IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 19, 2020, accepted August 24, 2020, date of publication September 9, 2020, date of current version September 23, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3022865*

# Did They Sense It Coming? A Pipelined Approach for Tsunami Prediction Based on Aquatic Behavior Using Ensemble Clustering and Fuzzy Rule-Based Classification

**NIKITA JAIN[1,2], DEEPALI VIRMANI[3], AJITH ABRAHAM[4], (Senior Member, IEEE), LORENZO SALAS-MORERA[5], (Member, IEEE), AND LAURA GARCIA-HERNANDEZ[5], (Member, IEEE)**

[1]University School of Information, Communication and Technology, GGSIPU, New Delhi 110078, India
[2]Bharati Vidyapeeth's College of Engineering, New Delhi 110063, India
[3]Department of Computer Science, Bhagwan Parshuram Institute of Technology, New Delhi 110089, India
[4]Machine Intelligence Research Labs (MIR Labs), Washington, WA 98071, USA
[5]Area of Project Engineering, University of Córdoba, 14071 Córdoba, Spain

Corresponding author: Nikita Jain (nikitajain1210@gmail.com)

This work was supported by the Junta de Andalucía, under Project UCO-FEDER 18 REF and Project 1265277 MD A1.

**ABSTRACT** Tsunami is one of the real feelings of dread among humanity. Designing an early and effective Tsunami Warning System (TWS) is an immediate goal, for which the scientific community is working. Underwater seismic responses sensed by different numerical expository techniques have resulted in various cautionary frameworks proving successful in predicting tsunamis. However, multiple instances in the past where these warning systems have failed to generate alerts in time, has raised concerns to design even more efficient, diverse, and multidisciplinary warning methods or systems. However, there have been many instances in the past where these warning systems have failed to generate alerts in time, raising concerns about designing/implementing more efficient, diverse, and multidisciplinary warning methods or systems. Therefore, we propose a sequenced ($EC_GF_C$) approach for designing a TWS, based on Ensemble Clustering ($EC_G$) and Classification for categorizing anomalous behavior in response to seismic perturbations, taking three aquatic animal behavioral datasets: Turtle, Earthworm, and Fish, as the input(s). $EC_G$ uses an existing state-of-the-art method bagged with Gaussian mixture model to label the dynamically changing behavioral data. The paper compares the results of the clustering ensemble used with baseline clustering methods on three behavior datasets as well as four benchmark datasets. The proposed sequenced ($EC_GF_C$) method is finally compared on three classification error metrics: MSE, MAE, and SMAPE on behavioral and existing ensemble frameworks in the state-of-the-art.

**INDEX TERMS** Tsunami warning system, ensemble clustering, FRBCS, behavioral data, tsunami, alert.

## I. INTRODUCTION

The 2004 Indian Ocean tsunami, popularly known as ''Boxing Day Tsunami'' was marked as one of the most devastating events in the history of disaster science (because of high underwater seismic activity). West course of Sumatra or Indian Ocean being the epicenter of the event, Indian communities suffered tremendous loss of life and property [1]. Since repeated occurrences of tsunami have affected countries like India, Srilanka, Japan, Thailand, and Indonesia, scientists and practitioners are taking cues from pre and post analysis of various tsunami events. This analysis has been presented in the form of analytical studies, algorithms, methods, simulations, and models describing the occurrence, prediction, or impact of such events. Different Tsunami Warning Systems (TWS) have been proposed, developed in research, and then deployed to predict seismic perturbations. The state-of-the art has broadly classified these systems as Physiological (based on geophysics) [2]–[4], Societal (based on inter-human interaction), [5]–[7] and Nature-based (based on

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

ecology) [8], [9]. While the former two, primarily rely on predicting seismic signals based on mathematical and computational analysis, the latter focuses on analyzing societal as well as nature's response towards seismic disturbances. Despite various TWS, there have been instances and observations prevailing in the literature depicting how this deadly event has incurred loss to humanity and nature [10]–[12]. These observations have underscored that TWS has struggled to produce timely warnings in the past. Accordingly, as indicated by the authors in [13], there is a need for advancement in tsunami science where approaches paving various disciplines, i.e. geophysics and ecology, can be more successful when appropriate.

Ecology, including plant and animal population, has responded in the form of unusual signals and responses to changing underwater conditions under seismic tremors. Animals can act as bio-sensors to predict natural disasters such as tsunami [14]. Reports [15]–[17], debated by researchers and scientists emphasize the presence of such signal or response production in aquatic animals towards seismic perturbations. Marine animals use their sensory receptors to navigate and breed. The varying underwater conditions such as oceanic flow, changing electromagnetic conditions across the flow can, therefore, affect or disrupt these sensory receptions [18]. Hence, the abnormal behavioral patterns in the form of unprocessed data can be tapped in real-time to help generate alarms after precise analysis.

The aforementioned literature citations clearly indicate that there are signs of anomaly in the behavior of underwater species whenever seismic perturbations reach the sea bed [19]–[21]. To analyze these signs, where the labels of the unprocessed data are unknown, computational intelligence analysis is needed. Hence, efficient machine learning paradigms [22], [23] can be used to classify such biological anomalous behavior datasets and help generate timed alerts. Here, the alert refers to a warning that can assist in the execution of safety measures and thereby avoid loss of life and property to humanity. Recently, a contribution [24] which reviews the impact of machine learning techniques to model complex behavioral data has attracted attention.

Machine learning techniques are categorized into two types: supervised and unsupervised. In unsupervised, clustering is one of the most critical unsupervised learning techniques that have found a wide application in data analysis [25]. The goal of clustering is to partition a dataset into several groups, such that data sampled in the same group are more similar than those in different groups [26]. Even after many improvements over existing base methods, various contributions in the state-of-the-art have proved that combining the strengths of various clustering termed as ensemble clustering (EC) can provide better insights towards data analysis [27]. EC is a process of combining a re-defined clustering method to obtain well defined and crisp partitions based on a weighted developed function [28].

The first aim of this work is to propose and evaluate an ensemble clustering algorithm ($EC_G$) based on a modified threshold K-means ($IT_KM$) followed by a bagging based ensemble of a Gaussian Mixture Model (GMM). For bagging, neural training is used that enables a weighted cluster head selection of three dynamic multi-source datasets (MSD's) coming from different aquatic species viz. *Sea Turtles(*D1), *Earthworms*(D2) *and Fish*(D3). Multi-source data (MSD) is a form of data taken from multiple sources integrated for any further analysis and inferences. For input, the three MSD's in this work are prepared by tapping the indicative behavioral attributes in response to the geophysical data which are merged under common timestamp. The motive behind such aggregation of information is to underline unlabeled trends of parallel changes in the behavior of aquatic species and varying geophysical conditions. This unlabeled data can be labeled and analyzed using ensemble clustering where soft clustering methods like GMM coupled with weight adjusted neural training can identify unknown nonlinear dependencies and interactions, across multiple variables and cluster them into groups. Supervised GMM helps in overcoming the (which helps in learning the class labels which were otherwise unknown) slow converging nature of unsupervised GMM and yet use its ability to cluster probabilistic data. Hence, pipelining an optimized K-means with GMM and neural training feedback with FRBCS, this is a scalable approach, yet efficient for alert classification forms the main rationale behind this work.

Though, using GMM for capturing semantic relationships and annotate behavioral data has been proved resilient against missing data and inaccuracies in [29], [30], combining output from such semantic relationships with seismic alert possibilities is a new attempt.

The second aim of this work is to focus on finding different settings that can classify more data by learning these relationships from the clustered groups. Guided by unsupervised clustering, a further pipelined classification can provide better insights for unlabeled data [31]. In this paper, a Fuzzy-rule–based-classification system (FRBCS) is pipelined after the clusters are labeled and hence a generic approach ($EC_GF_C$) is proposed. The aim behind using FRBCS is to understand the continuous probabilistic behavior of data within the clusters. The approach uses an optimized K-Means [79] method based on a sorted and dynamic centroid allocation technique Also, FRBCS provide a basis of converting human observed linguistic variables into a well-defined knowledge rule base. Such rule base can be used to design generic mathematical model that can flag alert situations based on any specific specie behavioral data. However, in this paper we evaluate the proposed generic approach for classification metrics (refer Section: Results).

The datasets analyzed here contains behavioral data values for days closer to tsunami days of the year 1997 for the Netherlands and 2004 for India. To understand intrinsic relationships between the existing data points, a class label is needed for mapping any consequent behavior as a response to seismic perturbations. Such response if observed well in time can easily help in generate timely alerts.

After the application/implementation of the proposed approach, the new clustered labels are categorized into two class labels i.e. one, which is the label for alert (having days closer to tsunami days) and zero for no-alert, which is a normal adaptive behavior. This paper provides an efficient pipelined (ensemble clustering & classification) approach to classify underwater specie behavior and use them as a precursor for future alerts. The proposed pipelined approach is initially evaluated on prepared datasets, and under 25 different settings. These settings are obtained by pipelining clustering and fuzzy rule based classification systems.

To ensure effectiveness, apart from evaluating the proposed algorithm ($EC_G$) on the prepared behavioral MSDs, the performance of the same has been tested on four benchmark datasets. The findings of this were outperformance of the initially proposed algorithm ($EC_G$) on benchmark datasets (refer Section: Results) in comparison to baseline clustering methods and the existing state-of-the-art method [33]. Here, the cluster statistical analysis based on Silhouette, Rand and Dunn Index [34] is used to provide empirical evidence to the results. The recent contributions [35]–[37] have used these benchmark datasets primarily for statistical clustering analysis and hence they form one of the bases of analysis in this article as well.

In addition, a comparison on six other benchmark datasets with existing ensemble clustering classification approaches is also drawn. The benchmark datasets selected here consist of three small scales and three large scale datasets, where the former comes from [32], identified as hard datasets. The large scale dataset comparison adds to the scalability property of the proposed approach.

The paper is an implementation of the scientific patent [38] published by the authors of this manuscript with the following novel contributions:

- An ensemble clustering algorithm ($EC_G$) is proposed and implemented based on modified threshold K-means ($IT_KM$) for labeling MSD's of three different aquatic species.
- The ($EC_G$) algorithm is a bagging based ensemble where a weighted cluster head selection is performed using a Gaussian Mixture Model and Neural Training.
- The existing baseline FRBCS have been sequenced to the above obtained grouped data for seismic alert classification.
- Real-time prediction using unlabeled data acquired. Till date, only unlabeled datasets are available – a transformation from unsupervised to supervised data

The rest of the paper is organized as follows: Section II describes the related work. Section III initially describes the dataset collection and pre-processing, followed by the base state-of-the-art methods for behavioral data clustering, and finally, the proposed methodology. The results from applying the base, as well as proposed methods on behavioral datasets, are shown in Section IV. The article concludes with Section V, which summarizes the main conclusions, identifies areas for future work

## II. RELATED WORKS

Various classifications, as well as clustering algorithms, have been used for categorizing animal behavioral data. Understanding animal behavior (specifically for aquatic species) for any ambiguity can be challenging. One of the reasons that account for such a challenge is that the data collection for identifying anomalous behavior trends needs continuous data recordings [39] from multiple sources. The lack of any such continuous records and multi-source data fusion, therefore, leads to an incomplete and missing value data set where a robust yet efficient learning algorithm is needed to classify or categorize behavior as per the constraints. The subsequent subsections explore the existing models in the respective domains extensively. Table 7, which summarizes the essential aspects, is also further discussed in conclusion, shows that the Ensemble of Clustering pipelined with a classification method for behavioral analysis goes beyond what was done in prior studies.

### A. ANIMALS AS TSUNAMI PRECURSORS

Animals, both terrestrial and aquatic, have shown ambiguous behavior before and after various natural disasters. Various studies have highlighted the post- tsunami impact on the animal population [40]. However, due to lack of deployed sensors and hence unavailability of data, pre-tsunami impact analysis on animals remains an open area. Various studies have discussed that certain animals like fish [41], toads [42], elephants [43], and whales [44] have shown unusual behavior as they could sense the pre-tsunami signals which humans and machines could not. Therefore, a global warning system on animal behavior data analysis that doesn't exist in the current-state-of-the-art to generate tsunami alerts is a dire need for progress in tsunami science.

### B. MACHINE LEARNING FOR ANIMAL BEHAVIOR CLASSIFICATION

Various machine learning algorithms have found applications in animal behavioral classification. Some on terrestrial animals such as cattle [68], sheep [69], while some for aquatic species [70], [71]. Followed by basic machine learning algorithms, some ensembles have also been applied for behavioral classification [72]. These applications have mainly been related to in identifying grazing or migration patterns. Hence analyzing change in aquatic animal activity due to seismic perturbations can be open area this article aims to explore.

### C. ENSEMBLE MULTI-SOURCE CLUSTERING: INSIGHTS

In remote sensing, data is tapped from multiple sources. Need for soft clustering methods to label such multi-view data is inevitable. The success of ensembles application in supervised classification tasks has motivated researchers to use the same in unsupervised tasks [45]. Lack of guidelines that define the selection of any individual clustering algorithms still exists. Jointly mining clusters from multiple data sources has been emerging as a novel direction in the domain of clustering analysis. Fern and Lin [46] designed three ensemble selection methods based on quality and diversity. Hong *et al.* [47] introduced a novel selective clustering ensemble method through resampling. Azimi and Fern [48] proposed an adaptive cluster ensembles method. However, designing a well-weighted consensus function is essential to clustering ensembles. Apart from voting [83], [98], bagging ensembles have also given promising results. Tsymbal *et al.* [50] presented an iterative clustering through a weighted scheme that outperformed many other selective voting or bagging methods. Another method, such as spectral clustering [51], which labels communities based on graphical linkages, is evaluated to be computationally expensive and hence needed a bagging method for quality improvement in label formation. The used ensemble here is different from Multi-view clustering (MVC) [52] as the former aims to find the cluster structure shared by multiple views of a particular dataset. From various improved clustering methods [53]–[56] proposed and discussed one of the methods for unsupervised labeling with a low computational cost is the Gaussian Mixture Model (GMM). Recently, GMM has found applications in various areas [57], [58] and hence is forms a basis in the proposed $EC_G$ method of this article as well.

### D. FRBCS FOR ANIMAL BEHAVIORAL CLASSIFICATION

Fuzzy models are used when a system cannot be defined in precise mathematical terms. The non-fuzzy or traditional representations require a well-structured model and well-defined model parameters. However, in practice, there may be uncertainties, unpredicted dynamics, and other unknown phenomena that cannot be mathematically modeled. The main contribution of the fuzzy modeling theory is its ability to handle many practical problems that cannot be adequately represented by conventional methods [97]–[102].

In this work, the input to the proposed method is a multiple animal behavioral dataset (D1, D2, and D3) exhibiting complex nature that can be modeled using fuzzy relations and rules [59]. Various methods have been used in the past, forming the optimized structure basis of the fuzzy model thus developed [60]–[62]. Since they are limited to specific objective functions, specific types of inference, and specific types of membership functions, this paper deals with some standard structured methods based on partitioning, and genetic algorithms have been used for FRBCS modeling [65]. A combination of clustering with FRBCS has found applications in various fields [66], [67].

### E. PIPELINING CLUSTERING & CLASSIFICATION

Under various existing algorithms, SVM and Random forests have been extensively used for animal behavioral classification. [75], [76]. The ensemble nature of Random forests, which use bagging and bootstrapping, has been cited as one of the reasons for its better performance. As stated by authors in [24], not every learning algorithm can perform for all behavioral categories. Therefore, multiple machine learning algorithms and their ensembles are needed to study a specific behavioral category, for which existing methods may have undesired performance. Various fields, such as text classification [73], credit scoring [74], image classification [77], etc., have also used a combination of pipelining of clustering and classification.

## III. METHODOLOGY

As described in previous sections, pipelining clustering and fuzzy classification have found broad applications in the field of data analytics and is thus the approach followed in this work for seismic prediction analysis. The GMM-based clustering ensemble ($EC_G$) used here allows to classify different relationships among the features taken and assigns alerts and no alert labels as 1 and 0, respectively The labeled clusters hence obtained, is the input to standard FRBCS for further classification hence a sequenced method is presented: $EC_GF_C$. The rule base hence received can be used as a knowledge base any further similar classification.

The first subsection describes one of the state-of-the-art clustering method already implemented in [78] from [79] and [33] on Sea Turtle behavioral data. The following subsection presents an intermediate method. The mentioned two clustering methods have been initially compared based on cluster statistical analysis (ref Section: Results). The better performing clustering method assessed using cluster indices is extended further and the $EC_G$ method is proposed, presented in the third subsection along with the algorithm implementation. Finally, the last subsection explains the sequenced approach $EC_GF_C$.

### A. IMPROVED CENTROID K-MEANS: $IC_KM$

The method $IC_KM$ improves the baseline K-means method to circumvent the latter's spherical nature. Figure 1 gives the complete workflow of the method. The adequacy of the method for labeling behavioral data lies in the fact that animals dynamically change their direction or location, along with their underwater count amid seismic perturbations. As data is continuous and complex, there is still a need for an effective algorithm to learn about cluster labels from the given unlabeled dataset. An improved clustering method is thus discussed [33] in the following subsection.

### B. IMPROVED THRESHOLD K-MEANS: $IT_KM$

This threshold-based clustering method is also another improved method that can be used to label behavioral data. The algorithm $IT_KM$ as shown below, takes marine behavioral dataset *(mabD)* as one of the inputs, having dimensions: $i \times j$ in which $i =$ number of entries and j is the feature set. Another input to this method is k which is the number of clusters. This method uses two functions Weighted_Score and CalThreshold.

$C'$, which is the output of the later, finally gives clustered data points based on an improved threshold of K-Means. The workflow for the same is shown in Figure 2.

**Algorithm 1** $IT_KM$

/*This function returns $ws_i$—the weighted score for each data point in the input dataset */
/***Input:** *mabD* is the input marine behavioral dataset */
/* **Input**: $N$ is the number of data points in the dataset */
/* **Output**: New set of refined cluster centers as $C'$*/
**function** Weighted_Score (*mabD*, N)
  array dp = [$dp_1$,$dp_2$,$dp_3$ . . . .$dp_m$] in *mabD*
  array N = [n1,n2,n3 . . . .nj] in *mabD*
  **for** i in 1 to n
    $w_i = \sum_{i=1}^{N} dp_i/N$
  **end for**
  **for** i in 1 to n
    $ws_i = \sum_{i=1}^{N} w_i * dp_i$
  **end for**
  **return** $ws_i$
**end function**
/*This function allocates each data point according to the newly identified centers */
**function** CalThreshold ($ws_i$, N, k)
  **for** i in 1 to k //here k defined the number of clusters
    **for** j in 1 to N //here N is data points
      centers $c_i$ = Threshold_Calculation (N, k)
      $min \sum_{i=0}^{k} \sum_{dp \in c_i}^{N} |dp - \frac{1}{|c_i|} \sum_{ws_j \in c_i} ws_j|$ //allocate each $c_1 \cup c_2 \cup c_3 \ldots \rho$ data point to the closest center
    **end for**
  **end for**
**end function**
/*This function calculates the threshold for each data point to allocate new centers */
**function** Threshold_Calculation (N, k)
//F (-U) = max(y such that #{r∈ $F$|r >=y} = U)
// F (U) = {y∈ F | y >= F (-U)}where X = N
  **for** i in 1 to n
    **if** $F_{max} < ws_i$
      set $F_{max} = ws_i$
    **end if**
  **end for**
  **for** i in 1 to n
    **if** $F_{min} > ws_i$
      set $T_{min} = ws_i$
    **end if**
  **end for**
  $T_{HV}$ (threshold value) = $(F_{max} - F_{min})/2$
  $C_1 = (F_{max} - T_{hv})/2$
  $C_2 = (T_{hv} - F_{min})/2$
  **for** i in 1 to K //where k is the number of clusters
    **for** j in 1 to $K_i$ the number of objects of the cluster i
      **return** G {$C_1$, $C_2$, $C_3$ . . . . . . $C_k$} = $K_i F_{ij-Oi}$
//$F_{ij}$ is the j-th object of the i-th cluster, and $O_i$ is the centroid of the i-th cluster, which is defined.
    **end for**
  **end for**
**end function**

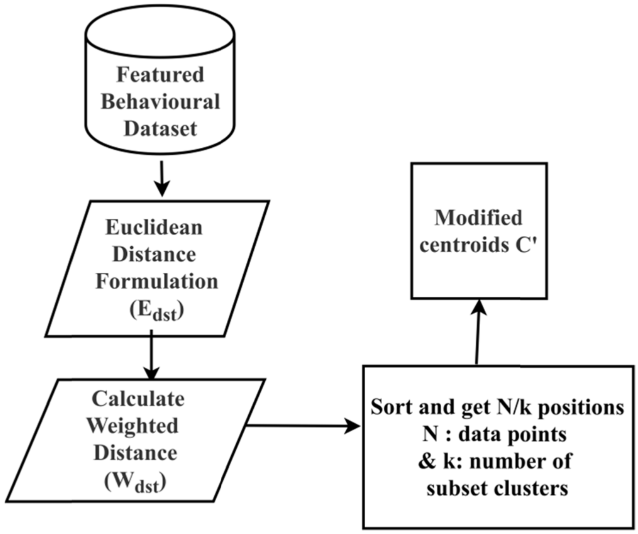


**FIGURE 1.** $IC_KM$ Workflow.

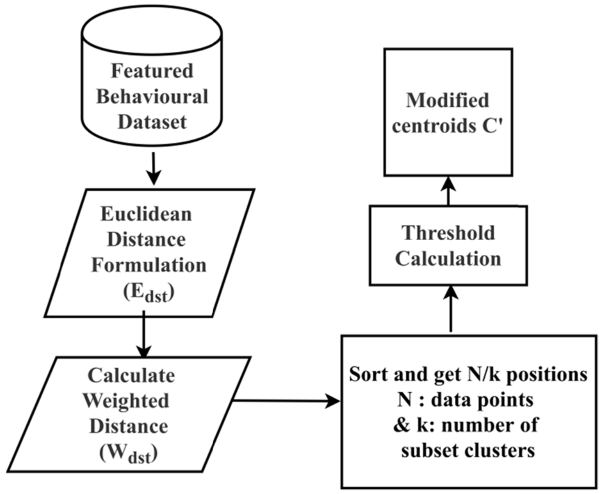


**FIGURE 2.** $IT_KM$ workflow.

### *C. PROPOSED ENSEMBLE CLUSTERING METHOD: $EC_G$*

This ensemble clustering technique uses the GMM, followed by the Fuzzy C means clustering. To illustrate, here, datasets D1 (Sea Turtle), D2 (Earthworms), and D3 (Fish) have animal behavioral feature values attributed in response to geophysical changes. An analysis of data can help to obtain the period, which was either alert or no-alert prone based on several parameters. The method at first identifies the data points pertaining to a particular period represented using weighted scores, evaluating whether the data point resides above the threshold level of risk or is below the threshold level. This computation follows from the previous algorithm $IT_KM$. The workflow is for the proposed method is shown in Figure 3. As, in this paper we deal with a fuzzy probabilistic data having data points of probability belonging to multiple clusters at a time, hence a soft clustering method is used. In terms of time taken the method GMM suffers from slow convergence and takes more time to cluster data as compared to K-Means even on small datasets. To speed up the process, a supervised

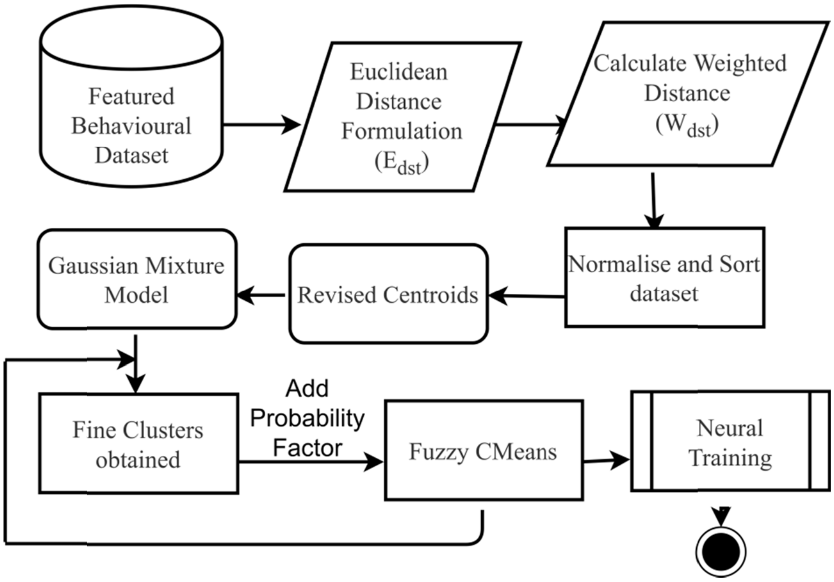


**FIGURE 3. Ensemble clustering method $EC_G$ workflow.**

GMM method can be used. By having a prior knowledge about cluster/class labels, GMM can cluster data in a time comparable to a baseline K-means method.

The algorithm $EC_G$ as shown below later refines the centroids of the classes and the belongingness of the data points with cluster centers and threshold levels that can help in the proper classification. The output of the clustering algorithm $IT_KM$ at, first step and the combined probability of belongingness of the data point to a cluster is the input to the GMM. GMM refines the cluster centers by using the threshold, the probability of belongingness of a data point, and the objective function. In the following step, the GMM model's output is allowed to pass through the fuzzy C means, which takes into consideration the probabilities as well as the fuzzy rules formed. In the last step, the cluster output is used to train the neural network helping improve the clustered data.

The ensemble here iteratively clusters the data point that is nearer to the centroid and has a high value of probability for belonging to that cluster whose centroid it is closer to and low likelihood for fitting to the latter cluster. The individual labels can be analyzed later on to check which parameter of the data point played a dominant role in its move towards a particular class label. For simulation, the R studio platform is used to implement the proposed $EC_G$ method. R environment is an open-source platform that allows re-implementing existing packages [80] to devise new ensembles as per the data analysis needed.

### D. PROPOSED SEQUENCED METHOD: $EC_GF_C$

Fuzzy rule-based classification systems have become a powerful tool in mining inferences from complex real-world problems using fuzzy concepts.

As animal behavioral data depicts non-linearity, FRBCS has been used to deal with such behavioral data values [81]. However, this approach may not be feasible when facing complex tasks or when human experts are not available. An effective alternative is to generate the FRBCS model automatically from data by using learning methods. Many methods have been proposed for this learning task, having clustering methods [82] as one of them. Hence, in this article, the approach $EC_GF_C$ uses the proposed method $EC_G$ (refer subsection: C) and FRBC for classification tasks. The FRBCS based on the default parameters are used for further classification, as shown in Table 1.

For the simulation requisites, the proposed pipelined approach workflow, as shown in Figure 4, is implemented using the R studio platform. For cluster statistics and baseline methods, R platform packages have been used.

Step 1: *Preprocessing*

The raw data for the three species mentioned is preprocessed to convert it into an attributed dataset (D1, D2, and D3)

Step 2: *$IT_KM$ & $EC_G$ application on Featured dataset*

The two methods $IC_KM$ (refer Subsection: A) and $IT_KM$ (refer Subsection: B) are initially applied to D1, D2, and D3.

As the latter produces better cluster quality (refer Section: IV) identified by certain cluster performance indices. $IT_KM$ is further improved to give $EC_G$(refer Subsection: C).Both $IT_KM$ & $EC_G$ are applied to the featured dataset.

Step 3: *Compare and Analyze*

Evaluate both $IT_KM$ *and* $EC_G$ based on cluster statistics using the following three equations: (1,2and 3) [34], [78]

$$S[i] = \frac{smad\left(dp_i\right) - ad(dp_i)}{\max\{smad\left(dp_i\right), b(dp_i)\})} \tag{1}$$

Here S[i] is the Silhouette Coefficient for the $i^{th}$ data point where for a given data point dpi, the smallest average distance of i to all points is given by smad ($dp_i$), where $dp_i$ does not

**Algorithm 2 $EC_G$**

/∗This function returns cluster centers for the dynamic data using soft clustering based on Gaussian Mixture model ∗/
/∗C' is the cluster centers obtained from ***Algorithm***: ITkMeans [$IT_KM$] ∗/
/∗ **Input:**r is the weighted score ∗/
/∗**Input:** N is the number of data points in the dataset ∗/
/∗ **Output**: New set of refined cluster centers as C' ∗/
**function** GMM ($C'$,r, N)
  **for** i in 1 to k
    **for** j in 1 to N
      $P_i = \frac{1}{\sqrt{(rj-C'j)^2}}$
    **end for**
  **end for**
R = append (r, Pi)
Q = Weighted Score (*mabD*, R) // Invoking function Weighted Score from ***Algorithm***: $IT_KM$
$f\ (Q) = \sum_{k=1}^{K} \alpha_k f_k\ (Q)$ //GMM function $\alpha_K$ is mixing weight for k$^{th}$ component where $\sum_{k=1}^{K} \alpha_k = 1$
**return** $f(Q)$
**end function**
/∗ This function performs a fuzzy partitioning through an iterative optimization of the objective function shown ∗/
FCM ($f(Q)$, r)
  **for** i in 1 to N
    **for** j in 1 to C'
      $u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c`_j\|}{\|x_i - c`_k\|}\right)^{\frac{2}{m-1}}}$
      $J_m = \sum_{i=1}^{N}\sum_{j=1}^{C`} u_{ij}^m \left\|x_i - c`_j\right\|^2$ where $1 \leq m < \infty$
    **end for**
  **end for**
**function** Neural_Train (r, N, *mabD*)
  $e_{b=\sum_{i=1}^{n} w_b(i)I(C_b(x_i)\neq y_i))}$ //$e_b$ is the error
  $w_{b+1}\ (i) = w_b(i)\exp(\alpha_b I\ (C_b\ (x_i) \neq y_i))$ // wb-1 is adjusting of backward weights & n is number of classes and αb= 1/2ln((1−eb)/eb + ln(n−1)
**end function**

belong to that taken cluster, and the average distance between dpi and all other data within the same cluster is given by ad($dp_i$). S[i] summed over-all points is termed here as SC (Silhouette Coefficient).

$$D = \frac{min.separation}{max.diameter} \tag{2}$$

where min.separation is separation within-cluster, and max.diameter is compactness within the cluster. Here the cluster separation is between two clusters labeled as 0 for no-alert and alert as per animal behavior in response to changing physical conditions.

A rand index to compute the ratio between agreements and no- agreements of the two classes here alert and no-alert.

$$R = \frac{a+b}{a+b+c+d} \tag{3}$$

Step 4: *$EC_G$ : improved clusters*

As per the results shown in Table 3, the proposed $EC_G$ performs better in terms of cluster statistics, as compared to $IT_KM$.

Step 6: *$EC_GF_C$ : The proposed sequenced method*

Finally, the above-evaluated ensemble clustering method is further given to FRBCS which gives the pipelined approach: $EC_GF_C$. a generic framework that can classify alert signals from a dynamically changing probabilistic unlabeled data capturing aquatic behavior (which needs soft clustering methods like GMM) and further capture a reduced knowledge fuzzy rule base by adjusted neural training. Studying rule base reduction, time reduction because of neural training introduced by employing proposed approach falls to be another future idea we aim to work for. Currently, the pipelined approach which is obtained in form of 25 different arrangements by permuting six baseline FRBCS with various baseline, existing and proposed clustering methods (refer Table 5) is evaluated for classification metrics.

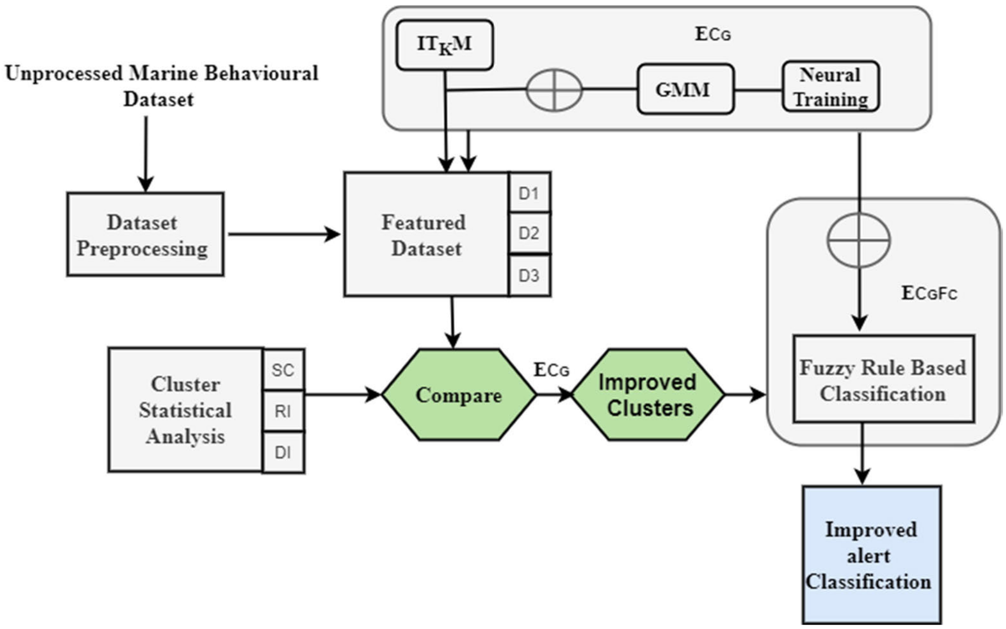


**FIGURE 4.** The workflow of complete proposed approach: $EC_GF_C$.

**TABLE 1.** Methods used for Fuzzy Rule-based Classification System with default methods.

| Method (Packag name in R) | Method Full Name | Parameter Description |
|---|---|---|
| FRBCS.W | fuzzy rule-based classification system with weight factor[65] | type.mf = "TRIANGLE", type.tnorm = "MIN", type.snorm = "MAX", type.implication.func = "ZADEH |
| FRBCS.CHI | fuzzy rule-based classification system using Chi's technique[65] | type.mf = "TRIANGLE", type.tnorm = "MIN", type.snorm = "MAX", type.implication.func = "ZADEH |
| SLAVE | structural learning algorithm on vague environment[65] | persen_cross = 0.6, persen_mutant = 0.3, max.iter = 10, max.gen = 10, num.labels, range.data.input, k.lower = 0.25, k.upper = 0.75, epsilon = 0.1 |
| GFS.GCCL | genetic cooperative-competitive learning[63] | persen_cross = 0.6, persen_mutant = 0.3, max.gen = 10 |
| FH.GBML | hybridization of genetic cooperative-competitive learning and Pittsburgh[64] | persen_cross = 0.6, persen_mutant = 0.3, max.gen = 10, p.dcare = 0.5, p.gccl = 0.5 |

## IV. RESULTS AND DISCUSSION

This section describes all aspects of data collection and pre-processing followed by the analysis and discussion.

### A. DATASET DESCRIPTION AND PRE-PROCESSING

The dataset(s) used in this work captures the behavioral activity of three underwater aquatic species. The three species are viz: sea turtles (Scientific name: *Caretta caretta*), earthworms (found on the underwater seabed), and fish. It is observed that animals or micro-organisms can behave as biological sensors to predict seismic disturbance on land and in water, based on information they receive from underground geophysical sensors, before many observable days extending up to 24 hours. [84], [19].The effect of a seismic-driven geophysical change on the three mentioned species is described in [85]. The dataset for varying geophysical values is modeled from [86], where it shows that a secondary induced magnetic field induces an electric current by Tsunami Flow.

**TABLE 2.** Sea Turtle, Earthworm and fish meta-data.

| Dataset | # of data points | # of parameters |
|---|---|---|
| Sea Turtles | 270 | 4 |
| Earthworms | 300 | 4 |
| Fish | 290 | 4 |

An intersecting time stamp fusion conducted for all three species offers the respective associated behavioral data values used in [86] for Sea Turtles and [87] for Earthworms and Fish. The aforementioned data source provides the raw latitude and longitude values in the form and the count underwater. The data is pre-processed by calculating two features Angle of Deflection (Day) and Angle of Deflection (Monthly) by using Haversine equations already used by authors in [88].

A group by clause over common species gives the underwater Count of the respective issue grouped under recurring specie id. Finally, the pre-processed dataset is mapped to varying electromagnetic values obtained under the time stamp of intersection. Table 2 shows the size of the collected datasets, while Table 3 displays the meta-data with the range of values of the data prepared for all organisms.

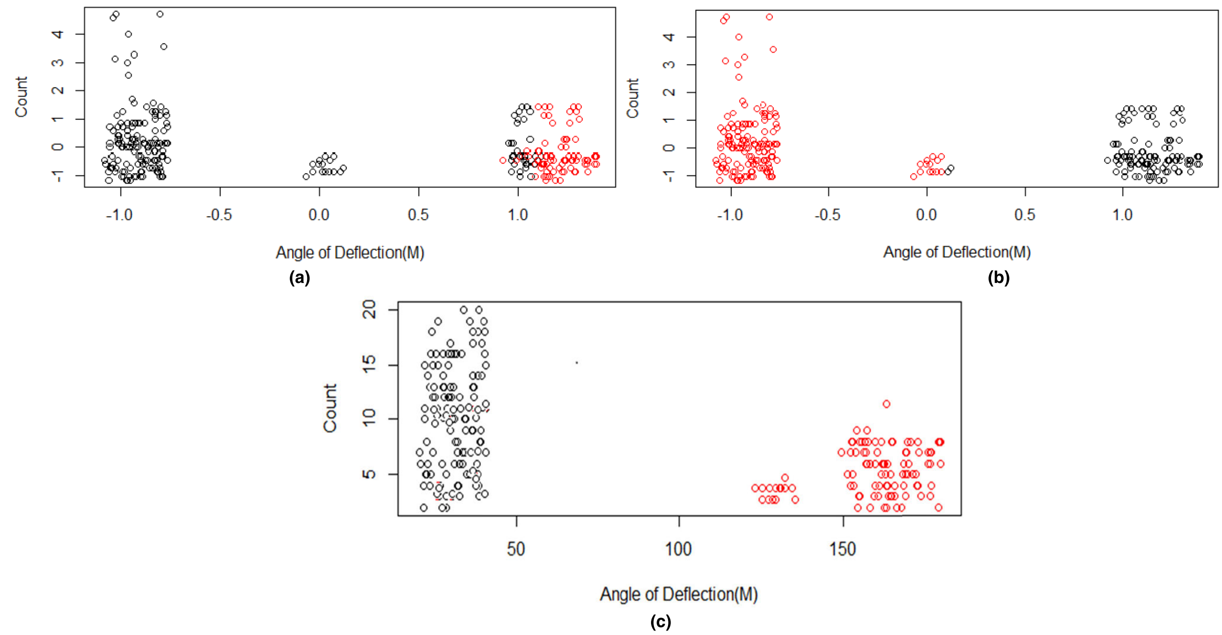


**FIGURE 5.** Shows clustering results for the two methods IT$_K$M (b) and EC$_G$ (c) used and compared on sea turtle dataset (D1) with K-Means (a). The ensemble clustering method EC$_G$ shows better clustering results visualized in terms of overlapping.

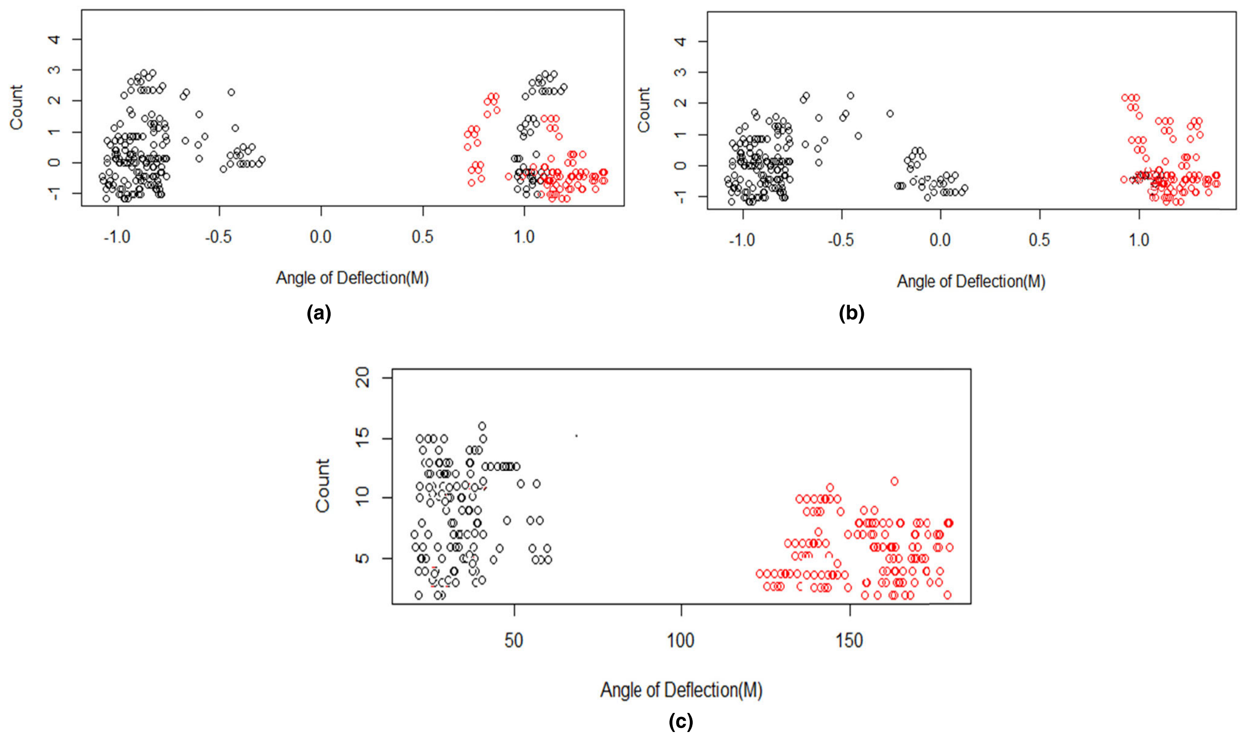


**FIGURE 6.** Shows clustering results for the two methods IT$_K$M (b) and EC$_G$ (c) used and compared on Earthworm dataset D2 with K-Means(a). The ensemble clustering method EC$_G$ shows better clustering results visualized in terms of overlapping.

### B. EC$_G$ ON BEHAVIORAL DATASETS: EVALUATION AND COMPARISON

We have evaluated the cluster statistical indices as given by equation 1, 2 and 3 for the initial method IC$_K$M, IT$_K$M as well as the proposed method (EC$_G$) on the three aquatic animal behavioral datasets Sea Turtle (D1), Earthworms (D2) and Fish (D3). IT$_K$M as explained is an intermediate method developed for further refinement achieved in the finally proposed method EC$_G$. The results show a comparison between these three clustering methods and the three baselines partitioning clustering methods viz. Hierarchical K-means, Fuzzy C-Means, and K-means [89]. Here, the proposed EC$_G$

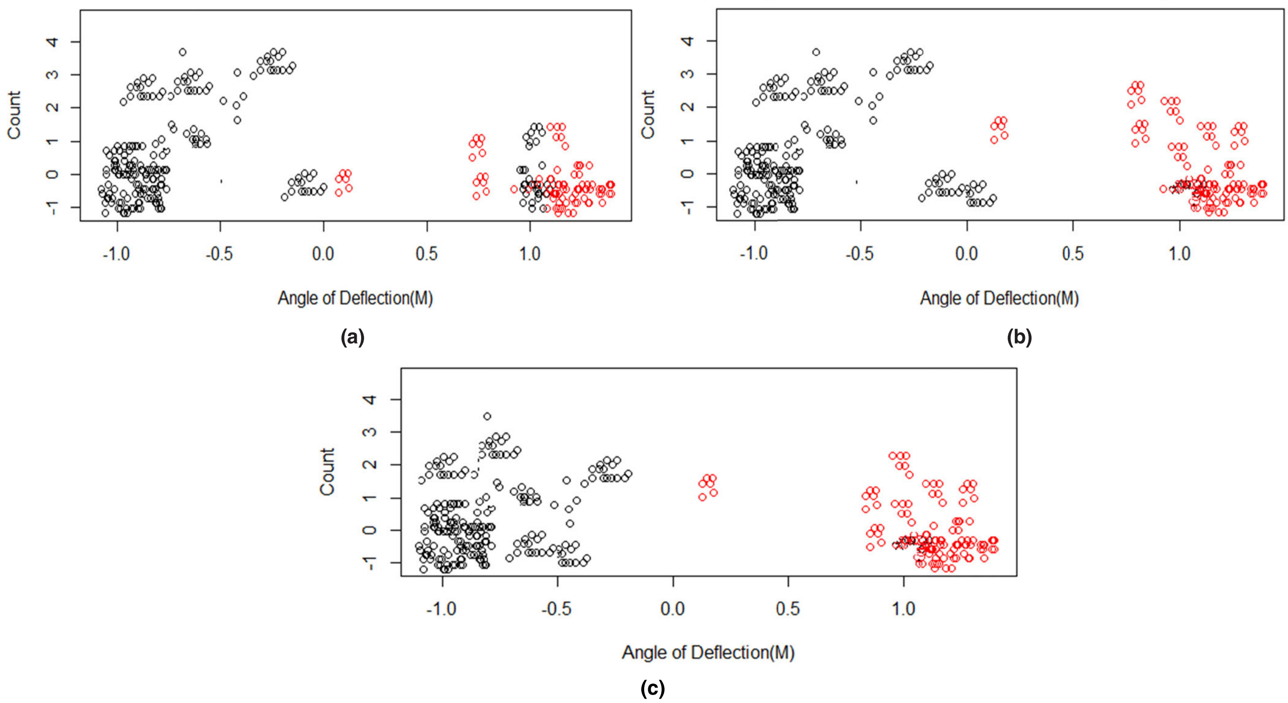


**FIGURE 7. Shows clustering results for the two methodsIT$_K$M (b) and EC$_G$ (c) used and compared on the Fish dataset (D3) with K-Means(a). The ensemble clustering method EC$_G$ shows better clustering results in terms of overlapping and inter-cluster distance.**

**TABLE 3. Sea Turtle, Earthworm and fish meta-data.**

| Parameter | Name | Rule extracted parameter variable | Range of Values Observed |
|---|---|---|---|
| p=1,2(In degree) | Angle of Deviation(Month) | Low | <= 100 ° |
| | Angle of Deviation(day) | Medium | >=100° & <=140° |
| | | High | >=140° & <160° |
| | | very high | >=160° |
| p =3 | Count | Low | <=3 |
| | | Medium | >=5 &<=7 |
| | | High | >=8 |
| p=4(In Tesla) | Electromagnetic Field | Low | <=1 nT |
| | | High | >= 1 nT till 4 nT |

outperforms all the baseline methods taken. The outperformance is empirically supported by the three cluster indices calculated: Silhouette Coefficient, Dunn and Rand Index. These indices are used to compare the performance of the clustering algorithms in terms of cluster quality, ability to accurately find the intrinsic groupings and agreement within the clusters [90].

In Table 4, the proposed method gives a high Silhouette Coefficient, Dunn and Rand Index value on all the three datasets as compared to other methods. A high score in silhouette signifies the ability of the method to effectively cluster the intrinsic relationships into crisp groups. Dunn index for EC$_G$ evaluates to be 0.55 (D1), 0.53 (D2) and 0.56 (D3) which is marginally higher as compared to baseline K-Means or Fuzzy C Means. A higher value of rand index which is an extrinsic validation index also provides evidence to the outperformance of the proposed method. There is also a variation across all three behavioral datasets, specifically for silhouette value, for all other methods there is a small difference between two other indices.

Hence, the unlabeled data values from all the three datasets can now be labeled to form the common clusters framed for each data point in sea turtle, earthworms, and fish behavioral dataset separately. The improvement in cluster identification of all three specie behavior, which is plotted between Count parameter of the particular specie and the Angle of Deviation in navigation shown monthly, is depicted in Figure 5, 6, and 7. The reason to choose these two parameters in comparison lies in the hypothesis discussed in [33] where a change in navigation in response to abnormal seismic activity brings in a change in specie population count abnormally.

### C. EC$_G$ ON BENCHMARK DATASETS: EVALUATION AND COMPARISON

This section describes the statistical cluster analysis performed for two baselines existing (K-Means and Hierarchical clustering), one intermediate (IT$_K$M) and the pro-

**TABLE 4.** Accuracy of the competing methods in terms of (a) Silhouette Coefficient, (b) Dunn-Index and (c) Rand Index There is no particular baseline which is the best across all datasets.

| Clustering Technique | Analysis Measure | Dataset 1(Sea Turtle) | Dataset 2 (Earthworms) | Dataset 3 (Fish) |
|---|---|---|---|---|
| K-Means | Silhouette Coefficient | 0.49 | 0.43 | 0.46 |
| | Dunn Index | 0.435 | 0.455478 | 0.475478 |
| | Rand Index | 0.32 | 0.335 | 0.316 |
| Hierarchial Clustering | Silhouette Coefficient | 0.50 | 0.5 | 0.49 |
| | Dunn Index | 0.206 | 0.312 | 0.335376 |
| | Rand Index | 0.33 | 0.34 | 0.35 |
| Fuzzy C Means | Silhouette Coefficient | 0.63 | 0.69 | 0.70 |
| | Dunn Index | 0.395376 | 0.435376 | 0.435376 |
| | Rand Index | 0.39 | 0.378 | 0.356 |
| $IC_KM$ | Silhouette Coefficient | 0.60 | 0.64 | 0.73 |
| | Dunn Index | 0.448378 | 0.48123 | 0.39878 |
| | Rand Index | 0.4992 | 0.4987 | 0.4887 |
| $IT_KM$ | Silhouette Coefficient | 0.69 | 0.71 | 0.72 |
| | Dunn Index | 0.348378 | 0.38123 | 0.31878 |
| | Rand Index | 0.5002 | 0.5034 | 0.5078 |
| $EC_G$ [Proposed Ensemble Clustering method] | Silhouette Coefficient | 0.87 | 0.80 | 0.77 |
| | Dunn Index | 0.55245 | 0.53678 | 0.56345 |
| | Rand Index | 0.51 | 0.492 | 0.5102 |

**TABLE 5.** Accuracy of the competing methods along with proposed in terms of (a) Silhouette Coefficient, (b) Dunn-Index and (c) Rand Index.

| Dataset | Analysis Measure | KMeans | Hierarchical Clustering | $IT_KM$ | $EC_G$ |
|---|---|---|---|---|---|
| Iris | Silhouette Coefficient | 0.49 | 0.52 | 0.68 | 0.98 |
| | Dunn Index | 0.0264 | 0.0315 | 0.0456 | 0.0513 |
| | Rand Index | 0.620 | 0.609 | 0.624 | 0.710 |
| Breast_cancer | Silhouette Coefficient | 0.45 | 0.27 | 0.56 | 0.88 |
| | Dunn Index | 0.85 | 0.71 | 0.874 | 0.889 |
| | Rand Index | 0.0248 | 0.0498 | 0.0678 | 0.0789 |
| Heart Disease | Silhouette Coefficient | 0.39 | 0.36 | 0.41 | 0.79 |
| | Dunn Index | 0.7061 | 0.6368 | 0.7376 | 0.8176 |
| | Rand Index | 0.69 | 0.532 | 0.756 | 0.878 |
| USarrests | Silhouette Coefficient | 0.63 | 0.39 | 0.71 | 0.87 |
| | Dunn Index | 0.558378 | 0.48123 | 0.67878 | 0.78123 |
| | Rand Index | 0.5992 | 0.5187 | 0.7881 | 0.896 |

posed method ($EC_G$) for UCI benchmark datasets [91]. Table 5 presents the results of the three cluster indices over four benchmark datasets for the mentioned methods. As shown in Table 4, intermediate method $IT_KM$ has outperformed $IC_KM$ in all clustering comparison metrics for all behavioral data sets considered for the given problem; hence the benchmark data sets are only evaluated on $IT_KM$ and $EC_G$ where the latter shows successful values for all five datasets. Say $EC_G$gives Silhouette Coefficient of 0.98 for iris, 0.88 for the breast_cancer dataset while $IT_KM$ gives 0.68 and 0.56 for the same respectively with K-Means as 0.49 and 0.45.

### D. PERFORMANCE EVALUATION TO SELECT BEST-SEQUENCED METHOD

For sequenced approach, we have permuted all feasible sequences of baseline standards 5 (five) FRBCS methods with 2 (two) standard clustering methods, two used and one proposed method ($IC_KM$, $IT_KM$, and $EC_G$) and select the best ensemble as $EC_GF_C$. While $IT_KM$ 's efficiency has outperformed $IC_KM$ in terms of cluster statistics (Table 4), it is still permissible to classify any improvements in the same after a fuzzy classification of the ensemble.

Table 6 presents the analysis to select the best sequence approach of clustering and classification ($EC_GF_C$). Of the

**TABLE 6.** Prediction accuracy of the fizzy rule-based classification for baseline FRBCS systems after ensemble clustering and classification ($EC_GF_C$).

| Ensemble Number | Ensemble Clustering and Classification Setting ↓ / Dataset → | Dataset 1 | | | Dataset 2 | | | Dataset 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | RMSE | MAE | SMAPE | RMSE | MAE | SMAPE | RMSE | MAE | SMAPE |
| 1 | KMeans+FRBCS.CHI | 0.18712 | 0.58821 | 19.67 | 0.381241 | 0.60213 | 17.78 | 0.377544 | 0.55724 | 18.93 |
| 2 | Hierarchical KMeans +FRBCS.CHI | 0.17251 | 0.57616 | 18.23 | 0.339912 | 0.59211 | 16.08 | 0.367456 | 0.52897 | 1 7.21 |
| 3 | IC$_K$M+FRBCS.CHI | 0.16652 | 0.54121 | 17.61 | 0.321897 | 0.58245 | 15.72 | 0.357589 | 0.50345 | 1 6.56 |
| 4 | IT$_K$M+FRBCS.CHI | 0.16001 | 0.53921 | 17.01 | 0.311241 | 0.57121 | 15.01 | 0.347521 | 0.49098 | 1 6.04 |
| 5 | EC$_G$+FRBCS.CHI | 0.15617 | 0.52143 | 16.62 | 0.301242 | 0.56211 | 14.98 | 0.307511 | 0.48729 | 1 5.90 |
| 6 | kMeans + GFS.GCCL | 0.12961 | 0.24614 | 11.21 | 0.336278 | 0.39821 | 13.86 | 0.299421 | 0.41828 | 1 2.34 |
| 7 | Hierarchical KMeans +GFS.GCCL | 0.14606 | 0.49721 | 10.977 | 0.350213 | 0.39987 | 14.18 | 0.226523 | 0.49924 | 1 3.83 |
| 8 | IC$_K$M +GFS.GCCL | 0.12112 | 0.44821 | 8.971 | 0.320149 | 0.40213 | 12.45 | 0.216523 | 0.46621 | 1 2.60 |
| 9 | IT$_K$M +GFS.GCCL | 0.10998 | 0.38821 | 8.11 | 0.301201 | 0.28211 | 11.98 | 0.209123 | 0.33726 | 1 1.08 |
| 10 | EC$_G$ +GFS.GCCL | 0.10112 | 0.30145 | 7.68 | 0.290117 | 0.22203 | 10.12 | 0.206521 | 0.30171 | 10 .93 |
| 11 | KMeans + FH.GBML | 0.12651 | 0.22815 | 10.46 | 0.323448 | 0.39203 | 12.71 | 0.217121 | 0.40721 | 1 0.31 |
| 12 | Hierarchical KMeans + FH.GBML | 0.13761 | 0.44725 | 9.676 | 0.331043 | 0.37211 | 13.73 | 0.206521 | 0.49624 | 1 1.92 |
| 13 | IC$_K$M + FH.GBML | 0.11656 | 0.40021 | 8.02 | 0.301208 | 0.39113 | 11.28 | 0.200123 | 0.41724 | 1 1.26 |
| 14 | IT$_K$M + FH.GBML | 0.10256 | 0.38825 | 7.97 | 0.281238 | 0.27113 | 10.59 | 0.196523 | 0.30621 | 1 0.28 |
| **15** | **EC$_G$+ FH.GBML** | **0.09482** | **0.30102** | **6.61** | **0.261218** | **0.20011** | **9.61** | **0.187671** | **0.29721** | **9.83** |
| 16 | KMeans + SLAVE | 0.17757 | 0.54825 | 16.67 | 0.351248 | 0.59213 | 17.78 | 0.317523 | 0.51724 | 1 6.93 |
| 17 | Hierarchical KMeans + SLAVE | 0.16756 | 0.54825 | 16.21 | 0.351121 | 0.59011 | 16.11 | 0.307121 | 0.50224 | 1 5.13 |
| 18 | IC$_K$M + SLAVE | 0.14341 | 0.52834 | 15.67 | 0.340008 | 0.58101 | 15.68 | 0.296511 | 0.51721 | 1 5.23 |
| 19 | IT$_K$M + SLAVE | 0.13417 | 0.51831 | 14.67 | 0.342142 | 0.52013 | 14.31 | 0.286613 | 0.49126 | 1 4.11 |
| 20 | EC$_G$+ SLAVE | 0.12217 | 0.50223 | 13.67 | 0.331248 | 0.50218 | 13.58 | 0.277523 | 0.48727 | 1 3.98 |
| 21 | KMeans + FRBCS.W | 0.14321 | 0.25616 | 14.21 | 0.320011 | 0.40167 | 15.61 | 0.320123 | 0.49734 | 1 4.93 |
| 22 | Hierarchical KMeans + FRBCS.W | 0.15211 | 0.45213 | 13.32 | 0.331456 | 0.40678 | 14.62 | 0.330011 | 0.48238 | 1 4.21 |
| 23 | IC$_K$M + FRBCS.W | 0.12101 | 0.42121 | 11.66 | 0.311289 | 0.39356 | 13.34 | 0.300513 | 0.46121 | 1 3.91 |
| 24 | IT$_K$M + FRBCS.W | 0.11721 | 0.40726 | 10.57 | 0.309311 | 0.37678 | 13.45 | 0.286521 | 0.45624 | 1 2.96 |
| 25 | EC$_G$+ FRBCS.W | 0.10757 | 0.39825 | 10.01 | 0.301234 | 0.36890 | 12.89 | 0.257512 | 0.39074 | 1 1.23 |

above mentioned permutations, 25 sequenced combinations are obtained for exhaustive analysis. The permuted ensembles (25) were compared and evaluated for three parameters: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and SMAPE (Symmetric Mean Absolute Percentage Error), as shown in equations 4, 5 and 6.SMAPE is defined as an accuracy measure which calculates a relative difference between actual ($A_t$) and forecasted ($F_t$) values [92].

$$\text{MAE (Mean absolute error)} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y_i}) \tag{4}$$

$$\text{RMSE (Root Mean Square Error)} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y_i}) \wedge 2} \tag{5}$$

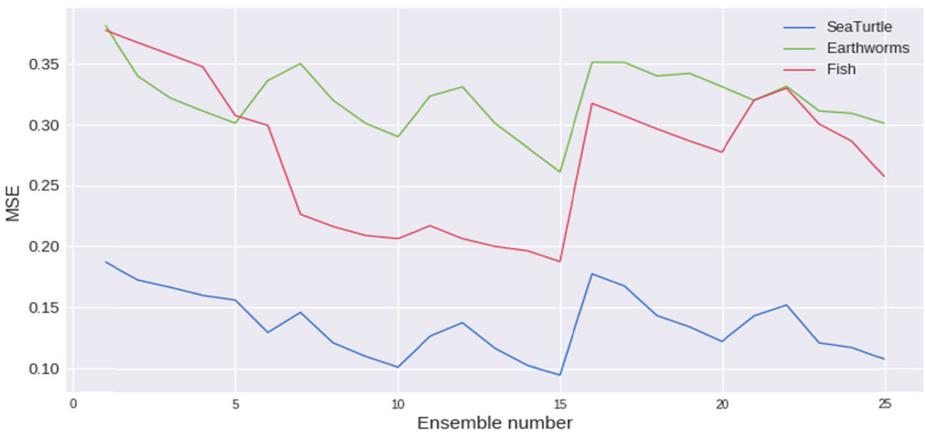


**FIGURE 8.** RMSE plotted for all three datasets: SeaTurtle, Earthworms, and Fish.

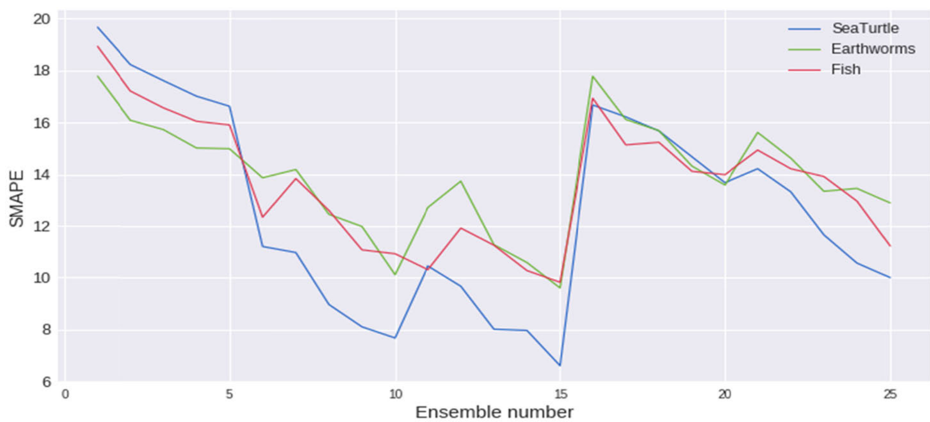


**FIGURE 9.** SMAPE plotted for all three datasets: SeaTurtle, Earthworms, and Fish.

SMAPE (Symmetric Mean Absolute Percentage Error)

$$= \frac{100\%}{n} \sum_{t=1}^{t=n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \quad (6)$$

For all three datasets which give behavior of sea turtle, earthworm, and fish, the devised $EC_GF_C$ having ITKM and GFS.GCCL have following values for RMSE,MAE and SMAPE.14756,.54345 and 13.67 for turtle behavior dataset (D1) and.251248,,.52213 and 12.78 for Earthworm behavioral dataset(D2). Similar lower error values are observed for the fish behavioral dataset (D3). The defined methods or the baseline methods such as K-Means, Hierarchal K-Means have also been sequenced with all standard FRBCS, as presented in Table 5 Evidently, FH.GBML and GFS.GCCL have low error values when bagged with baseline clustering methods. RMSE provides Out of all 25 different settings, two have performed best having the proposed ensemble clustering methods ($IC_KM$, $IT_KM$, and $EC_G$).

Figure 8 and Figure 9 show RMSE and SMAPE plots for all three behavioral datasets: Sea Turtle, Earthworms, and Fish for all 25 ensembles numbered here from 1 to 25.Sea turtle displays, as shown, the least RMSE values for all permuted ensembles while fish shows the most. The mentioned metrics: RMSE & MAE provide generalized probable accuracy criteria for unseen and unlabeled data [49]. Sea Turtles may indeed serve mostly as fairly powerful predictor in detecting any seismic activity underwater, as opposed to fish or earthworms. Nonetheless, it can be inferred from the study of high SMAPE values that more efficient methods can be planned and implemented for higher precision performance.

### *E. SEQUENCED METHODS WITH EXISTING ENSEMBLES*

From Table 6, the best-performing combination of clustering and classification ($EC_GF_C$) is $EC_G$ sequenced with GFS.GCCL. For all three behavioral datasets, high reliable results make the above-mentioned combination one of the most powerful ensembles to predict anomaly patterns in aquatic animal activity prior to seismic disturbance. The FRBCS used here creates a knowledge base of specific fuzzy rules based on labeled data to identify any similar abnormality in the future. Four benchmark datasets are analyzed to further validate the selected ensemble performance in order to compare obtained sequenced $EC_GF_C$ approach with current state-of-the-art ensembles. The results are shown in Table 7. Because the benchmark datasets are already

**TABLE 7.** Performance comparison of proposed $EC_GF_C$ method with existing ensembles on the Benchmark datasets.

| Dataset | No. of Instances | Agrawal et al. [93] | F. Farahbod et al. [95] | Proposed Approach | [103] | [37] |
|---|---|---|---|---|---|---|
| **Iris** | **150** | 96 | 96.6 | **96.6** | 94 | X |
| **Wisconsin Breast_Cancer** | **683** | 96.2 | **98.7** | 98.2 | 96.20 | 90.78 |
| **Wine Dataset** | **178** | 90 | **97.5** | 96.4 | 94.94 | X |
| **Car Evaluation** | **1787** | X | 83.13 | **85.67** | X | X |
| **RingNorm** | **7400** | X | X | **91.48** | 89.46 | 55.27 |
| **Magic** | **19020** | X | X | **81.40** | 79.81 | 77.08 |

**TABLE 8.** Comparison with state-of-the-art models for behavioral classification.

| Study/ Experiment | Animal type | Method | Classification metric | Result | Application |
|---|---|---|---|---|---|
| [72] | Cattle | Ensemble Classification | Accuracy | 96% | Quantitative health assessment |
| [95] | Cattle | Variable segmentation & Ensemble Classification | Accuracy | 96% | Livestock Monitoring |
| [96] | Dolphin | Ensemble modeling | Accuracy (season dependent) | 41-82% | Identify dolphin conservation areas |
| [39] | Lemon Sharks | Voting Ensemble | 0.888 Macro-Averaged-F Measure | 0.888 | Predicting prey capture behavior |
| [92] | Brown Hare | Random Forest Ensemble | Accuracy | 89% | Classify behavior |
| This paper | Sea Turtles, Fish & Earthworms | GMM ensemble Clustering with sequenced FRBCS | MSE, MAE & SMAPE | Table 5 | Classify behavior for tsunami alert identification |

clustered, the existing labels are removed here and the results will be evaluated after $EC_GF_C$ application. Different sets and combined frameworks have been proposed in state-of-the-art terms; however, only clustering classification approaches are selected for the comparison presented. The existing frameworks [93], [94] have already been used for breast cancer profile identification and some other benchmark datasets respectively. The comparison here is identified in terms of the percentage of data instances correctly classified after applying $EC_GF_C$. The sequenced method $EC_GF_C$ give higher accuracy results for the considered benchmark datasets: Iris, Wisconsin Breast cancer, Heart Disease and Car Evaluation [91].

In order to understand the performance statistics in reference to the scalability, the benchmark datasets were selected having different scales. Here both, a small iris dataset (150 data points) were considered for performance comparison and evaluation along with one mid-scale dataset (instances $< 2000$) and two large scale datasets (instances $> 2000$) taken from UCI repository [91]. The explanation for evaluating small-scale data sets was based on the fact that the behavioral datasets collected and then prepared were low-scale, hence the empirical evaluation and validation of the proposed method was carried out initially on small-scale benchmark datasets. However to comment about the scalability yet efficiency and analyze the trade-off between the two, one of the recent big fuzzy data algorithm (Chi-BD [104]) has been used for the evaluation from [37]. The comparison subsequently reaffirms the need to pipeline an $EC_G$ pre-processing clustering method prior to the FRBCS (Fuzzy Rule Based Classification Systems) baseline where $EC_G$provided optimized quick labeling and further assisted in improving classification. Even on large dataset, the proposed sequenced approach outperformed the current ensembles. Compared to 77.08 reported by authors in [37] using one of the big data fuzzy classification algorithm, the percentage of data instances correctly classified after application of $EC_GF_C$ accounts to be 89.46.

Table 8 summarizes the most relevant related articles that have employed Ensemble clustering or classification on animal behavioral datasets. A fair comparison is not possible, as this work is a new attempt to integrate behavioral response analytics on a prepared MSD for seismic alert classification. However, even though the behavioral study of both marine and terrestrial animals has been in literature for many other applications. In the presented comparison, ensemble classification has provided insights about cattle monitoring [72] and conservation area identification in dolphins [97] cattle.

The state-of-the-art, as shown, differs from the current analysis in terms of applications, including the animal type case taken. It can evidently be summarized that an integrated scalable approach focusing on finding marine behavioral patterns which can assist seismic alerts in the form of a global system is an unexplored application area from the exhaustive analysis presented at the moment.

## V. CONCLUSION

In this paper, three behavioral datasets have been used to identify any pattern that can help in retrieving real-time alerts on seismic disturbances. The identified species for the same are Sea turtles, Earthworms, and Fish. The data here is prepared by combining both biological and geophysical sources. From the prepared dataset, a classification identifying alert or no-alert situations can be performed based on ambiguous behavioral signals exhibited by marine species in response to seismic perturbations. This article focuses on the labeling of prepared data by using a proposed ensemble clustering method $EC_G$ which is evaluated on four benchmark datasets and three behavioral datasets in terms of clusters statistics viz. Silhouette Coefficient, Dunn, and Rand Index. For classifying alert and no-alert data points, the sequencing of the proposed $EC_G$ method is done with state-of-the-art fuzzy rule-based classification methods. A total of 25 sequenced combinations are evaluated on all three behavioral datasets for MSE, MAE, and SMAPE. One of the advantages of the proposed sequenced approach its generic adaptability towards behavioral datasets. The results coming from cluster indices empirically support the ability of ensemble approaches for similar application areas.

Sea turtle dataset hence prepared, showed the best MSE, MAE, and SMAPE values. The sequence method having proposed $EC_G$ and GFS.GCCL showed the least error statistics across all three datasets. Therefore, using such optimized sequenced methods, behavioral datasets can be easily classified. Such improved classification can form a basis to program various global epidemic systems to raise alerts from animal behavior based on mined patterns learned.

The goal of this research is to progressively move towards a global TWS using marine behavioral statistical analysis. Finding more categories of species (both terrestrial and aquatic) that can act as biosensors for seismic alert classification is one of the areas this research opens to. More data sources need to be evolved that can help in devising more insights using the machine and deep learning methods. In the future, we plan to use other fuzzy classification methods recently proposed for both big and small-scale data, where further validation will be sought. Furthermore, how the proposed solution allows for a reduced base of rules, and how time reduction shapes another field of this research.

## REFERENCES


[1] P.-C. Athukorala and B. P. Resosudarmo, "The Indian Ocean tsunami: Economic impact, disaster management, and lessons," *Asian Econ. Papers*, vol. 4, no. 1, pp. 1–39, Jan. 2005.

[2] R. S. C. Arce, M. Onuki, M. Esteban, and T. Shibayama, "Risk awareness and intended tsunami evacuation behaviour of international tourists in Kamakura City, japan," *Int. J. Disaster Risk Reduction*, vol. 23, pp. 178–192, Aug. 2017.

[3] B. Pérez, E. Álvarez Fanjul, S. Pérez, M. de Alfonso, and J. Vela, "Use of tide gauge data in operational oceanography and sea level hazard warning systems," *J. Oper. Oceanogr.*, vol. 6, no. 2, pp. 1–18, Aug. 2013.

[4] A. Romano, G. Bellotti, and M. Di Risio, "Wavenumber-frequency analysis of the landslide-generated tsunamis at a conical island," *Coastal Eng.*, vol. 81, pp. 32–43, Nov. 2013.

[5] F. Ai, L. K. Comfort, Y. Dong, and T. Znati, "A dynamic decision support system based on geographical information and mobile social networks: A model for tsunami risk mitigation in Padang, Indonesia," *Saf. Sci.*, vol. 90, pp. 62–74, Dec. 2016.

[6] T. Saito, Y. Ito, D. Inazu, and R. Hino, "Tsunami source of the 2011 Tohoku-Oki earthquake, Japan: Inversion analysis based on dispersive tsunami simulations," *Geophys. Res. Lett.*, vol. 38, no. 7, pp. 1–5, Apr. 2011.

[7] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "#Earthquake: Twitter as a distributed sensor system," *Trans. GIS*, vol. 17, no. 1, pp. 124–147, Feb. 2013.

[8] K. Seike, T. Kitahashi, and T. Noguchi, "Sedimentary features of Onagawa Bay, northeastern Japan after the 2011 off the Pacific coast of Tohoku Earthquake: Sediment mixing by recolonized benthic animals decreases the preservation potential of tsunami deposits," *J. Oceanogr.*, vol. 72, no. 1, pp. 141–149, Feb. 2016.

[9] G. Bankoff, "Learning about disasters," in *Learning and Calamities, Practices, Interpretations, Patterns*, H. Egner, M. Schorch, and M. Voss, Eds. New York, NY, USA: Routledge, 2015, pp. 42–55.

[10] I. Farías, "Misrecognizing tsunamis: Ontological politics and cosmopolitical challenges in early warning systems," *Sociol. Rev.*, vol. 62, no. 1, pp. 61–87, Jun. 2014.

[11] E. Marris, "Inadequate warning system left Asia at the mercy of tsunami," *Nature*, vol. 433, no. 7021, pp. 3–6, 2005.

[12] P. M. Mai, "Supershear tsunami disaster," *Nature Geosci.*, vol. 12, no. 3, pp. 150–151, Mar. 2019.

[13] B. R. Röbke and A. Vött, "The tsunami phenomenon," *Prog. Oceanogr.*, vol. 159, pp. 296–322, Dec. 2017.

[14] Y. Bar-Cohen, "Biological senses as inspiring model for biomimetic sensors," *IEEE Sensors J.*, vol. 11, no. 12, pp. 3194–3201, Dec. 2011.

[15] T. Waltham, "The Asian tsunami disaster, December 2004," *Geol. Today*, vol. 21, no. 1, pp. 22–26, 2005.

[16] R. D. Sagarin and T. Taylor, Natural security: How biological systems use information to adapt in an unpredictable world," *Secur. Inform.*, vol. 1, no. 1, pp. 1–9, 2012.

[17] H. Woith, G. M. Petersen, S. Hainzl, and T. Dahm, "Can animals predict earthquakes?" *Bull. Seismol. Soc. Amer.*, vol. 108, no. 3A, pp. 1031–1045, 2018.

[18] W. K. Dodds, C. T. Robinson, E. E. Gaiser, G. J. A. Hansen, H. Powell, J. M. Smith, N. B. Morse, S. L. Johnson, S. V. Gregory, T. Bell, T. K. Kratz, and W. H. McDowell, "Surprises and insights from long-term aquatic data sets and experiments," *BioScience*, vol. 62, no. 8, pp. 709–721, Aug. 2012.

[19] J. L. Kirschvink, "Earthquake prediction by animals: Evolution and sensory perception," *Bull. Seismol. Soc. Amer.*, vol. 90, no. 2, pp. 312–323, Apr. 2000.

[20] T. Rikitake, "Biosystem behaviour as an earthquake precursor," *Tectonophysics*, vol. 51, nos. 1–2, pp. 1–20, Nov. 1978.

[21] T. Richa and T. Sudhanshu, "Animals:Anatural messenger for disasters," *J. Nature Prod.*, vol. 4, pp. 3–4, 2011.

[22] H. Pirim and S. E. Seker, "Ensemble clustering for biological datasets," in *Bioinformatics*, H. Pérez-Sánchez, Eds. London, U.K.: IntechOpen, 2012, pp. 2320–2333.

[23] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson, "JAABA: Interactive machine learning for automatic annotation of animal behavior," *Nature Methods*, vol. 10, no. 1, pp. 64–67, Jan. 2013.

[24] J. J. Valletta, C. Torney, M. Kings, A. Thornton, and J. Madden, "Applications of machine learning in animal behaviour studies," *Animal Behaviour*, vol. 124, pp. 203–220, Feb. 2017.

[25] R. M. Alguliyev, R. M. Aliguliyev, and L. V. Sukhostat, "Efficient algorithm for big data clustering on single machine," *CAAI Trans. Intell. Technol.*, vol. 5, no. 1, pp. 9–14, Mar. 2020.

[26] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[27] Y. Shi and J. Nan, "Improved FCM algorithm based on initial center optimization method," *J. Intell. Fuzzy Syst.*, vol. 32, no. 5, pp. 3487–3494, Apr. 2017.

[28] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining partitionings," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

[29] J. Garriga, J. R. B. Palmer, A. Oltra, and F. Bartumeus, "Expectation-maximization binary clustering for behavioural annotation," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, Art. no. e0151984.

[30] A. Farrell, G. Wang, S. A. Rush, J. A. Martin, J. L. Belant, A. B. Butler, and D. Godwin, "Machine learning of large–scale spatial distributions of wild turkeys with high–dimensional environmental data," *Ecol. Evol.*, vol. 9, no. 10, pp. 5938–5949, May 2019.

[31] N. V. Chawla and G. Karakoulas, "Learning from labeled and unlabeled data: An empirical study across techniques and domains," *J. Artif. Intell. Res.*, vol. 23, pp. 331–366, Mar. 2005.

[32] C. Luan and G. Dong, "Experimental identification of hard data sets for classification and feature selection methods with insights on method selection," *Data Knowl. Eng.*, vol. 118, pp. 41–51, Nov. 2018.

[33] D. Virmani, N. Jain, A. Srivastav, M. Mittal, and S. Mittal, "An enhanced binary classifier incorporating weighted scores," *Eng., Technol. Appl. Sci. Res.*, vol. 8, no. 2, pp. 2853–2858, 2018.

[34] G. Brock, V. Pihur, S. Datta, and S. Datta, "ClValid: AnRPackage for cluster validation," *J. Stat. Softw.*, vol. 25, no. 4, 2011.

[35] V. Mehta, S. Bawa, and J. Singh, "Analytical review of clustering techniques and proximity measures," *Artif. Intell. Rev.*, May 2020.

[36] A. Kassambara, *Practical Guide to Cluster Analysis in R: Unsupervised machine learning*, vol. 1. Warsaw, Poland: STHDA, 2017.

[37] M. Elkano, H. Bustince, and M. Galar, "Do we still need fuzzy classifiers for small data in the era of big data?" in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2019, pp. 1–6.

[38] D. Virmani and N. Jain, "Ensemble multi-stage system and method to predict underwater seismic rate on aquatic behaviour," Indian Patent Office, New Delhi, India, India Appl. 202011009855, Mar. 2020.

[39] L. R. Brewster, J. J. Dale, T. L. Guttridge, S. H. Gruber, A. C. Hansell, M. Elliott, I. G. Cowx, N. M. Whitney, and A. C. Gleiss, "Development and application of a machine learning algorithm for classification of elasmobranch behaviour from accelerometry data," *Mar. Biol.*, vol. 165, no. 4, p. 62, Apr. 2018.

[40] J. Urabe, T. Suzuki, T. Nishita, and W. Makino, "Immediate ecological impacts of the 2011 Tohoku earthquake tsunami on intertidal flat communities," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e62779.

[41] J. Shoji and M. Morimoto, "Changes in fish community in seagrass beds in Mangoku-Ura bay from 2009 to 2014, the period before and after the tsunami following the 2011 off the pacific coast of Tohoku earthquake," *J. Oceanogr.*, vol. 72, no. 1, pp. 91–98, Feb. 2016.

[42] R. A. Grant and T. Halliday, "Predicting the unpredictable; evidence of pre-seismic anticipatory behaviour in the common toad," *J. Zool.*, vol. 281, no. 4, pp. 263–271, 2010.

[43] M. Garstang and M. Kelley, "Understanding animal detection of precursor earthquake sounds," *Animals*, vol. 7, no. 12, p. 66, Aug. 2017.

[44] B. W. Levin and M. A. Nosov, "General information on tsunami waves, seaquakes, and other catastrophic phenomena in the ocean," in *Physics of Tsunamis*, Cham, Switzerland: Springer, 2016, pp. 1–34.

[45] T. Alqurashi and W. Wang, "Clustering ensemble method," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 6, pp. 1227–1246, 2019.

[46] X. Z. Fern and W. Lin, "Cluster ensemble selection," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 1, no. 3, pp. 128–141, Nov. 2008.

[47] Y. Hong, S. Kwong, H. Wang, and Q. Ren, "Resampling-based selective clustering ensembles," *Pattern Recognit. Lett.*, vol. 30, no. 3, pp. 298–305, Feb. 2009.

[48] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in *Proc. 21st Int. Joint Conf. Artif. Intell. (IJCAI)*, 2009, pp. 992–997.

[49] X. Wang, C. Yang, and J. Zhou, "Clustering aggregation by probability accumulation," *Pattern Recognit.*, vol. 42, no. 5, pp. 668–675, May 2009.

[50] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Inf. Fusion*, vol. 6, no. 1, pp. 83–98, Mar. 2005.

[51] A. H. R. Ko, R. Sabourin, and A. S. Britto, Jr., "From dynamic classifier selection to dynamic ensemble selection," *Pattern Recognit.*, vol. 41, no. 5, pp. 1718–1731, May 2008.

[52] Y. Yang and H. Wang, "Multi-view clustering: A survey," *Bid Data Mining Anal.*, vol. 1, no. 2, pp. 83–107, Jun. 2018.

[53] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proc. 29th Int. Conf. Very Large Database*, vol. 29, 2003, pp. 81–92.

[54] J. Gama, P. P. Rodrigues, and L. Lopes, "Clustering distributed sensor data streams using local processing and reduced communication," *Intell. Data Anal.*, vol. 15, no. 1, pp. 3–28, Jan. 2011.

[55] M. R. Ackermann, M. Märtens, C. Raupach, K. Swierkot, C. Lammersen, and C. Sohler, "StreamKM++: A clustering algorithm for data streams," *J. Experim. Algorithmics*, vol. 17, pp. 2–4, May 2012.

[56] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The ClusTree: Indexing micro-clusters for anytime stream mining," *Knowl. Inf. Syst.*, vol. 29, no. 2, pp. 249–272, Nov. 2011.

[57] J. Diaz-Rozo, C. Bielza, and P. Larrañaga, "Clustering of data streams with dynamic Gaussian mixture models: An IoT application in industrial processes," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3533–3547, Oct. 2018.

[58] Z. Ji, Y. Xia, Q. Sun, Q. Chen, and D. Feng, "Adaptive scale fuzzy local Gaussian mixture model for brain MR image segmentation," *Neurocomputing*, vol. 134, pp. 60–69, Jun. 2014.

[59] O. R. Bidder, O. Arandjelović, F. Almutairi, E. L. C. Shepard, S. A. Lambertucci, L. A. Qasem, and R. P. Wilson, "A risky business or a safe BET? A fuzzy set event tree for estimating hazard in biotelemetry studies," *Animal Behaviour*, vol. 93, pp. 143–150, Jul. 2014.

[60] V. Rozin and M. Margaliot, "The fuzzy ant," *IEEE Comput. Intell. Mag.*, vol. 2, no. 4, pp. 18–28, Nov. 2007.

[61] E. Tron and M. Margaliot, "Mathematical modeling of observed natural behavior: A fuzzy logic approach," *Fuzzy Sets Syst.*, vol. 146, no. 3, pp. 437–450, Sep. 2004.

[62] D. Gu and H. Hu, "Using fuzzy logic to design separation function in flocking algorithms," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 4, pp. 826–838, Aug. 2008.

[63] H. Ishibuchi, T. Nakashima, and T. Murata, "Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 29, no. 5, pp. 601–618, May 1999.

[64] H. Ishibuchi, T. Yamamoto, and T. Nakashima, "Hybridization of fuzzy GBML approaches for pattern classification problems," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 35, no. 2, pp. 359–365, Apr. 2005.

[65] L. S. Riza, C. N. Bergmeir, F. Herrera, and J. M. B. Sánchez, *Frbs: Fuzzy Rule-Based Systems for Classification and Regression in R*. Alexandria, VA, USA: American Statistical Association, 2015.

[66] R. Avogadri and G. Valentini, "Fuzzy ensemble clustering based on random projections for DNA microarray data analysis," *Artif. Intell. Med.*, vol. 45, nos. 2–3, pp. 173–183, Feb. 2009.

[67] Z. Yu, H. Chen, J. You, G. Han, and L. Li, "Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 657–670, May 2013.

[68] D. Smith, B. Little, P. I. Greenwood, P. Valencia, A. Rahman, A. Ingham, G. Bishop-Hurley, S. Shahriar, and A. Hellicar, "A study of sensor derived features in cattle behaviour classification models," in *Proc. IEEE Sensors*, Nov. 2015, pp. 1–4.

[69] N. Mansbridge, J. Mitsch, N. Bollard, K. Ellis, G. Miguel-Pacheco, T. Dottorini, and J. Kaler, "Feature selection and comparison of machine learning algorithms in classification of grazing and rumination behaviour in sheep," *Sensors*, vol. 18, no. 10, p. 3532, Oct. 2018.

[70] E. N. de Souza, K. Boerder, S. Matwin, and B. Worm, "Improving fishing pattern detection from satellite AIS using data mining and machine learning," *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0158248.

[71] O. R. Bidder, H. A. Campbell, A. Gómez-Laich, P. Urgé, J. Walker, Y. Cai, L. Gao, F. Quintana, and R. P. Wilson, "Love thy neighbour: Automatic animal behavioural classification of acceleration data using the K-Nearest neighbour algorithm," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e88609.

[72] R. Dutta, D. Smith, R. Rawnsley, G. Bishop-Hurley, J. Hills, G. Timms, and D. Henry, "Dynamic cattle behavioural classification using supervised ensemble classifiers," *Comput. Electron. Agricult.*, vol. 111, pp. 18–28, Feb. 2015.

[73] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, no. 2, pp. 330–348, Feb. 2017.

[74] H. Xiao, Z. Xiao, and Y. Wang, "Ensemble classification based on supervised clustering for credit scoring," *Appl. Soft Comput.*, vol. 43, pp. 73–86, Jun. 2016.

[75] R. Nathan, O. Spiegel, S. Fortmann-Roe, R. Harel, M. Wikelski, and W. M. Getz, "Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: General concepts and tools illustrated for griffon vultures," *J. Experim. Biol.*, vol. 215, no. 6, pp. 986–996, Mar. 2012.

[76] G. Wang, "Machine learning for inferring animal behavior from location and movement data," *Ecol. Informat.*, vol. 49, pp. 69–76, Jan. 2019.

[77] M. Hassan, A. Chaudhry, A. Khan, and J. Y. Kim, "Carotid artery image segmentation using modified spatial fuzzy C-means and ensemble clustering," *Comput. Methods Programs Biomed.*, vol. 108, no. 3, pp. 1261–1276, Dec. 2012.

[78] N. Jain and D. Virmani, "Ensemble learning using fast rule based fuzzy $k$-means pre clustering and classification for aquatic behavior-extracted tsunami prediction," *J. Inf. Optim. Sci.*, vol. 40, no. 2, pp. 441–453, Feb. 2019.

[79] D. Virmani, N. Jain, K. Parikh, S. Upadhyaya, and A. Srivastav, "Proficient normalised fuzzy K-means with initial centroids methodology," *Int. J. Knowl. Discovery Bioinf.*, vol. 8, no. 1, pp. 42–59, Jan. 2018.

[80] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. (2019). *Cluster: Cluster Analysis Basics and Extensions. R Package Version 2.1.0*. [Online]. Available: https://cran.r-project.org/web/packages/cluster/citation.html

[81] P. Korondi, B. Korcsok, S. Kovács, and M. Niitsuma, "Etho-robotics: What kind of behaviour can we learn from the animals?" *IFAC-PapersOnLine*, vol. 48, no. 19, pp. 244–255, 2015.

[82] H. Ishibuchi, K. Nozaki, and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," *Fuzzy Sets Syst.*, vol. 52, no. 1, pp. 21–32, Nov. 1992.

[83] D. Devi, S. Namasudra, and S. Kadry, "A boosting-aided adaptive cluster-based undersampling approach for treatment of class imbalance problem," *Int. J. Data Warehousing Mining*, vol. 16, no. 3, pp. 60–86, Jul. 2020, doi: 10.4018/IJDWM.2020070104.

[84] D. Virmani and N. Jain, "Intelligent information retrieval for tsunami detection using wireless sensor nodes," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 1103–1109.

[85] N. Jain and D. Virmani, "Feature classification for underwater seismic prediction using wireless sensor nodes," in *Proc. Int. MultiConf. Eng. Comput. Sci.*, vol. 1, 2017, pp. 151–156.

[86] N. Jain, D. Virmani, and A. Abraham, "Overlap function based fuzzified aquatic behaviour information extracted tsunami prediction model," *Int. J. Distrib. Syst. Technol.*, vol. 10, no. 1, pp. 56–81, Jan. 2019.

[87] P. Halpin, A. Read, E. Fujioka, B. Best, B. Donnelly, L. Hazen, C. Kot, K. Urian, E. LaBrecque, A. Dimatteo, J. Cleary, C. Good, L. Crowder, and K. D. Hyrenbach, "OBIS-SEAMAP: The world data center for marine mammal, sea bird, and sea turtle distributions," *Oceanography*, vol. 22, no. 2, pp. 104–115, Jun. 2009.

[88] N. Jain and D. Virmani, "A novel navigation based decision algorithm to generate tsunami alerts using sensor nodes," *Compusoft*, vol. 7, no. 12, p. 2945. 2018.

[89] A. Filchenkov, S. Muravyov, and V. Parfenov, "Towards cluster validity index evaluation and selection," in *Proc. IEEE Artif. Intell. Natural Lang. Conf. (AINL)*, Nov. 2016, pp. 1–8.

[90] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[91] A. Asuncion and D. Newman, (2007). *UCI Machine Learning Repository*. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[92] L. Lush, S. Ellwood, A. Markham, A. I. Ward, and P. Wheeler, "Use of tri-axial accelerometers to assess terrestrial mammal behaviour in the wild," *J. Zoology*, vol. 298, no. 4, pp. 257–265, Apr. 2016.

[93] U. Agrawal, D. Soria, C. Wagner, J. Garibaldi, I. O. Ellis, J. M. S. Bartlett, D. Cameron, E. A. Rakha, and A. R. Green, "Combining clustering and classification ensembles: A novel pipeline to identify breast cancer profiles," *Artif. Intell. Med.*, vol. 97, pp. 27–37, Jun. 2019.

[94] F. Farahbod and M. Eftekhari, "A new clustering-based approach for modeling fuzzy rule-based classification systems," *Iranian J. Sci. Technol. Trans. Electr. Eng.*, vol. 37, no. E1, p. 67, 2013.

[95] M. L. Williams, W. P. James, and M. T. Rose, "Variable segmentation and ensemble classifiers for predicting dairy cow behaviour," *Biosyst. Eng.*, vol. 178, pp. 156–167, Feb. 2019.

[96] N. Zanardo, G. Parra, C. Passadore, and L. Möller, "Ensemble modelling of southern Australian bottlenose dolphin Tursiops sp. Distribution reveals important habitats and their potential ecological function," *Mar. Ecol. Prog. Ser.*, vol. 569, pp. 253–266, Apr. 2017.

[97] S. Namasudra, D. Devi, S. Kadry, R. Sundarasekar, and A. Shanthini, "Towards DNA based data security in the cloud computing environment," *Comput. Commun.*, vol. 151, pp. 539–547, Feb. 2020.

[98] S. Namasudra, P. Roy, P. Vijayakumar, S. Audithan, and B. Balusamy, "Time efficient secure DNA based access control model for cloud computing environment," *Future Gener. Comput. Syst.*, vol. 73, pp. 90–105, Aug. 2017.

[99] S. Namasudra and P. Roy, "Time saving protocol for data accessing in cloud computing," *IET Commun.*, vol. 11, no. 10, pp. 1558–1565, Jul. 2017.

[100] S. Namasudra, "An improved attribute-based encryption technique towards the data security in cloud computing," *Concurrency Comput., Pract. Exper.*, vol. 31, no. 3, p. e4364, Feb. 2019.

[101] S. Namasudra, "Taxonomy of DNA-based security models," in *Advances of DNA Computing in Cryptography*, S. Namasudra and G. C. Deka, Eds. New York, NY, USA: Taylor & Francis, 2018, pp. 53–68.

[102] S. Namasudra and G. C. Deka, "Introduction of DNA computing in cryptography," in *Advances of DNA Computing in Cryptography*, S. Namasudra and G. C. Deka, Eds. New York, NY, USA: Taylor & Francis, 2018, pp. 17–34.

[103] G. Lucca, G. P. Dimuro, B. Bedregal, J. A. Sanz, and H. Bustince, "A proposal for tuning the alpha parameter in a copula function applied in fuzzy rule-based classification systems," in *Proc. 5th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2016, pp. 367–372.

[104] M. Elkano, M. Galar, J. Sanz, and H. Bustince, "CHI-BD: A fuzzy rule-based classification system for big data classification problems," *Fuzzy Sets Syst.*, vol. 348, pp. 75–101, Oct. 2018.

[105] L. Yao and Z. Ge, "Scalable semisupervised GMM for big data quality prediction in multimode processes," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3681–3692, May 2019.



**NIKITA JAIN** received the M.Tech. degree from IIIT Delhi (A State University under the Government of NCT Delhi), with a specialization in Information Security, in 2014. She is currently an Assistant Professor with the Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering. She has been a Ph.D. Research Scholar in computer science specializing in Remote Sensing and Information Retrieval with Guru Gobind Singh Indraprastha University, since 2016. She has more than 25 publications in various National, International Conferences, and Journals of high repute.



**DEEPALI VIRMANI** received the B.Tech. degree in computer science from MDU, Rohtak, the M.Tech. degree in information technology from GGSIPU, and the Ph.D. degree in computer science from Delhi University, India. She has an innovative work experience of more than 19 years in both research and academics. She is currently the Head of the Department and a Professor with the Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology affiliated to GGSIPU, New Delhi. She has published more than 90 research articles in International journals/National journals/International conferences of repute. She also works in a multi-disciplinary environment involving sensor networks, web intelligence, data mining and intelligent information retrieval systems, and machine learning applied to various real-world problems. She has more than 500 academic citations index as per Google scholar. She has guided more than 70 B.Tech. Projects. She is also guiding many Ph.D. scholars registered with reputed universities, such as GGSIPU and UPTU. She is Branch Counselor of the BPIT-IEEE student chapter and BPIT-CSI student branch. She is an Associate Editor of *Journal Open Computer Science* (De Grutyer) and also on the Reviewer Board of various IEEE, Elsevier, and springer journals. She has organized many professional activities, such as FDPs, workshops, expert lectures, and conferences. She has been the Session Chair in national/international conferences.

**AJITH ABRAHAM** (Senior Member, IEEE) received the Ph.D. degree in computer science from Monash University, Melbourne, VIC, Australia, in 2001, and the M.Sc. degree from Nanyang Technological University, Singapore, in 1998. He is currently the Director of the Machine Intelligence Research Labs (MIR Labs), a Not-for-Profit Scientific Network for Innovation and Research Excellence connecting Industry and Academia. The Network with HQ in Seattle, WA, USA, has currently more than 1 000 scientific members from more than 100 countries. As an Investigator/Co-Investigator, he has won research grants worth more than 100 Million U.S.$ from Australia, USA, EU, Italy, Czech Republic, France, Malaysia, and China. He is also works in a multi-disciplinary environment involving machine intelligence, cyber-physical systems, the Internet of Things, network security, sensor networks, Web intelligence, Web services, data mining, and applied to various real-world problems. In these areas, he has authored or coauthored more than 1 400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated to Japanese and few other articles were translated to Russian and Chinese. About more than 1 200 publications are indexed by Scopus and more than 900 are indexed by Thomson ISI Web of Science. Some of the articles are available in the ScienceDirect Top 25 hottest articles. He has more than 1 200 coauthors originating from more than 40 countries. He has more than 38 000 academic citations (H-index of 93 as per Google scholar). He has given more than 100 plenary lectures and conference tutorials (in more than 20 countries). For his research, he has won Seven Best Paper awards at prestigious International conferences held in Belgium, Canada Bahrain, Czech Republic, China, and India. Since 2008, he has been the Chair of IEEE Systems, Man, and Cybernetics Society Technical Committee on Soft Computing (which has more than 200 members) and has served as a Distinguished Lecturer of the IEEE Computer Society representing Europe, from 2011 to 2013. He is the Editor-in-Chief of Engineering Applications of Artificial Intelligence (EAAI) and serves/served the Editorial Board of more than 15 International Journals indexed by Thomson ISI. He is actively involved in the organization of several academic conferences, and some of them are now annual events.

**LORENZO SALAS-MORERA** (Member, IEEE) received the M.Sc. degree in agricultural engineering and the Ph.D. degree in agricultural engineering from the University of Córdoba, Spain, in 1989 and 1993, respectively. He is currently a Professor with the University of Córdoba in the area of projects engineering. He is the Vice-Chancellor of Digital University and Strategic Planning of Córdoba University. His primary areas of research are engineering design optimization, intelligent systems, machine learning, user adaptive systems, interactive evolutionary computation, project management, risk prevention in automatic systems, and educational technology.

**LAURA GARCIA-HERNANDEZ** (Member, IEEE) received the M.Sc. degree in computer science from the Universitat Oberta de Catalunya, Spain, in 2007, and the European Ph.D. degree in engineering from the University of Córdoba, Spain, and also from the Institut Français de Mécanique Avancée, Clermont-Ferrand, France, in 2011. She has been an Invited Professor during a semester with the Institut Français de Mécanique Avancée, Clermont-Ferrand. She is currently an Associate Professor in project management with the University of Córdoba. Her primary areas of research are engineering design optimization, intelligent systems, machine learning, user adaptive systems, interactive evolutionary computation, project management, risk prevention in automatic systems, and educational technology. In these fields, she has authored or coauthored more than 70 international research publications. She has given several invited talks in different countries. She has realized several postdoctoral internships in different countries with a total duration of more than two years. She received the prestigious National Government Research Grant José Castillejo for supporting their Postdoctoral Research during six months with the University of Algarve, Portugal. She has been an Investigator Principal in two Spanish research projects and has also been an Investigator Collaborator in some research contracts and projects. She is an Expert Member of ISO/TC 184/SC working team and the National Standards Institute of Spain (UNE). Moreover, she is a member of the Spanish Association of Engineering Projects (IPMA Spain). Considering her research, she received the Young Researcher Award granted by the Spanish Association of Engineering Projects (IPMA), Spain, in 2015. Additionally, she received two times the General Council of Official Colleges Award at prestigious International Conference on Project Management and Engineering both 2017 and 2018 editions. She is the Co-Editor-in-Chief of the *Journal of Information Assurance and Security*. She is an Associate Editor of the following ISI journals: *Applied Soft Computing* and the *Journal of Intelligent Manufacturing*.