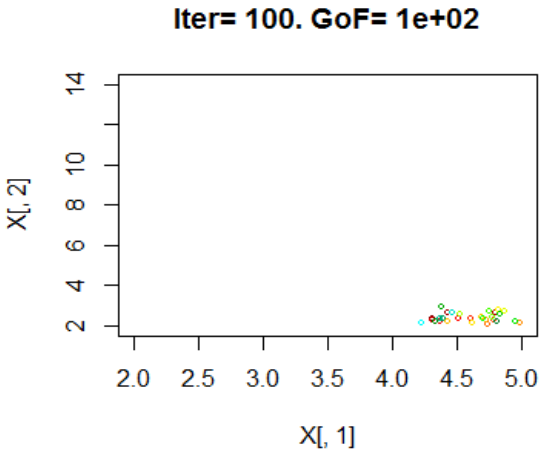
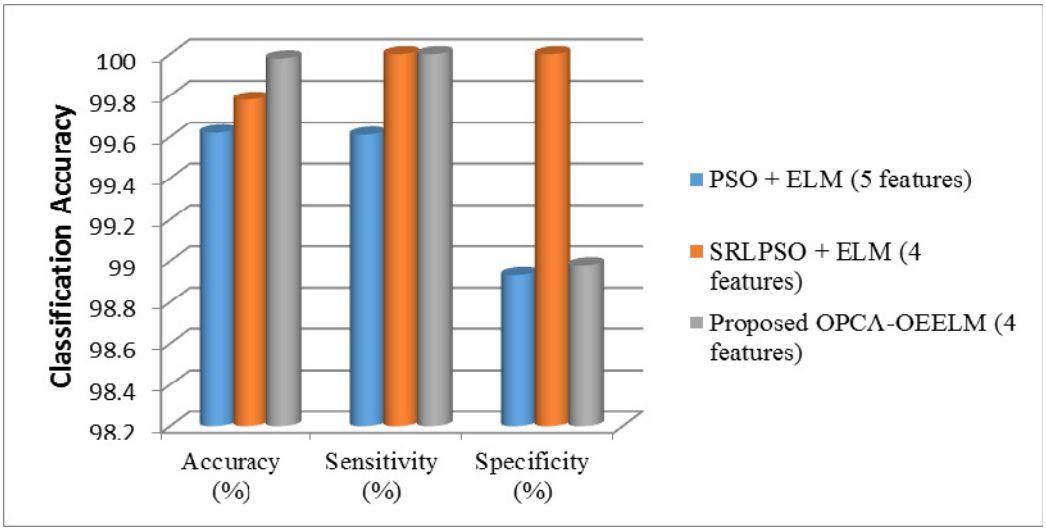


Table 7 Classification results with the breast cancer data set

Methodology adopted	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Number of selected features*
Optimised LVQ (10 × CV) (Goodman et al., 2002)	96.70	91.29	92.34	3
Big LVQ (10 × CV) (Goodman et al., 2002)	96.80	95.23	96.10	3
AIRS (10 × CV) (Goodman et al., 2002)	97.20	96.92	95.00	4
Supervised fuzzy clustering (10 × CV) (Abonyi and Szeifert, 2003)	95.57	98.23	97.36	5
Fuzzy-AIS-knn (10 × CV) (Sahan et al., 2007)	99.14	99.56	100	5
F-score + support vector machine (Akay, 2009)	99.51	99.24	98.61	4
Association rule + neural network (Karabatak and Ince, 2009)	97.4	93.12	91.26	5
Artificial metaplasticity neural network (Marcano-Cedeño et al., 2011)	99.26	100	97.89	5
Mean selection method (Jaganathan and Kuppuchamy, 2013)	95.99	93	97	4
Half selection method (Jaganathan and Kuppuchamy, 2013)	96.71	94	98	5
Neural network for threshold selection (Jaganathan and Kuppuchamy, 2013)	97.28	94	99	7
PSO + ELM (Subbulakshmi and Deepa, 2015)	99.62	99.61	98.93	5
SRLPSO + ELM (Subbulakshmi and Deepa, 2015)	99.78	100	100	4
Proposed OPCA-OEELM	99.98	100	98.98	4

Figure 7 Classification results with breast cancer data set



In Pima Indians Diabetes Data set, while comparing the proposed approach with other classification algorithms as shown in Table 8, it is observed that OPCA-OEELM achieved a good increase in classification accuracy because it was increased from 93.09% to 96.11 with only three principal components and these results are graphically represented in Figure 8.

In Heart-Statlog Data set, Table 9 includes the comparison results between the proposed approach (OPCA-OEELM) and many different feature selection algorithms, which were applied to the Heart-Statlog data set. From these results, it is observed that OPCA-OEELM had the best classification accuracy parameters in terms of accuracy rate, sensitivity, and specificity, as shown in Figure 9.

The results of applying the common techniques and the proposed approach on the hepatitis data set are presented in Table 10. These results obtained by using only four principal components achieved a little increase in all classification accuracy parameters. However, this small increase makes any medical application more accurate as graphically represented in Figure 10.

Table 8 Classification results with Pima Indians diabetes data set

Methodology adopted	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Number of selected features*
PCA-ANFIS (10 × FC) (Polat and Gunes¸ 2007)	89.47	70	71.1	5
LS-SVM (10 × FC) (Polat et al., 2008)	78.21	73.91	80	4
GDA-LS-SVM (10 × FC) (Polat et al., 2008)	79.16	79.1	83.33	5
MLNN with LM (10 × FC) (Temurtas et al., 2009)	79.62	70	70.31	4
PNN (10 × FC) (Temurtas et al., 2009)	78.05	71	70.5	3
LDA-MWSVM (Calisir and Dogantekin, 2011)	89.74	83.33	93.75	5
Mean selection method (Jaganathan and Kuppuchamy, 2013)	76.04	71	78	3
Half selection method (Jaganathan and Kuppuchamy, 2013)	75.91	69	79	4
Neural network for threshold selection (Jaganathan and Kuppuchamy, 2013)	76.04	71	78	3
PSO + ELM (Subbulakshmi and Deepa, 2015)	91.27	85.26	94.10	4
SRLPSO + ELM (Subbulakshmi and Deepa, 2015)	93.09	91.47	96.29	3
Proposed OPCA-OEELM	96.11	95.65	96.11	3

Figure 8 Classification accuracy parameters for Pima Indians diabetes data set

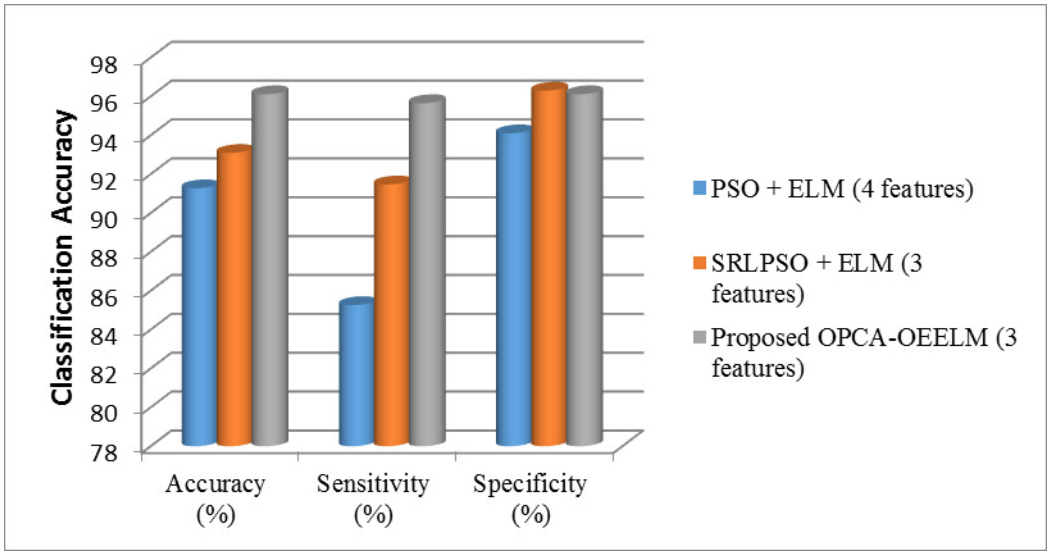


Table 9 Classification results with the Heart-Statlog data set

Methodology adopted	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Number of selected features*
Evolutionary sigmoidal unit neural network (ESUNN) (Martínez-Estudillo et al., 2008)	83.22	84.32	81.65	5
Evolutionary product unit neural network (EPUNN) (Martínez-Estudillo et al., 2008)	81.89	83.67	84.91	4
Multi-logistic regression + EPUNN (Hervás-Martínez, et al., 2008)	83.12	78.15	80.59	5
Mean selection method (Jaganathan and Kuppuchamy, 2013)	84.44	85	84	6
Half selection method (Jaganathan and Kuppuchamy, 2013)	84.81	85	84	7
Neural network for threshold selection (Jaganathan and Kuppuchamy, 2013)	85.19	85	86	4
PSO + ELM (Subbulakshmi and Deepa, 2015)	85.88	86.00	86.03	4
SRLPSO + ELM (Subbulakshmi and Deepa, 2015)	89.96	87.79	88.42	3
Proposed OPCA-OEELM	93.78	92.98	91.00	3

Figure 9 Classification accuracy parameters for Heart-Statlog data set

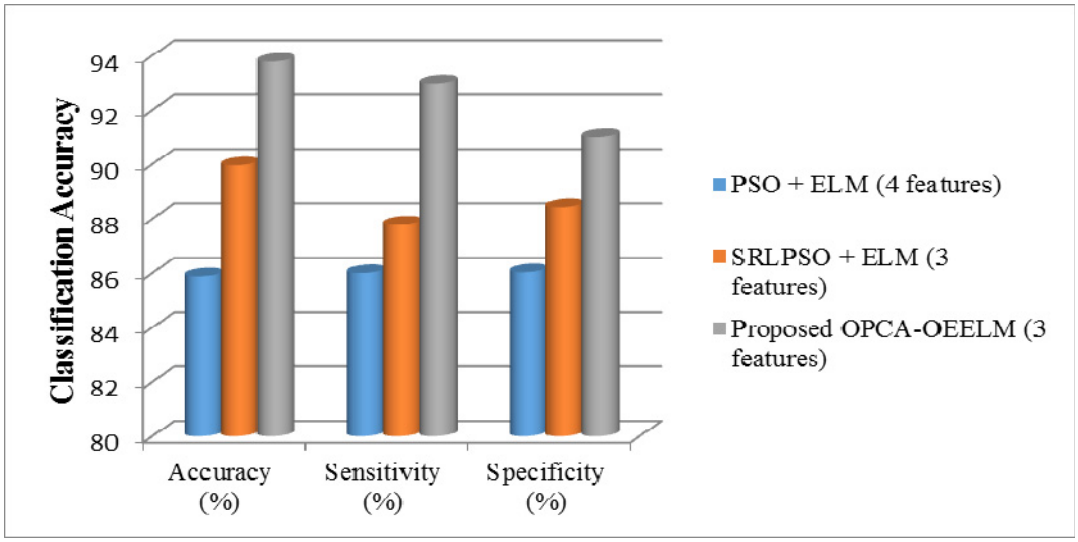
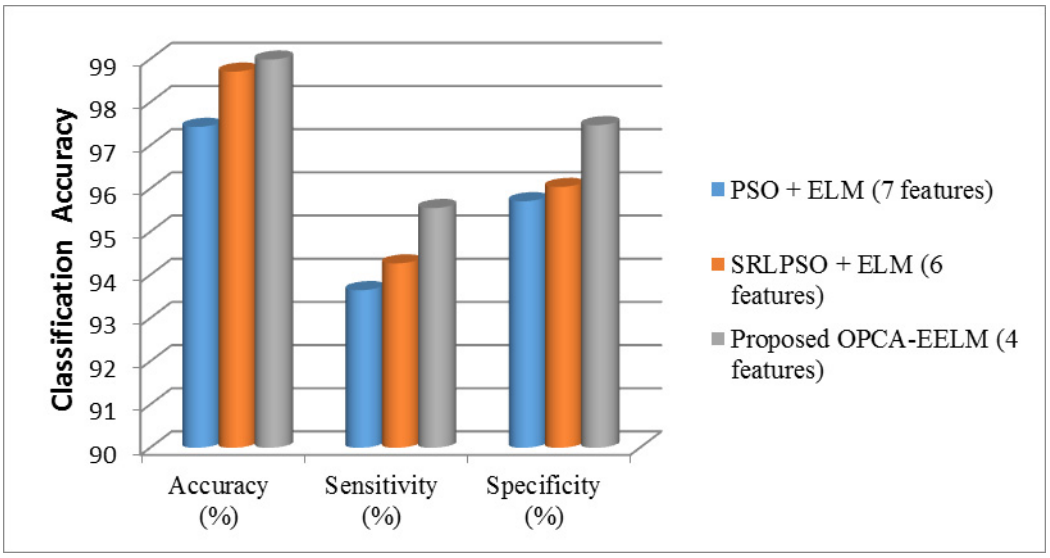


Table 10 Classification results with the Hepatitis data set

Methodology adopted	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Number of selected features*
Conventional artificial neural network (Reibnegger et al., 1991)	97.00	92.31	94.5	19 (All)
Mean selection method (Jaganathan and Kuppuchamy, 2013)	82.58	87	60	8
Half selection method (Jaganathan and Kuppuchamy, 2013)	85.16	90	66	10
Neural network for threshold selection (Jaganathan and Kuppuchamy, 2013)	85.16	90	66	10
PSO + ELM (Subbulakshmi and Deepa, 2015)	97.43	93.65	95.71	7
SRLPSO + ELM (Subbulakshmi and Deepa, 2015)	98.71	94.27	96.04	6
Proposed OPCA-OEELM	98.99	95.55	97.47	4

Figure 10 Classification accuracy parameters for Hepatitis data set



In Cleveland Heart Disease Data set, all used previous medical applications are considered binary-class classification problems. However, this problem is the only multi-class classification problem with five classes; therefore, Table 11 and Figure 11 illustrate the high, increasing percentage in classification accuracy that was improved from 91.33% to 96.54% using the proposed approach (OPCA-OEELM)

Table 11 Classification results with the Cleveland Heart Disease data set

Methodology adopted	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Number of selected features*
C4.5 (Cheung, 2001)	81.11	77.23	76.58	13 (All)
Naive Bayes (Cheung, 2001)	81.48	80.97	81.22	4
BNND (Cheung, 2001)	81.11	82.13	80.42	3
BNNF (Cheung, 2001)	80.96	76.93	75.81	5
AIRS (Polat et al., 2005)	84.50	75.34	72.96	13 (All)
Hybrid neural network (Kahramanli and Allahverdi, 2008)	87.40	93.00	78.50	6
Neural networks ensemble (Das et al., 2009)	89.01	80.95	95.91	13 (All)
Mean selection method (Jaganathan and Kuppuchamy, 2013)	81.75	82	82	6
Half selection method (Jaganathan and Kuppuchamy, 2013)	83.44	84	83	7
Neural network for threshold selection (Jaganathan and Kuppuchamy, 2013)	84.46	82	82	3
PSO + ELM (Subbulakshmi and Deepa, 2015)	89.47	94.49	96.02	5
SRLPSO + ELM (Subbulakshmi and Deepa, 2015)	91.33	95.46	97.29	3
Proposed OPCA-OEELM	96.54	95.68	97.67	6

Figure 11 Classification accuracy parameters for Cleveland Heart Disease data set

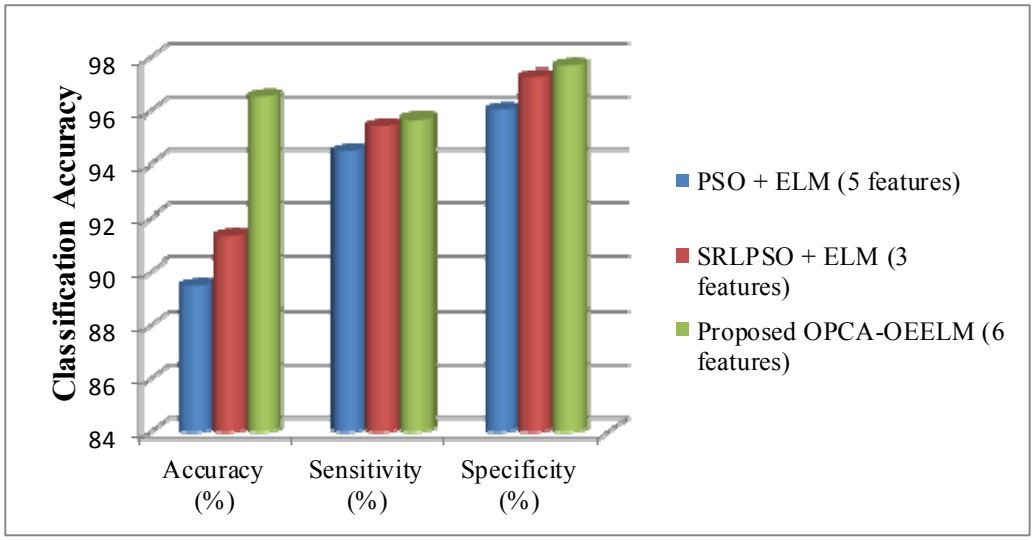


Table 12 Comparison results of the proposed approach with other previous work in the domain of improving the number of hidden nodes

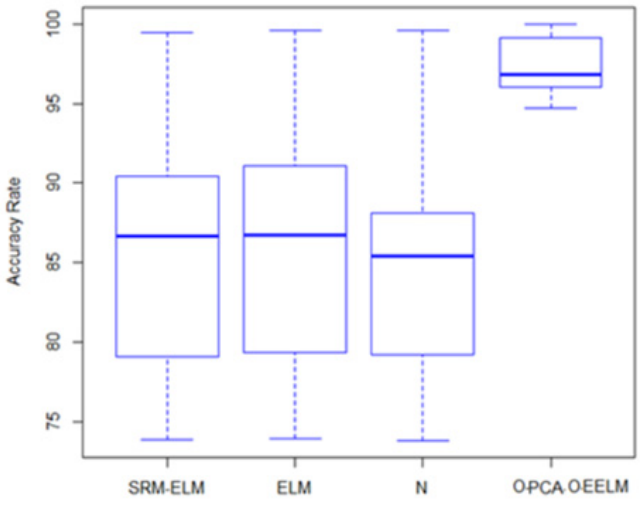
	SRM-ELM		*ELM*		$N=\sqrt{a+b+c}$		*OPCA-OEELM*	
	NHN	*Accuracy (%)*	*Range of NHN*	*Accuracy (%)*	*Range of NHN*	*Accuracy (%)*	*NHN*	*Accuracy (%)*
Haberman	7±1	73.86	4~9	73.95	3~12	73.81	7±4	96.00
Blood	17±3	86.73	15~45	86.92	3~12	86.13	8±2	97.52
Pima	17±2	79.09	10~17	79.36	4~13	79.18	9±5	96.11
Ionosphere	25±5	90.40	25~55	91.09	7~16	85.64	12±4	99.11
Breast Cancer	15±6	99.45	10~50	99.60	4~13	99.60	8±2	99.98
Australian Credit	17±3	86.53	13~23	86.53	5~14	85.18	13±6	94.73

From the result of applying the proposed approach on five medical data sets and comparing it with other previous algorithms, the proposed approach (OPCA-OEELM) has proven its effectiveness in binary and multi-class classification tasks with just 100 iterations, otherwise, most previous works use a maximum iteration of 500 as the termination criterion.

Especially, in the domain of optimising the number of hidden nodes, the performance of OPCA-OEELM was evaluated using the classification accuracy rate and the number of hidden nodes, and these results are compared with other algorithms in Table 12 (Huang and Lai, 2012). The first is SRM-ELM, which used the value of position p as hidden node number to be the key link from PSO to ELM then at every iteration position p and v will be updated; the second is the standard ELM with ten-fold cross-validation; and the third as executed by the cut and try work $N=\sqrt{a+b+c}$ where N is the number of hidden nodes for ELM, a and b are the input and output nodes and c is a random number in [1:10].

Moreover, for the purpose of explanation, Figure 12 shows the box plot, which represents the percentage accuracy over three different algorithms over all data sets using SRM-ELM, ELM, the cut-try work to calculate N (number of hidden nodes) for ELM and. It is obvious from Figure 12 that the OPCA-OEELM is located at the upper side of the figure, which indicates that the proposed approach had the best accuracy scores of all the other algorithms.

Figure 12 Boxplot for all optimisation techniques for the number of hidden nodes



5.2.2 The proposed OPCA-EELM2SS for big data processing

Data sets are gradually growing larger in size. As a result, many techniques have obstacles for analysing data sets to obtain useful knowledge (García-Pedrajas and de Haro-García, 2014). The high computation times and storage requirements of the existing classification algorithms make them inapplicable for these huge data sets (Hernandez-Leal et al., 2013). However, the size of the data set by selecting a representative subset has two main advantages: it reduces the optimisation time required to get the proper number of the PCs, and it accelerates the classification algorithms within reduced data (Dornaika and Aldine, 2015).This subset is generated using stratified random sampling, which involves dividing the entire population into homogeneous groups, which are called strata (singular is stratum), then random samples are then selected from each stratum. A stratified sample is carefully selected to be the input of the optimisation technique (PSO) without minimising the predictive power of the classifier trained with such a subset (Nanni and Lumini, 2011).

In the domain of improving the number of feature selection, feature selection is an important phase for real-world classification problems. PCA has been used for the dimension reduction phase to remove all irrelevant or redundant features. However, most classification techniques are expensive computations. Therefore, PCA is combined with an Enhanced Extreme Learning Machine (EELM). However, PCA depends on the main parameter, which is the number of principal components, to transform the original data from the high-dimensional space to the low-dimensional space. The proposed approach in dimension reduction phase, experiments conducted over five big data sets, and the results showed that the proposed approach (OPCA-EELM2SS) is capable of producing the optimal number of the general feature subsets (principal components). Hence, PCA is able to remove irrelevant, noisy, or redundant features while preserving the classification accuracy rate. These results were compared with the classical PCA-EELM classifier and the classical ICA + EELM classifier, as shown in Table 13.

As an outline of the experimentation with large size data sets, the proposed approach achieved competitive results in terms of accuracy, sensitivity, and specificity, F1, MCC, AUC. Finally, considering execution time, the algorithms presented in this paper were able to classify the data sets much faster than the proposed approach, but within the worst classification accuracy. OPCA-EELM2SS achieved the maximum classification accuracy rate with the larger computation time due to the optimisation phase (i.e., For SUSY data set requires 849 s for optimisation phase, although the proposed approach requires 32.19 s to run classification within the optimised value of PCs).

Table 13 The results of optimisation of the reduction phase

	Model	*Data set*				
		DIABETIC	*DOTA*	*COVERTYPE*	*HEPMASS*	*SUSY*
PCA+EELM	Accuracy (%)	84.55	44.10	85.65	42.58	52.31
	Sensitivity (%)	75	54.88	86.57	44.51	51.67
	Specificity (%)	94.515	66.69	85.71	73.56	52.65
	F1	69.825	43.67	84.39	77.43	51.40
	MCC	0.6790	0.3480	0.8571	0.3504	0.4843
	AUC	0.8540	0.6305	0.8647	0.3566	0.526
	Time(s)	1.95	22.80	48.30	64.72	224.21
ICA+EELM	Accuracy (%)	35.26	41.98	33.48	45.45	94.81
	Sensitivity (%)	34.17	43.07	34.22	46.14	95.56
	Specificity (%)	64.93	67.67	73.43	56.96	95.67
	F1	31.74	35.52	13.10	34.10	95.84
	MCC	0.3057	0.3408	0.1636	0.3408	0.9212
	AUC	0.3015	0.3166	0.2410	0.2702	0.9567
	Time(s)	2.68	37.16	66.70	111.42	158.13
OPCA+EELM2SS	Accuracy (%)	90.45	86.95	91.47	94.58	96.15
	Sensitivity (%)	91.56	87.55	90.99	95.21	96.00
	Specificity (%)	92.12	88.14	92.58	95.47	96.65
	F1	91.57	85.88	91.66	93.33	95.78
	MCC	0.9299	86.64	0.9101	94.01	0.9457
	AUC	0.9001	0.8523	0.9233	0.9539	0.9677
	Time(s)	129+0.90	355 + 4.46	490+13.80	620 + 21.9	849 + 32.19

The performance of the proposed OPCA-EELM2SS approach is proved by applying it to five genuine big data set classification problems (UCI Machine Learning Repository). The specification of these problems is listed in Table 10. Figure 14 demonstrates the mean performance parameters of PCA-EELM, ICA-EELM, and OPCA-EELM2SS techniques for all big data sets based on the enhanced extreme learning machine within various feature selection algorithms. It reveals the effectiveness of OPCA-EELM2SS over PCA-EELM and ICA-EELM approaches since it shows the PSO based approach provides higher accuracy value. The results strongly suggest that the proposed hybrid approach can aid in the dimensionality reduction. However, OPCA-EELM2SS takes a large amount of time for finding the optimal value of the number of PCs as illustrated in Figure 13; the PSO-based methods applied in this work are the best way to reduce the time needed for executing the algorithm after the optimisation phase.

Finally, the experimental results show that OPCA-OEELM and OPCA-EELM2SS is a competitive approach in the classification tasks with optimal parameters, which can be employed efficiently in various fields.

Figure 13 Comparison of computation time in milliseconds across big data sets

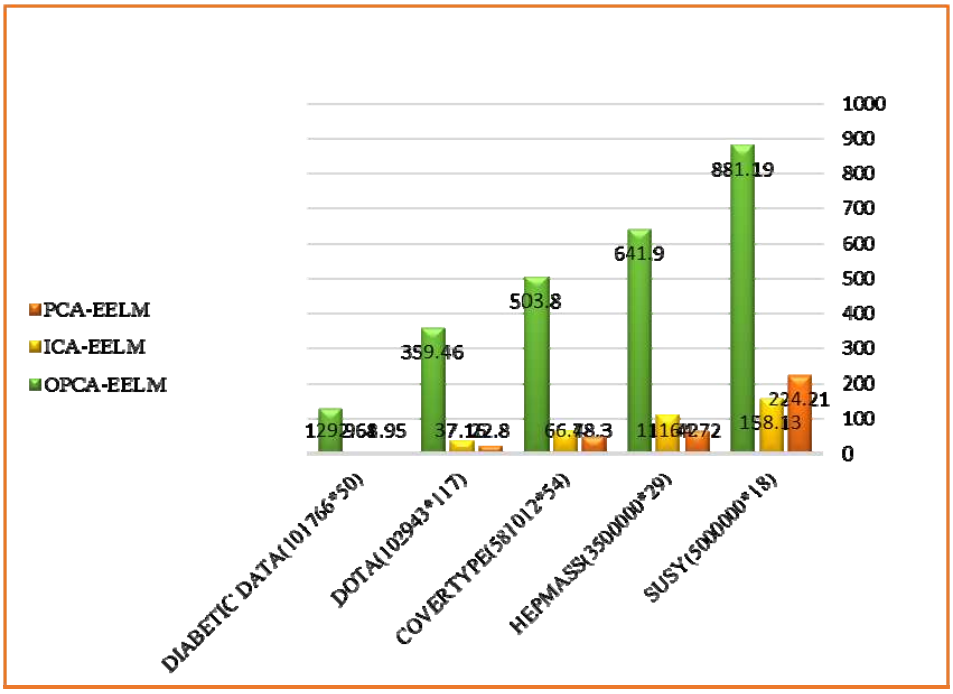
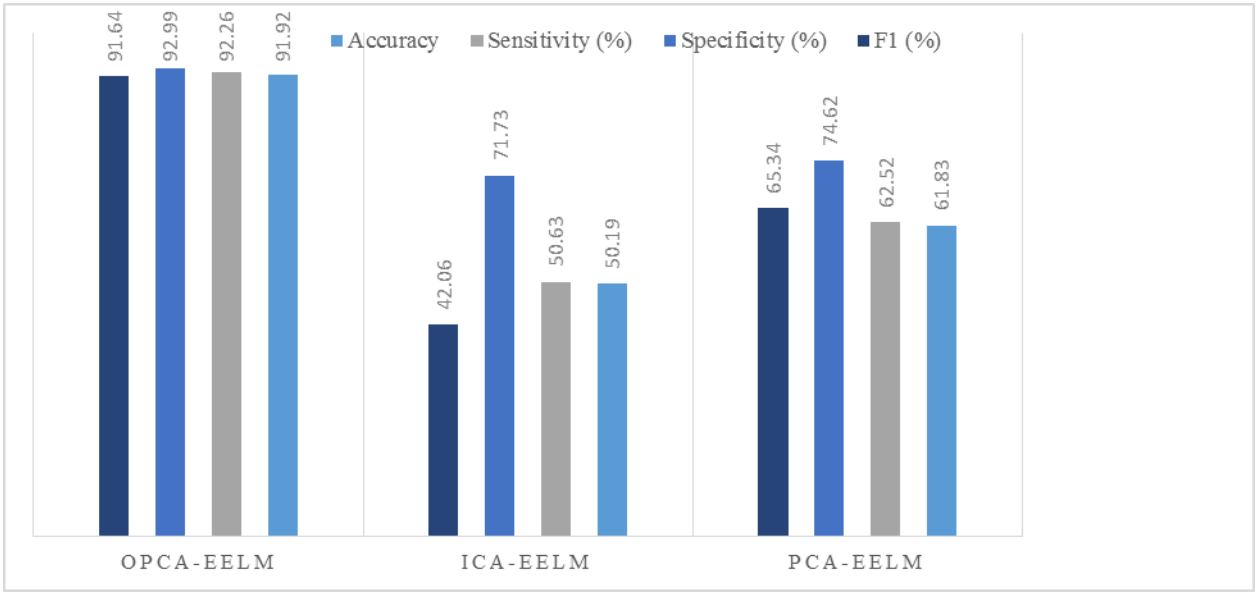


Figure 14 Mean performance parameters of OPCA-EELM2SS across big data sets



6 Conclusions and future work

Due of the excellent generalisation performance of PCA-EELM revealed from the experimental results across many data sets to withstand the computational complexity of ELM in SLFN, therefore, in this paper, two new hybrid approaches that integrate particle swarm optimisation (PSO) algorithm with the principal component analysis-enhanced extreme learning machine (PCA-ELM) for classification and reduction problems are presented for obtaining the optimisation versions of approach PCA-EELM. The first version named OPCA-OEELM is proposed for improving data processing by using PSO in two dimensions, one for determining the optimal number of principal components and the other for selecting the proper number of hidden nodes in SLFN. The limitation of this proposed approach is that all experimental results are applied on supervised classification problems, what about regression and unsupervised problems. The second approach is OPCA-EELM2SS which optimise the number of principal components for big data sets by applying the optimisation phase on a perfect stratified sample generated from the original data set. However, from experimental results of OPCA-EELM2SS obtained from good classification performance parameters across several data sets, it was concluded that this approach is also giving better results in terms of time computation compared to the results that are obtained, but, after optimisation phase. So, in the future work, optimisation phase will be executed on any parallel distributed environments such as Hadoop or Spark to overcome the time computation problem for real-world applications.

In the future, the approaches can be enhanced in the following three directions. In the first direction, PCA can be improved to treat the nonlinear dimensionality reduction issues by using nonlinear functions for PCA or integration of the kernel technique with PCA. In the second direction, this study presents the concept of improvement for reduction and classification phases to achieve the perfect analysis of data, however, for just numerical data sets, so that EELM can be improved to have the ability to process any type of data sets. In the third direction, a single layer of EELM will be replaced by a multi-layer feedforward neural network for testing the effectiveness of the multi-layer of EELM.

References

Abonyi, J. and Szeifert, F. (2003) 'Supervised fuzzy clustering for the identification of fuzzy classifiers', *Pattern Recognition Letters*, Vol. 24, No. 14, pp.2195–2207.

Akay, M.F. (2009) 'Support vector machines combined with feature selection for breast cancer diagnosis', *Expert Systems with Applications*, Vol. 36, No. 2, pp.3240–3247.

Alba, E., Garcia-Nieto, J., Jourdan, L. and Talbi, E. (2007) 'Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms', *Proceedings of the IEEE Congress on Evolutionary Computation*, No.7, pp.284–290.

Al-Rajab, M. and Lu, J. (2014) 'Algorithms implemented for cancer gene searching and classifications', *Bioinformatics Research and Applications*, pp.59–70.

Bishop, F.R. and Christopher, M. (2006) 'Pattern recognition and machine learning', *Linear Models for Classification*, Springer, pp.179–210.

Calisir, D. and Dogantekin, E. (2011) 'An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier', *Expert Systems with Applications*, Vol. 38, No. 7, pp.8311–8315.

Cao, J., Chen, T. and Fan, J. (2014) 'Fast online learning algorithm for landmark recognition based on bow framework', *IEEE Conference on Industrial and Electronics Applications*, IEEE, China.

Cao, J., Zhang, K., Luo, M., Yin, C. and Lai, X. (2016) 'Extreme learning machine and adaptive sparse representation for image classification', *Neural Networks*, Elsevier Ltd, Vol. 81, pp.91–102.

Castaño, A., Fernández-Navarro, F. and Hervás-Martínez, C. (2013) 'PCA-ELM: a robust and pruned extreme learning machine approach based on principal component analysis', *Neural Processing Letters*, Vol. 37, No. 3, pp.377–392. Doi: 10.1007/s11063-012-9253-x.

Cheung, N. (2001) *Machine Learning Techniques for Medical Analysis*. [M.S. thesis], University of Queensland.

Chyzhyk, D., Savio, A. and Graña, M. (2014) 'Evolutionary ELM wrapper feature selection for Alzheimer's disease CAD on anatomical brain MRI', *Neurocomputing*, Vol. 128, pp.73–80.

Das, R., Turkoglu, I. and Sengur, A. (2009) 'Effective diagnosis of heart disease through neural networks ensembles', *Expert Systems with Applications*, Vol. 36, No. 4, pp.7675–7680.

Dornaika, F. and Aldine, I.K. (2015) 'Decremental sparse modeling representative selection for prototype selection', Pattern Recognit, Vol. 48, No. 11, pp.3714–3727. Doi: 10.1016/j.patcog.2015.05.018.

Elbably, D.L. and Fouad, K.M. (2018) 'A hybrid approach for improving data classification-based on PCA and enhanced ELM', *International Journal of Advanced Intelligence Paradigms*, Vol. 10, No. 1. Doi: 10.1504/IJAIP.2018.10013881.

Feng, G., Huang, G.B., Lin, Q. and Gay, R. (2009) 'Error minimized extreme learning machines with the growth of hidden nodes and incremental learning', *IEEE Transactions on Neural Network*, Vol. 20, No. 8, pp.1352–1357.

Foody, G.M. (2002) 'Status of land cover classification accuracy assessment', Remote Sensing of Environment, Vol. 80, No. 1, pp.185–201.

Fouad, K.M. (2018) 'A hybrid approach of missing data imputation for upper gastrointestinal diagnosis', *International Journal of Advanced Intelligence Paradigms*, Forthcoming articles.

Gani, A. Siddiqa, A., Shamshirband, S. and Hanum, F. (2015) 'A survey on indexing techniques for big data: taxonomy and performance evaluation', *Knowledge and Information Systems*, Vol. 46, No. 2, pp.1–44.

García-Pedrajas, N. and de Haro-García, A. (2014) 'Boosting instance selection algorithms', *Knowledge-Based Systems*, Vol. 67, pp.342–360. Doi: 10.1016/j.knosys.2014.04.021.

Goodman, D.E., Boggess, L.C. and Watkins, A.B. (2002) 'Artificial immune system classification of multiple-class problems', *Proceedings of the Artificial Neutral Networks in Engineering Conference (ANNIE'02)*, pp.179–184.

Grzegorowski, M. and Stawicki, S. (2015) 'Window-based feature extraction framework for multi-sensor data: a posture recognition case study', Paper presented at *the* Federated Conference on *Computer Science and Information Systems (FedCSIS)*, IEEE, Poland.

Hassan, A.R. and Haque, M.A. (2015) 'Computer-aided gastrointestinal hemorrhage detection in wireless capsule endoscopy videos', *Computer Methods and Programs in Biomedicine*, Vol. 122, No.12, pp.341–353.

Hassan, A.R. and Subasi, A. (2016) 'Automatic identification of epileptic seizures from EEG signals using linear programming boosting', *Computer Methods and Programs in Biomedicine*, Vol. 136, No. 11, pp.65–77.

Hassan, A.R., Siuly, S. and Zhang, Y. (2016) 'Epileptic seizure detection in EEG signals using tunable-Q factor wavelet transform and bootstrap aggregating', *Computation Methods Programs in Biomed*icine, Vol. 137, No. 12, pp.247–259.

Heckenbergerova, J., Musilek, P., Marek, J. and Rodway, J. (2014) 'Principal component analysis for evaluation of wind ramp event probability,' *Electrical Power and Energy Conference*, IEEE, Canada.

Hernandez-Leal, P., Carrasco-Ochoa, J., Martínez-Trinidad, J. and Olvera-López, J. (2013) 'Instance rank based on borders, for instance, selection', *Pattern Recognition*, Vol. 46, No. 1, pp.365–375. Doi: 10.1016/j.patcog.2012.07.007.

Hervás-Martínez, C., Martínez-Estudillo, F.J. and Carbonero-Ruz, M. (2008) 'Multilogistic regression by means of evolutionary product-unit neural networks', *Neural Networks*, Vol. 21, No. 7, pp.951–961.

Huang, G.B. (2013) *Extreme Learning Machine*, Springer.

Huang, G.B. (2014) 'An insight into extreme learning machines: random neurons, random features, and kernels', *Cognitive Computation*, Vol. 6, No. 3, pp.3376–390.

Huang, G.B., Chen, L. and Siew, C. (2006) 'Universal approximation using incremental constructive feedforward networks with random hidden nodes', *Neural Network*, Vol. 17, pp.879–892.

Huang, G.B., Li., M.B., Chen, L. and Siew, C. (2008) 'Incremental extreme learning machine with fully complex hidden nodes', *Neurocomputing*, Vol. 71, pp.576–583.

Huang, G.B., Zhu, Q.Y. and Siew, C.K. (2006) 'Extreme learning machine: theory and applications', *Neurocomputing*, Vol. 70, No.1–3, pp.489–501.

Huang, Y. and Lai, D. (2012) 'Hidden node optimization for extreme learning machine', *ASSRI Conference on Modeling, Identification, and Control*, Vol. 3, pp.375–380. Doi: 10.1016/j.aasri.2012.11.059.

Huo, X.M. and Smith, A.K. (2008) 'A survey of manifold-based learning methods', *Mining of Enterprise Data*, pp.691–745.

Iosifidis, A. (2015) 'Extreme learning machine based supervised subspace learning', *Neurocomputing*, Vol. 167, pp.158–164.

Iosifidis, A., Tefas, A. and Pitas, I. (2015) 'Drop ELM: fast neural network regularization with dropout and dropconnect', Neurocomputing, Vol. 162, pp.57–66.

Jaganathan, P. and Kuppuchamy, R. (2013) 'A threshold fuzzy entropy-based feature selection for medical database classification', *Computers in Biology and Medicine*, Vol. 43, No. 12, pp.2222–2229.

Jin X., Wah, B.W., Cheng, X. and Wang, Y. (2015) 'Significance and challenges of big data research', *Big Data Research*, Vol. 2, No. 2, pp.59–64.

Kahramanli, H. and Allahverdi, N. (2008) 'Design of a hybrid system for the diabetes and heart diseases', *Expert Systems with Applications*, Vol. 35, Nos. 1/2, pp.82–89.

Karabatak, M. and Ince, M.C. (2009) 'An expert system for detection of breast cancer based on association rules and neural network', *Expert Systems with Applications*, Vol. 36, No. 2, pp.3465–3469.

Kirar, J.S. and Agrawal, R.K. (2017) 'Composite kernel support vector machine-based performance enhancement of brain-computer interface in conjunction with spatial filter', *Biomedical Signal Processing and Control*, Vol. 33, No. 3, pp.151–160.

Li, S., You, Z-H., Guo, H., Luo, X. and Zhao, Z-Q. (2016) 'Inverse-free extreme learning machine with optimal information updating', *IEEE Transactions on Cybernetics*, Vol. 46, No. 5, pp.1229–1241.

Lohr, S. (2008) 'The age of big data', *New York Times*, Vol. 16, No. 4, pp.10–15.

Marcano-Cedeño, A., Quintanilla-Domínguez, J. and Andina, D. (2011) 'WBCD breast cancer database classification applying artificial metaplasticity neural network', *Expert Systems with Applications*, Vol. 38, No. 8, pp.9573–9579.

Martínez-Estudillo, F.J., Hervás-Martínez, C., Gutierrez, P.A. and Martínez-Estudillo, A.C. (2008) 'Evolutionary product-unit neural networks classifiers', *Neurocomputing*, Vol. 72, Nos. 1/3, pp.548–561.

Matias, T., Souza, F., Aráujo, R. and Antunes, C.H. (2014) 'Learning of a single hidden layer feedforward neural network using an optimized extreme learning machine', *Neurocomputing*, Vol. 129, pp.428–436.

Mohamad, M.S., Omatu, S., Deris, S., Misman, M.F. and Yoshioka, M. (2009) 'A multi-objective strategy in genetic algorithms for gene selection of gene expression data', *Artificial Life and Robotics*, Vol. 13, No. 2, pp.410–413.

Nanni, L. and Lumini, A. (2011) 'Prototype reduction techniques: a comparison among different approaches', *Expert Systems with Applications*, Vol. 38, No. 9, pp.11820–11828. Doi: 10.1016/j. eswa.2011.03.070.

Polat, K. and Gunes¸ S. (2007) 'An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease', *Digital Signal Processing*, Vol. 17, No. 4, pp.702–710.

Polat, K., Gunes¸ S. and Arslan, A. (2008) 'A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine', *Expert Systems with Applications*, Vol. 34, No. 1, pp.482–487.

Polat, K., Sahan, S., Kodaz, H. and Günes, S. (2005) 'A new classification method to diagnosis heart disease: supervised artificial immune system (AIRS)', *Proceedings of the Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'05)*. Doi: 10.1109/SIU.2005.1567647.

Porkodi, R. and Suganya, G. (2015) 'A comparative study on classification algorithms in data mining using microarray dataset of colon cancer', *International Journal of Advance Research Computational Sciences Sowftware Engineerin*, Vol. 5, No.5, pp.1768–1777.

Raihana, F., Jailani, R., Hassan, S. and Tahir, N. (2016) 'Analysis of sparse PCA using high dimensional data', *Proceedings of the IEEE 12th International Colloquium on Signal Processing and its Applications (CSPA2016)*, 4–6 March 2016, IEEE, Melaka, Malaysia.

Reibnegger, G., Weiss, G., Werner-Felmayer, G., Judmaier, G. and Wachter, H. (1991) 'Neural networks as a tool for utilizing laboratory information: comparison with linear discriminant analysis and with classification and regression trees', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 88, No. 24, pp.11426–11430.

Rizvi, S., Mohammadpour, J., Tóth, R. and Meskin, N. (2016) 'A kernel-based PCA approach to model reduction of linear parameter-varying systems', *IEEE Transactions on Control Systems Technology*, Vol., 24, pp.1883–1891.

Sahan, S., Polat, K., Kodaz, H. and Gunes, S. (2007) 'A new hybrid method based on fuzzy-artificial immune system and *k*-algorithm for breast cancer diagnosis', *Computers in Biology and Medicine*, Vol. 37, No. 3, pp.415–423.

Salmon, J., Harmany, Z., Deledalle, C-A. and Willett, R. (2014) 'Poisson noise reduction with non-local PCA', *Journal of Mathematical Imaging and Vision*, Vol. 48, No.2, pp.279–294.

Sarveniazi, A. (2014) 'An actual survey of dimensionality reduction', *American Journal of Computational Mathematics*, Vol. 4, No. 2, pp.55–72.

Shah, S. and Kusaik, A. (2007) 'Cancer gene search with data-mining and genetic algorithms', *Computers in Biology and Medicine*, Vol. 37, No. 2, pp.251–261.

Sharma, N. and Saroha, K. (2015) 'A novel dimensionality reduction method for cancer dataset using PCA and feature ranking', *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, India.

Snijders, C., Matzat, U. and Reips, U.D. (2012) 'Big data: big gaps of knowledge in the field of internet science', *International Journal of Internet Science*, Vol. 7, No. 1, pp.1–5.

Subbulakshmi, C.V. and Deepa, S.N. (2015) 'Medical dataset classification: a machine learning paradigm integrating particle swarm optimization with extreme learning machine classifier', *Scientific World Journal*, Doi: 10.1155/2015/418060.

Tang, Y. (2013) 'Deep learning using linear support vector machines', *International Conference on Machine Learning*, Atlanta, Georgia, USA, pp.1–5.

Temurtas, H., Yumusak, N. and Temurtas, F. (2009) 'A comparative study on diabetes disease diagnosis using neural networks', *Expert Systems with Applications*, Vol. 36, No. 4, pp.8610–8615.

UCI Repository (2017) Machine learning repository. Available online at: https://archive.ics.uci.edu/ml/datasets.php (accessed on January 2017).

Wang, H., Chen, M., Shi, X. and Li, N. (2016) 'Principal component analysis for normal-distribution-valued symbolic data', *IEEE Transactions on Cybernetics*, Vol. 46, No. 2, pp.356–365.

Wu, Z., Li, Y., Li, J., Xiao, F. and Wei, Z. (2015) 'Parallel and distributed dimensionality reduction of hyperspectral data on cloud computing architectures', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 9, No. 6, pp.2270–2278.

Yadav, C., Wang, S. and Kumar, M. (2013) 'Algorithm and approaches to handle large Data', *IJCSN International Journal of Computer Science and Network*, Vol. 2, No. 3, pp.2277–5420.

Zhai, Y., Ong, Y.S. and Tsang, I.W. (2014) 'The emerging big dimensionality', *Computational Intelligence Magazine*, IEEE, Vol. 9, No. 3, pp.14–26.

Zhang, Y., Cai, Z., Wu, J., Wang, X. and Liu, X. (2015) 'A memetic algorithm based extreme learning machine for classification', *International Joint Conference on Neural Networks (IJCNN)*, IEEE, Ireland.

Zhang, Y., Wu, J., Cai, Z., Zhang, P. and Chen, L. (2016) 'Memetic extreme learning machine', *Pattern Recognition*, Vol. 58, pp.135–148.