IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 23, 2021, accepted March 15, 2021, date of publication April 9, 2021, date of current version April 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072120

Emotion Recognition Using Explainable Genetically Optimized Fuzzy ART Ensembles

WEI SHIUNG LIEW[1], CHU KIONG LOO[1], (Member, IEEE), AND STEFAN WERMTER[2], (Member, IEEE)

[1]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Informatics, Knowledge Technology Institute, University of Hamburg, 22527 Hamburg, Germany

Corresponding author: Chu Kiong Loo (ckloo.um@um.edu.my)

This work was supported in part by the Georg Forster Research Fellowship for Experienced Researchers from the Alexander von Humboldt-Stiftung/Foundation, in part by the German Research Foundation (DFG) on Crosssmodal Learning under Grant TRR169, and in part by the Impact Oriented Interdisciplinary Research Grant (IIRG) from the University of Malaya under Grant IIRG002C-19HWB.

ABSTRACT There is a growing demand for explainability in complex artificial intelligence solutions to support critical applications' decision-making processes. Barriers to explainable processes include black-box classifiers, such as deep learning, and noisy datasets. Affect recognition involving neural networks attempts to map complex human emotions onto Arousal and Valence scales based on physiological signal measurements. Datasets collected for this purpose are inherently noisy and may contain outliers and imbalanced classes, hindering accurate classification. In our approach, these issues are addressed using Fuzzy ART (FA) for clustering data samples into more condensed memory templates, introducing stochastic resonant noise to amplify signal-to-noise ratio, and SMOTE sampling to generate synthetic minority samples. A genetic algorithm is developed for FA optimization and ensemble model selection. Clusters obtained from the resulting ensembles are then used to train an ensemble of boosted decision trees for classification and to visualize the decision-making processes. Individual features such as heart rate variability and EEG band power, as well as feature interactions between pairs of features, may contain critical information as human affect indicators. Contributions of individual features and feature interactions toward describing human affect are quantified and interpreted using Shapley additive explanation values. Three established affect recognition datasets were considered for mapping physiological features onto a binary classification of Low/High Arousal and Positive/Negative Valence. Our framework was able to achieve good generalization for both classification tasks as well as provide detailed insights into the contributions of physiological features towards describing Arousal and Valence affects.

INDEX TERMS Affective computing, decision support systems, genetic algorithms, hybrid intelligent systems, knowledge-based systems, pattern clustering, regression analysis.

I. INTRODUCTION

A. MOTIVATION

With increasingly complex algorithms being used in everyday applications, it is difficult for humans, even expert users, to keep track of machine reasoning. Artificial intelligence (AI) in particular has achieved superhuman capabilities in some highly complex domains such as strategy games [1], medical diagnosis [2], and character recognition [3]. As similar systems are used in critical domains that would influence human lives and well-being, there is a need for humans to be able to understand machine reasoning on how such decisions are achieved [4]. Machine-generated and human-interpretable explanations that are generated alongside decisions would help to gain users' trust [5] and support post-hoc debugging and correction when the system produces a harmful decision. There is a growing need for understandable explanations from black-box methods such as convolutional networks and effective deep learning algorithms in complex domains. Therefore, explainability requirements would have to be balanced against a model's performance [6], [7].

B. PROBLEM STATEMENT

In domains where knowledge is based on inexact data, generating definitive and well-defined explanations for decisions

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

is a problematic proposition [8]. For example, physiological signals are commonly used for affect recognition due to higher data granularity and are less affected by cultural norms [9]. Several corpora of physiological signals were recorded by volunteers who were shown emotional stimuli to provoke specific affective states, for example, DEAP [10], DREAMER [11], and AMIGOS [12]. The datasets were made available for researchers to conduct affect recognition experiments, typically by performing signal processing and analysis followed by classification using neural networks.

A common challenge in emotion recognition research is trying to model the affect ground truth that has to date been done by simplifying human emotions using dimensionality descriptors such as Arousal and Valence (AroVal) scales in Russell's circumplex model [13]. Several factors have to be considered when selecting an emotion model to classify or quantify emotions during data collection. A simple model such as the circumplex reduces the complexity of emotion modeling into AroVal metrics. At the same time, emotional granularity is low, resulting in highly similar but distinct emotions being grouped in the same cluster [14]. On the other hand, using more specific emotion descriptors may encounter significant inter-individual differences, and specific emotion is more challenging to induce under laboratory conditions [15].

Other challenges include dealing with inter-person physiological differences when capturing sensory information and the lack of a universal methodology for processing the recorded signals to extract relevant features [16]. Trying and testing every known feature extraction runs the risk of high dimensionality. The combination of factors means that the datasets may not fully represent the participants' affective information, resulting in less-than-perfect affect recognition rates.

In addition, datasets gathered under laboratory conditions often have a limited number of data points due to time and resource constraints. A phenomenon known as class imbalance may occur when a dataset has an uneven distribution of labeled samples; i.e., samples belonging to the minority class are observed significantly less than majority samples. Classifiers trained with an imbalanced dataset may produce poor generalization performance with regards to the minority group.

Dataset pre-processing includes identifying salient and outlier data samples and reducing the dataset's dimensionality for faster processing. Clustering creates a topology that places highly similar objects closer to each other while dissimilar objects are placed further apart. Data redundancy is reduced by using a single representative for multiple highly similar samples. Self-organizing neural networks such as ARTs have inherent clustering capability. Instead of pre-defining a set number of clusters, self-organized clusters were allowed to form as a byproduct of the network's hyper-parameter settings. Depending on the hyper-parameter settings, however, the resulting topology of clusters may vary. The vigilance parameter in ART-based networks, for example, controls the clusters' quantity and granularity. Hyper-parameter tuning is usually necessary to achieve optimal clustering and varies between datasets.

C. CONTRIBUTION

This study proposes a holistic framework for generating explainable and usable information from physiological signals in the context of recognizing human affect. The framework includes pre-processing techniques to improve the dataset's quality before training a decision tree (DT) ensemble model for generating explanations. Using Shapley additive explanation values, individual contributions of samples and features can be observed to fine-tune the classification model by excluding poor-quality samples and features.

This work is organized as follows: the state-of-the-art approaches and fundamentals for explainable affect recognition are summarized in Section 2, Section 3 proposes a new methodology for explainable affect recognition, Section 4 outlines the conducted experiments and the results for different affect recognition model settings, and finally, Section 5 concludes our research.

II. EXPLAINABLE AFFECT RECOGNITION

Studies of affect or emotional states typically involve recording and statistical analysis of physiological signals. Annotations of affective states provide ground truth information, allowing researchers to locate specific affect indicators from features extracted from the signals. Datasets such as DEAP, DREAMER, and AMIGOS were obtained by recording physiological signals of volunteers who were shown multimedia stimuli to provoke affective responses. Annotations were then provided post-hoc by the volunteers using the AroVal scales. Affect recognition studies use a variety of approaches to correlate the physiological signals to the annotated affect information.

There have been several frameworks for affect recognition in the literature. Most studies focused on developing novel feature computation methods [17]–[19] or deep learning models for feature generation [20]–[22]. Convolutional methods are popular as there is no need for prior signal processing to generate highly relevant features. However, these are black-box methods, and hamper the generation of an explainable relationship between the measured physiological signals and affect. Common pre-processing methods were used including baseline normalization and feature selection using statistical methods [10]–[12], or signal filtering [23].

There are a few studies implementing explainability techniques for affect recognition. Most studies stopped at training a regression model such as DT to provide a visualized example [24], while little to no post-hoc analysis of the model was provided. Lin *et al.* [25] trained multiple convolutional models, each dedicated to a single physiological modality. Explainable information was then provided by observing each model's predictive output to determine which physiological modality is still important. This method, however, does not provide a detailed explanation of exactly which feature is essential. The development of efficient computation of Shapley additive explanation (SHAP) values [26] has introduced a powerful tool for generating useful explanations. Yang *et al.* [27] used SHAP to provide an extensive analysis of the relationship between EEG features to different annotation techniques for affect recognition. Shapley values can also be used in place of feature selection metrics [28] to determine the importance of individual features and data samples. So far, few studies compare cross-dataset findings using high-level explanations. Sarkar *et al.* [29] utilized a convolutional network-based method for self-supervised ECG learning for emotion recognition from multiple datasets but did not elaborate on the interactions between the features and affect. Yang *et al.* [27] focused on the AMIGOS dataset.

Before the classification step, the signals require significant pre-processing for dimensionality reduction and to extract information-dense features for computation. One issue in affect recognition is that a common affective stimulus or descriptor may induce different subjective emotional experiences across individuals [30]. In addition to using self-annotation to mitigate this effect from the subjective annotation perspective, self-organizing mapping (SOM) of the extracted physiological features can be useful for grouping similar affective responses and reduce inter-subject variance [31]. SOMs have been used as a pre-processing method for exploratory analysis of physiological signals in preparation for affect recognition in numerous studies [32]–[34].

Certain neural network models possess self-organizing qualities as a side-effect of the learning process. Adaptive Resonance Theory (ART)-based neuro-fuzzy networks [35] are particularly useful for abstract classification from noisy physiological data [36], [37], or for use as a clustering technique [38]. FAs utilize fuzzy logic for enhanced generalization but is statistically inconsistent due to training order dependency. Methods proposed for optimizing the performance of FAs typically include tuning the hyper-parameters such as vigilance and learning rate and determining the given dataset's optimum training order [39]–[41].

Physiological signals are inherently noisy, whether due to inter-subject variances and data collection methodology or for categorizing abstract concepts with inconsistent annotations. Feature extraction and selection may not represent fully the target domain (i.e., affect classes). The signal-to-noise ratio may be increased using the stochastic resonance phenomenon [42]. This technique has been used to enhance physiological signals for affect recognition [43]–[45]. Simulations with various clustering algorithms obtained a significant improvement in convergence speed when an amplitude-tuned noise signal was added to the clustered signal [46]. Tuning the noise signal is required to maximize signal gain as resonant noise varies across different datasets.

For self-annotation methods, class imbalance may occur when the annotations are unevenly distributed across the entire scale. For the DEAP and DREAMER datasets, class imbalance occurred when AroVal scores were divided into Low and High classes by setting the threshold at the center of the scale. The synthetic minority oversampling technique (SMOTE) [47] addresses this problem by generating new data points from minority classes, and has been used in applications with sparse [48] and noisy [49] samples and multimodal physiological signals for affect classification [50]–[52]. In some cases [50], [51], models trained with SMOTE-generated data did not perform well due to the so-called ''cold start'' problem. When the minority samples' population was low and the data dimensionality was high, adding synthetic samples to the dataset had a detrimental effect on model accuracy [53].

Tuning the hyper-parameters of computational techniques is often required to obtain better results than obtained by common or default settings. When a large number of hyper-parameters is involved along with an uncertain or complex fitness evaluation, the solution space may be highly complex and it may be difficult to determine the optimum combination(s) of settings with a cursory glance. Genetic algorithms (GAs) are effective due to their population-based approach to examine multiple points in solution space simultaneously and a mutation operator to allow the algorithm to move forward from local optima. GAs have been previously used for FA optimization [39]–[41], improving the performance of SMOTE [54], [55] and stochastic resonance [56], [57], and for ensemble model selection [41], [58].

From the review of literature, implementing a framework for generating explainable information for affect recognition comes up with several challenges. Initially, feature extraction and pre-processing is a necessary step to reduce data dimensionality for faster computation. Inter-subject variances, unreliable self-annotation, signal noise, and class imbalance may negatively affect any classifier's ability to map the extracted features to the affect labels. And lastly, explainable information was typically generated by training a regression model with little-to-no post-hoc analysis to relate affect to the measured physiological data.

III. METHODOLOGY

Given an affect recognition dataset, data augmentation was conducted using three techniques: stochastic resonant noise was added to the extracted features to increase the signal-to-noise ratio, SMOTE was performed to reduce class imbalance, and self-organized clustering was then conducted by training a FA classifier using the augmented dataset. The FA was then tested with unmodified held-out data samples and was evaluated based on the testing accuracy and the complexity of the self-organized clusters.

An optimum combination of hyper-parameters for stochastic resonance, SMOTE, and FA was required to generate a good cluster, i.e., a cluster having the least complex topology while yielding the best cross-validation testing performance. GAs were used for population-based multi-objective optimization. FA classifiers can be arranged in parallel in a classifier ensemble to outperform any single classifier. A second

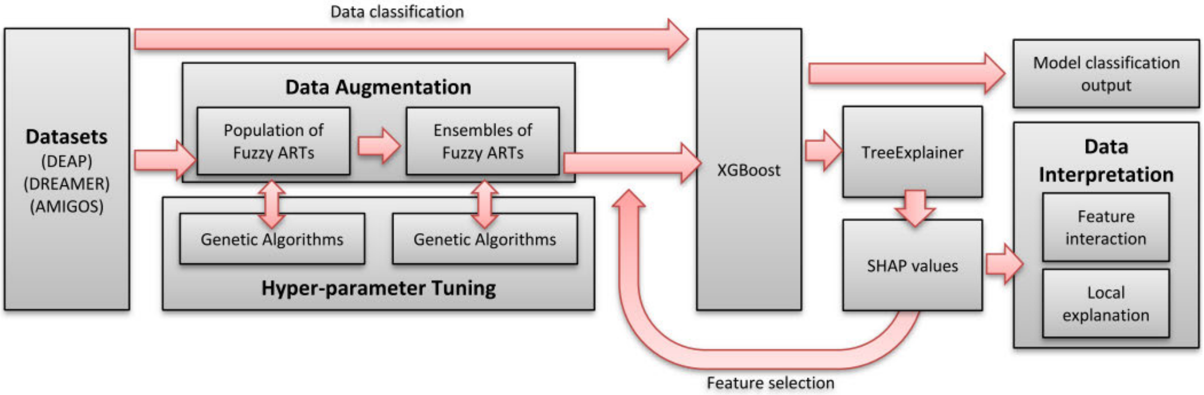

FIGURE 1. **The model consists of several modules. Emotion recognition datasets were clustered using genetic-optimized ensembles of Fuzzy ART classifiers. Hyper-rectangles extracted from the selected FAs were used for training an interpretable XGBoost decision tree model. SHAP values were computed from the DT, and feature selection was conducted on the basis of those values before retraining the DT with the reduced dataset. The final XGBoost DT was used for classifying testing data. Explainable information can be generated by visualizing the tree model, the SHAP feature interactions and the SHAP feature scores.**

GA was applied for selecting a subset of FAs to achieve minimal topological complexity with the best ensemble generalization performance.

The best-performing ensemble is considered as having learned the dynamics of the affect recognition dataset in the form of a more compact representation, the hyper-rectangles encoded in the component FA models in the ensemble. Hyper-rectangles are n-dimensional shapes where n represents the number of features or attributes in the dataset. Each shape embodies a unique learned category or template. Grouping similar data samples under one hyper-rectangle reduce redundancies and lessen the impact of noisy signals. Outliers are identified from hyper-rectangles that are formed from a minimal number of data samples.

The hyper-rectangles can then be extracted and assembled into a pseudo-dataset, optimized to eliminate redundancies, noise, and class imbalance in the previous steps. Explainable information was generated using the pseudo-dataset to train an extreme gradient boosted (XGBoost) DT model [59]. Shapley additive explanation (SHAP) values [26] were then used for evaluating the XGBoost model, generating SHAP scores for individual samples and features in the pseudo-dataset. Significant features were considered as having higher-than-average SHAP scores, indicating a high contribution to the classification. Subsequently, feature selection was performed in terms of high SHAP scores. The feature-selected pseudo-dataset was then used to train a final XGBoost model.

This section details the individual techniques used in the framework:

1) Stochastic resonance for signal-to-noise amplification.
2) Synthetic minority oversampling to reduce the imbalance between minority and majority classes.
3) Fuzzy ART for self-organizing clustering.
4) Genetic algorithms for tuning the hyper-parameters of the previous techniques and selecting multiple models for a classifier ensemble.

A. STOCHASTIC RESONANCE

Physiological signals are often noisy due to involuntary muscle movement or inexact electrode placement. When using signal filtering techniques such as band-pass, it is necessary to know the signal's approximate frequency range to be amplified while filtering all other signals. Another approach leverages the stochastic resonance phenomenon whereby the signal-to-noise ratio can be improved using additive white Gaussian noise (AWGN). The common frequencies in the noisy signal and the added white noise signal resonate to produce an amplified signal, which can then be isolated from the other noisy signals that were not boosted to the same degree.

Osoba *et al.* [46] demonstrated the benefits of additive noise in a variety of clustering applications, including unsupervised competitive learning (UCL). Assuming a UCL algorithm using a two-layer neural network topology, d-dimensional input patterns x making up the first layer and J competing neurons in the second layer. Simple distance matching is used to approximate competitive neuron dynamics in a winner-take-all connective topology, similar to ART-based networks. UCL node learning is performed by shifting the winning node's vector to become more similar to an incoming pattern. During simulations, an AWGN signal $n \sim N(0, \Sigma_\sigma(t))$ was added to the pattern vector x, producing the augmented training sample $y = x + n$. A scaled identity matrix was used for the noise covariance matrix $\Sigma_\sigma(t)$, with a standard deviation $\sigma > 0$ controlling the noise signal amplitude during the learning process. The variance decreases following a schedule:

$$\Sigma_\sigma(t) = (t^{-2}\sigma)I \tag{1}$$

Simulations for different datasets showed that the optimum convergence time was achieved for different values of σ [46].

B. SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

Class imbalance in a dataset occurs if each unique class label is not equally represented, i.e., under-represented class labels belong to minority classes and over-represented class labels belong to majority classes. The predictive accuracy of machine learning algorithms is significantly affected by data imbalance, which can be addressed using data resampling methods. Simple oversampling replicates minority samples with replacement until they are equal to the total number of samples in the majority classes, which, however, does not significantly improve minority class recognition [60].

Another oversampling approach known as SMOTE [47] creates synthetic examples of minority classes instead of direct duplicates. Synthetic samples are generated as follows: one assumes a target minority sample and k nearest neighbour minority samples selected for consideration, each with feature length d. For each feature, the lower and upper bounds are identified, and a random point is selected between them. When repeated for the entire feature vector, the newly-generated synthetic sample would effectively be located in the region between the nearest neighbours' extremes. This approach ensures that the decision region of the minority samples becomes wider, thus, more general.

Degrees of SMOTE were determined using the two parameters oversampling rate and number of nearest neighbors. The oversampling rate determines how many synthetic samples will be generated. For instance, setting a rate of 200% would produce two synthetic samples for each minority sample. A low value may not fully compensate imbalance in highly imbalanced datasets, while a high value leads to the "cold-start" problem. This problem occurs if the population of the minority samples is low while the dimensionality of samples is high. Generating many synthetic samples, in this case, would negatively affect model accuracy [53]. The number of nearest neighbours determines how many minority samples will be used as references for generating synthetic minority samples. A low neighbourhood value produces highly redundant or duplicated synthetic samples, while a high neighbourhood value produces samples that are less representative of the actual samples. Therefore, careful tuning of both parameters was required to achieve optimum model performance.

C. FUZZY ART CLUSTERING

With a dataset with an arbitrarily large number of observations, cluster analysis serves as a pre-processing step to achieve two goals: identifying which observations are highly similar and dissimilar to each other and suitable dimensionality reduction. Clustering methods can reduce a broad set of observations into smaller representative samples to enable faster and easier computation. In this study, Fuzzy ART was selected as the clustering method for several advantages, including its self-organizing and self-supervised incremental training capability, which allows ART networks to overcome the stability-plasticity dilemma.

FA networks are regulated by the use of three parameters determined before the training process.

1) The **choice parameter** α is used for selecting candidate weight vectors based on their similarity to the input vector to be trained.
2) The **learning rate** β represents the momentum where the winning weight vector is adjusted in response to the input. Fast learning is implemented by setting it to a value close to 1, meaning the weights are highly dynamic. However, this may result in significant catastrophic forgetting as older knowledge would be overwritten by newer information.
3) The **vigilance threshold** ρ sets the minimum matching value to determine whether to create a new hyper-rectangle in response to the input or to activate and update a pre-existing one. Setting a high threshold encourages the creation of more granular rectangles.

Given an M-dimensional input vector x_i, complement coding was performed by rescaling x to [0, 1] and augmenting the input with its complement:

$$x'_i = [x_i, 1 - x_i] \tag{2}$$

Fuzzy membership values were then computed between x'_i and existing rectangles w_j:

$$T_j = \frac{|x'_i \wedge w_j|}{\alpha + |w_j|} \tag{3}$$

where the choice parameter $\alpha \in [0.0, 1.0]$. The fuzzy AND operator $\wedge$ was given as:

$$(p \wedge q) \equiv \min(p_m, q_m) \tag{4}$$

while the norm was defined as

$$|p| = \Sigma_{m=1}^{M} p_i \tag{5}$$

If there were no pre-existing hyper-rectangles (i.e., at the start of the training), then the next step was skipped, and the input was immediately added as a new hyper-rectangle. Hypothesis testing was then conducted in sequential order, starting from the hyper-rectangle j with the highest membership value T_j:

$$\frac{|x'_i \wedge w_j|}{|x'_i|} \geq \rho \tag{6}$$

where $\rho \in (0.0, 1.0]$ was the vigilance threshold. If Equation 6 was satisfied, the input vector was found to match with the hyper-rectangle j and learning was performed:

$$w_j^{new} = (1 - \beta) w_j^{old} + \beta(x' \wedge w_j^{old}) \tag{7}$$

with $\beta \in (0.0, 1.0]$ as learning rate.

Otherwise, if the hypothesis testing failed, the process was repeated for the next best matching hyper-rectangle until all hyper-rectangles were tested or no suitable matches were found. If no match was found, x'_i was then added to the layer as a new hyper-rectangle.

While the FA was used primarily for unsupervised learning, each hyper-rectangle's categorical information is further needed. Thus, instead of using a mapping module for classification a la ARTMAP, an associative matrix was used to store and map the association between hyper-rectangles and their respective labels. This method has been used in similar self-organizing networks to map associations between learned categories and their respective class labels while not interfering with the unsupervised learning process [61]. Whenever a hyper-rectangle j was activated in response to an input with a label l, the corresponding entry in the matrix $H(j, l)$ was incremented by a value equal to the fuzzy membership value:

$$\Delta H(j, l) = T_j \tag{8}$$

Unlike ARTMAPs, labels were not taken into account before learning to preserve the FA's self-organizing quality. In applications with high inter-subject variability, the supervised learning approach may produce many hyper-rectangles, creating an overfitting problem. In contrast, the unsupervised mapping approach mitigates the effect of abnormal subjective affect annotations in favour of hyper-rectangle similarity.

When the hyper-rectangle j was activated in response to the input, the label with the strongest association was selected as the predicted label:

$$\text{label} = \underset{l \in L}{\text{argmax}}\, H(j, l) \tag{9}$$

Validating the effectiveness of the trained network was performed using hold-out validation sets. Each sample in the dataset was assigned to one fold. The samples in one fold were designated as the hold-out, while the samples in all other folds were used for training an FA network. The process was repeated using a different fold as hold-out and training a new FA network each time. Validation accuracy was then averaged to obtain a generalization score.

D. HYPER-PARAMETER OPTIMIZATION USING GENETIC ALGORITHMS

The incremental learning nature of the FA networks makes them susceptible to ordering effects. Several approaches reported either averaging classifiers with multiple arbitrary training sequences [62] or used an optimization technique such as GA to find the best sequence [40]. In addition, hyper-parameter tuning is necessary to ensure that the network can cluster the training data with minimal information loss.

Several hyper-parameters were subject to optimization as follows:

1) **Sample importance score**. Assuming a dataset with N samples, sample importance was represented by a string of values $\{s_1, \ldots, s_N\}$, where $\Sigma_{n=1}^{N} s_n = 1$. The sequence in which samples were presented during training was determined by ordering the sample importance scores in descending order. Training order affects performance of FAs [40], [63].
2) **Sample subset selection**. Represented by a single floating point in the (0.0,1.0] range representing the fraction of dataset samples to be selected for training and testing. When combined with the sample importance score, samples were selected for training from the most important to the least important. A subset value of less than 1.0 will exclude the least important samples and improve training time. Subset selection does not affect the samples selected for hold-out testing.
3) **Fuzzy ART hyper-parameters**. The FA was initialized for a given choice value, learning rate, and vigilance threshold, represented by floating points in the range (0.0,1.0].
4) **Perturbation**. A parameter in the range [0.0,1.0] controls the stochastic noise's maximum amplitude to be introduced. A white noise signal was generated using a Gaussian function and was added to samples before training. Additive noise was neither accumulated nor carried over when the same data sample was used for training in other cross-validation sets or phenotypes. Noise was further not added to samples designated for hold-out testing.
5) **SMOTE parameters**. Two parameters govern the behaviour of SMOTE for generating synthetic samples. The oversampling ratio determines how many synthetic samples will be generated for each minority sample and is set to an integer in the [1,5] range. The nearest neighbour parameter determines how many nearest neighbouring minority samples are used as references for generating synthetic samples and is set to an integer in the [1,10] range. The hyper-parameter ranges were selected based on the findings of Elreedy *et al.* [53].

The combination of hyper-parameters results in a highly complex solution space with potentially multiple Pareto-optimal solutions. Two target objectives were defined: to achieve the maximum testing generalization performance with the least complex clustering topology. Genetic algorithms were used for hyper-parameter tuning as an extension from our previous work [63].

Genetic algorithms is a population-based method for hyper-parameter optimization. A candidate solution or phenotype represents a possible configuration for initializing a FA classifier's parameters and conducting training with a specific sequence of data samples. The phenotype fitness was evaluated using a fitness function such as the trained FA's hold-out classification accuracy. From an initial population of randomized phenotypes, a GA incrementally traverses the solution space by discarding low-fitness phenotypes and using biologically-inspired genetic operators to generate new phenotypes by combining genetic traits from the remaining high-fitness survivors. Repeating the process over multiple generations would evolve the population towards optimal solutions.

The string of values within a single phenotype determines an FA model's initialization and for preparing the data samples for training and testing. When initializing the GA, a population of randomly-generated phenotypes was generated. Fitness testing was conducted for each phenotype as follows:

1) The sequence of data samples was reordered following the sample importance scores in descending order to ensure that important samples were presented much earlier to the FA.
2) Samples were split into training and testing following the K-fold cross-validation scheme.
3) When performing training for each fold, the sample subset selection parameter chose a fraction of data samples with the highest importance scores for training.
4) Selected training samples were augmented by artificially generated samples using SMOTE [47] to reduce class imbalance. Newly generated training samples were concatenated below the originally selected training samples to minimize their impact on the ordering effect. Altogether, the number of augmented and unaugmented training samples in the dataset was denoted as D.
5) The training samples were augmented with a white Gaussian noise signal with maximum amplitude with respect to the perturbation parameter.
6) A new FA model was initialized using the vigilance, choice, and learning rate hyper-parameters defined in the phenotype. The training was conducted using the augmented training dataset and then tested using the hold-out samples.
7) Fitness was determined by an index score combining a performance metric and a small penalty score proportional to the sum of all hyper-rectangles in the FA:

$$\mathrm{fit}_k = F1 - \lambda_g \frac{c}{D} \tag{10}$$

where $F1$ is the testing fitness from the hold-out samples, c is the number of hyper-rectangles in the FA, and λ_g sets the importance of the penalty function, ideally set to a sufficiently small value so that a small difference in the number of hyper-rectangles will not overly penalize the fitness score.
8) Steps (3) to (7) were repeated for K cross-validation folds. The overall phenotype fitness was then averaged across all folds.

GA traverses the solution space using genetic operators selection, reproduction, and mutation. Selection determines which phenotype with high fitness can propagate to the next generation while low-fitness phenotypes are discarded. A new phenotype is created for each discarded phenotype through genetic reproduction, essentially averaging two randomly selected parent phenotypes' genetic values. The next iteration of the population thus consists of high-fitness phenotypes and their newly generated offspring. The generation counter is then incremented, and the offspring are fitness-tested to be compared against older phenotypes.

Genetic convergence occurs when newly generated phenotypes are highly similar to pre-existing phenotypes, typically due to combining two highly similar parents. If unchecked, convergence would result in the entire population consisting of exact clones of one another. Early convergence is undesirable as the GA may unduly focus on the local maxima while ignoring the rest of the solution space. Premature convergence is mitigated by introducing a genetic mutation parameter, defined as the probability of a gene to be mutated. In this case, genetic mutation was introduced by adjusting the value of a parameter by adding a small positive or negative number.

The GA was said to have converged onto an optimum point after fulfilling one or more stopping criteria or until an arbitrarily large number of generations have passed. Therefore, the phenotype with the highest fitness score was considered the fittest hyper-parameters to train the FA.

The consistency and efficiency of the GA to reach the global optimum depends on several parameters.

1) **Population size** maintains a set number of phenotypes at any point in time. A low value restricts the initial search space while a high value increases computation time.
2) **Genetic selection** determines the proportion of phenotypes to be carried over from one generation to the next. A low value may prematurely discard potentially optimum solutions, while a high value negatively impacts the GA's ability to traverse the search space.
3) **Genetic mutation** introduces a small probability of changing a parameter value within a phenotype. A low value negatively impacts the GA's ability to escape a local optimum point, while a high value negatively affects the GA's ability to converge to the optimal solution.

Adaptive control of the selection and mutation parameters was implemented for efficient searching [64]. The genetic selection was set to a low value, and genetic mutation was set to a high value to widen the solution space at early stages. As the GA converges, the genetic selection was increased to retain more optimal solutions, and the genetic mutation was decreased for smaller incremental steps in the solution space.

The optimization process may produce highly overfitted solutions, i.e., classifiers that generalize well only in a narrow scope. Combining multiple classifiers in an ensemble model may produce more robust predictive ability [39], [65].

E. ENSEMBLE MODEL SELECTION USING GENETIC ALGORITHMS

A classifier ensemble is a paradigm where multiple classifiers are cooperatively used for solving a problem. Decision-level fusion methods take the predictive outputs from multiple classifiers to be combined to select the best predictive output, typically using some kind of decision-fusion or voting scheme. After optimizing a classifier population, a second

GA was employed to find the optimum combination of FAs to achieve maximum generalization.

Model selection was conducted using a simple binary phenotype of dimension equal to the number of FAs K, where 0 or 1 indicated a classifier's membership in the ensemble. Decision fusion was performed using probabilistic plurality voting [63]. The ensemble fitness function was designed to maximize ensemble generalization with the least number of hyper-rectangles across all constituent FAs:

$$\text{fit} = F1 - \lambda_h \frac{\Sigma_{n=1}^{K} k_n C_n}{N} \tag{11}$$

where $F1$ is the testing fitness of the combined ensemble predictive output, k_n is the membership of the n^{th} classifier, C_n is the number of hyper-rectangles in the n^{th} classifier, N is the sample size of the dataset introduced as a normalization factor, and λ_h sets the weight of the penalty function of the ensemble size in relation to the fitness score. Ideally, λ_h is set to a sufficiently small value so that adding an extra member to the ensemble with a minor overall improvement to the fitness score can be avoided.

A GA was then used in a similar manner as in the previous section: a randomized population of phenotypes was initialized, whereby each phenotype represented the membership of an ensemble of classifiers. Phenotype fitness was judged based on ensemble generalization and size. The optimum solution would ideally consist of an ensemble with the minimum number of hyper-rectangles from FAs to maximize ensemble accuracy. While each FA consists of a condensed representation of the same dataset, the ordering effect and different parameter settings may produce sufficiently distinct hyper-rectangles.

F. FEATURE INTERPRETATION USING XGBoost AND SHAP

DT models data in a flowchart-like structure where each successive node marks a feature threshold splitting the decision path. An input vector to be classified can be represented by a continuous path through the DT starting from the root node and ending at a leaf node representing an outcome class label. XGBoost is a highly efficient and scalable DT methodology encompassing an ensemble of DT models and has been favorably benchmarked against deep learning techniques in several applications [66].

Given a dataset with data sample x_i and target class y_i, the XGBoost ensemble of trees performs classification as follows:

$$\hat{y}_i = \Sigma_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{12}$$

where K represents the number of decision tree models in the ensemble, f_k is an independent tree structure from the space of all possible regression trees F. The set of trees in the ensemble was determined by optimizing the objective function:

$$\text{obj} = \Sigma_{i=1}^{n} l(y_i, \hat{y}_i) + \Sigma_{k=1}^{K} \Omega(f_k) \tag{13}$$

where l is a loss function to quantify the dissimilarity between the model outcome $\hat{y}_i$ and the target class y_i, and Ω is the penalty weight for the complexity of the model f_k. During the training process, XGBoost uses an efficient greedy algorithm to evaluate all features to find the best possible split. From the completed ensemble of trees, the feature's importance can be estimated from the frequency of splits using that feature. This method is inconsistent, however, and the importance score of a feature may not fully reflect its impact on the model [26].

The TreeExplainer [66] provides a consistent explanation method for DT with the advantage of efficient computation of local explanations using Shapley additive explanation (SHAP) values. A global overview of feature importance towards classification is observed from combining the local explanations across the whole dataset. By focusing on features with significant SHAP values, feature selection and interaction analysis can be conducted on a narrower set of data to further refine the classification model.

IV. EMOTION RECOGNITION BENCHMARK EXPERIMENT

Three affect recognition datasets were used as benchmarks in this study: DEAP, DREAMER, and AMIGOS. The datasets incorporate volunteers' physiological signals in response to emotional multimedia stimulus chosen to evoke a specific affective state. The recorded signals' affective ground truth was obtained by asking the volunteers to self-assess their affective state after each stimulus was shown. Affect recognition was conducted using neural networks to classify the extracted physiological features and the corresponding ground truth affective scores. Table 1 highlights the main differences between the datasets.

The respective methodologies used for collecting the data of each dataset were mostly similar: affective stimuli were selected from a larger group of stimuli that were previously scored on the AroVal scales by volunteers. Extreme stimuli of the AroVal scales were then hand-selected to maximize affective response from participants. Four groups of stimuli were assembled from the four quadrants consisting of Low or High Arousal combinations with Negative or Positive Valence. To define the terminologies: Arousal represents the intensity of the affect, while Valence is the intrinsic quality of the affect (i.e., pleasant or unpleasant emotions).

Stimuli were then presented to participants one at a time while devices measured their physiological signals simultaneously. After viewing each stimulus, participants were instructed to score their affect for several scales including Arousal, Valence, Liking (individual preference for the stimuli), and Dominance (influence exerted by the stimuli over the individual).

Electroencephalogram (EEG) was recorded from specific electrode positions following standard conventions. Cardiac activity was recorded using blood volume pulse (BVP) in DEAP and electrocardiogram (ECG) electrodes in other datasets. Other modalities include galvanic skin response (GSR), electrooculogram (EOG), electro-myogram (EMG), respiration (Rsp), and skin temperature (Tmp). As the

TABLE 1. Summary of the emotion recognition datasets used in the experiment section.

	DEAP	DREAMER	AMIGOS
Participants	32	23	40
Stimulus	40 one-min. music videos.	18 one-min. music videos.	16 short (<4 min.) and 4 long (>14 min.) movie clips.
Physiological channels	EEG, BVP, EMG/EOG, GSR, Rsp, Tmp	EEG, ECG	EEG, ECG, GSR
Feature normalization	Baseline features extracted from 5 seconds prior to experiment, and subtracted from stimulus features.	Baseline features extracted from final 4 seconds of control recording. Stimulus features were divided by baseline features.	Stimulus features normalized to [-1, +1] by feature, recording session, and subject.
Affect self-scoring	Arousal, Valence, Dominance, Liking on [1,9] scale.	Arousal, Valence, Dominance on [1,5] scale.	Arousal, Valence, Dominance, Liking, Familiarity on [1,9] scale.
Binary label threshold	5.0	3.0	Median of population scores for each scale.
Training and testing	Leave-one-participant-out cross-validation	Ten-fold cross-validation using stimuli grouped into folds	Leave-one-participant-out cross-validation

datasets provided the actual physiological signal recordings from the experiments, feature extraction was performed manually following the instructions provided by the respective authors of each dataset. Likewise, feature normalization was performed differently for each dataset, as shown in Table 1.

Two types of affect recognition experiments were formulated: classifying Arousal as Low or High and classifying Valence as Negative or Positive. Participant scores were discretized into two class labels for the Arousal and Valence metrics, respectively, using score thresholds shown in Table 1. DEAP and DREAMER datasets set the threshold to the center of the scale regardless of the population distribution, while AMIGOS used population-based statistical scores as the threshold. Training and testing were conducted using the cross-validation strategies following the methodology outlined by respective authors. DEAP and AMIGOS used subject-dependent cross-validation while DREAMER used stimuli-dependent cross-validation.

Each dataset was first used as inputs to the GA for FA optimization. A second GA was then developed for ensemble model selection. Ensembles with the highest fitness scores were selected, and hyper-rectangles were extracted from their constituent FAs for training an XGBoost classifier. SHAP values for each feature were computed after converting the ensemble of DT into a TreeExplainer model. Significant features were identified by summing up the absolute SHAP values and selecting features with above-average values. The reduced feature set was then used for training the GAEFA-XGB.

While there have been newer ART models since the introduction of Fuzzy ART, preliminary benchmark tests with Bayesian ART (BA) [67] models showed that FAs achieved superior performance. A comparison of the generalization performance of both models will be given in the results section. Similar to FAs, hyper-parameter optimization for BAs was performed using GA [68], [69].

Feature extraction, FA, and GA were run using MATLAB, while XGBoost and SHAP were implemented using Python. Population size was set to 50 for both GAs, the λ penalty weight was set to 0.1 for both fitness functions at Equations 10 and 11. The GA's progression was tracked by observing the phenotypes' fitness scores at each generation. The GA was stopped when the population fitness does not improve over several consecutive generations.

TABLE 2. F1-scores for binary Arousal and Valence recognition for genetic-optimized individual (GAFA) and ensembles of FAs (GAEFA) and for individual and ensembles of BAs (GABA, GAEBA).

Dataset	Exp.	Aro	Val
DEAP	GAFA	0.560	0.591
	GAEFA	**0.615**	**0.620**
	GABA	0.557	0.559
	GAEBA	0.568	0.583
DREAMER	GAFA	0.607	0.571
	GAEFA	**0.644**	0.608
	GABA	0.569	0.569
	GAEBA	0.600	**0.610**
AMIGOS	GAFA	0.603	0.559
	GAEFA	**0.659**	**0.636**
	GABA	0.572	0.554
	GAEBA	0.583	0.598

V. RESULTS AND DISCUSSION

This section reports the performance metrics computed from the methodologies used in this study. Metrics of GAFA, GAEFA, and GAEFA-XGB, before and after feature selection, are benchmarked against contemporary studies using the same feature sets and cross-validation methodology used by the datasets' respective authors. Lastly, significant feature interactions are highlighted for consideration.

A. CLUSTERING ANALYSIS

The F1-score metric was used for evaluating the fitness of the classifier models, in part so that the findings of this study can be compared to other studies using the same affect recognition datasets. Table 2 reports the best F1-scores of individuals and ensembles of FAs and BAs after genetic optimization, using the cross-validation strategies outlined in Table 1. FA ensembles outperformed BA ensembles in five of the six classification tasks, suggesting that FA models are more suitable for clustering the affective datasets.

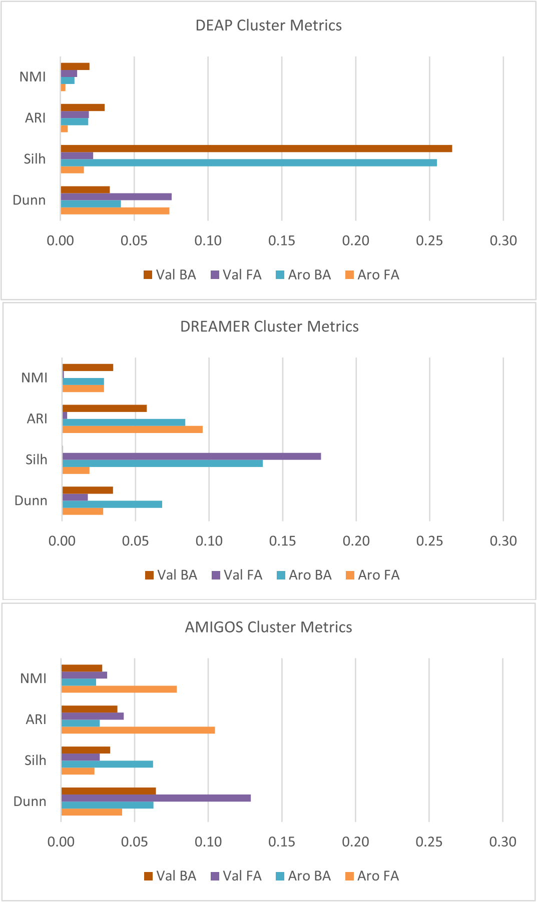

FIGURE 2. **Cluster metrics for benchmark datasets, measured using Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Silhouette coefficient (Silh), and Dunn index.**

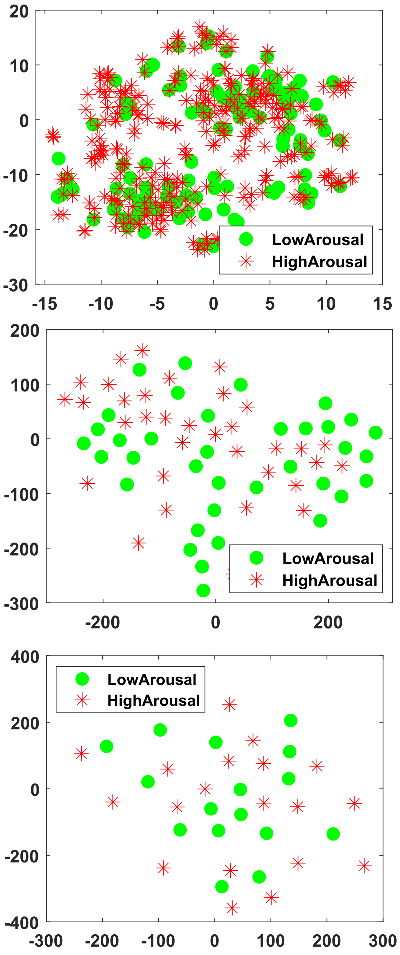

FIGURE 3. **Distribution of training data samples (top) and the FA-clustered exemplars (center) and BA-clustered exemplars (bottom) for the DREAMER dataset, visualized using TSNE.**

The hyper-volume clusters obtained from the FA and BA ensembles were assessed using four metrics that have been used previously for evaluating self-organizing networks [70], [71]. Dunn index [72] and silhouette coefficient [73] were used for internal cluster evaluation, measuring intra- and inter-cluster similarities. Normalized Mutual Information (NMI) [74], and Adjusted Rand Index (ARI) [75] were used for external evaluation, measuring how well the clusters could classify held-out data points and benchmarked against ground truth.

Fig. 2 shows the cluster metrics for the best-performing FA and BA ensembles after GA optimization. For the DEAP dataset, BA clusters were shown to outperform FA clusters for all metrics except the Dunn index. In contrast with the silhouette coefficients that show that BA clusters were well-separated compared to FA clusters, the Dunn indices appear to show that FA clusters were otherwise more dense than BA clusters. When clustering Valence for the DREAMER dataset, BA outperformed FA in all metrics except silhouette, indicating some overlap among the BA clusters, while FA clusters displayed significantly less overlap. For Arousal classification, BA and FA roughly had an equal performance for the external metrics (NMI and ARI). BA clusters were superior to FA clusters for internal metrics (silhouette and Dunn), suggesting that both clusters displayed the same classification ability, although BA clusters were significantly denser and separated. When clustering Valence for the AMIGOS dataset, FA and BA clusters were approximately equal except for the Dunn index indicating that FA clusters may be denser than BA. For Arousal clusters, FA significantly outperformed BA for external metrics. At the same time, BA was better for internal metrics, suggesting that the FA clusters were better at classifying but were topologically distributed. In contrast, BA clusters were less accurate when used for classification but were denser and well-separated.

Fig. 3 demonstrates how an optimized FA was able to condense the DREAMER dataset into a much smaller and distinctive set of exemplars, visualized using TSNE [76]. The top diagram shows the distribution of the data samples for the binary Arousal classification task. The distribution of the hyper-rectangles of the best-performing GAEFA is shown in the center diagram, displaying significant dimensionality reduction and distinct separation between the two class labels.

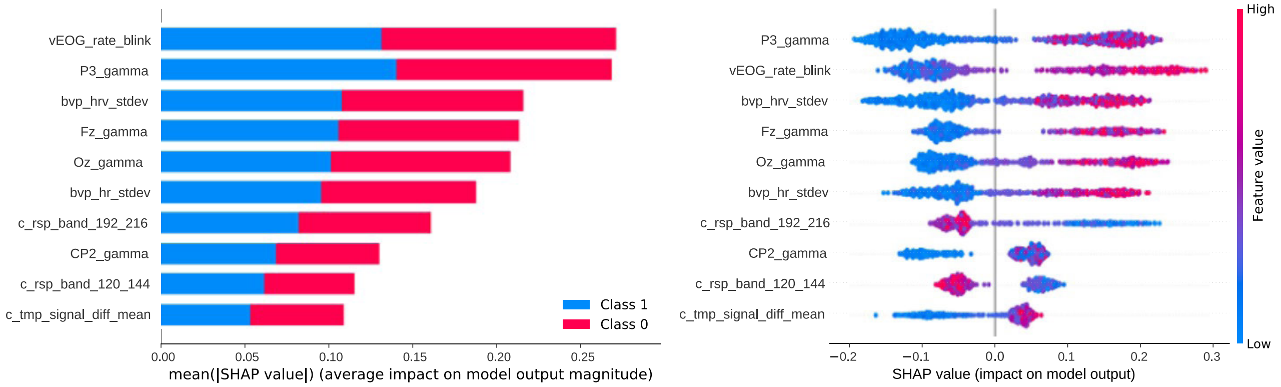

FIGURE 4. SHAP summary plots for Arousal classes (left) and High Arousal class (right) for DEAP dataset.

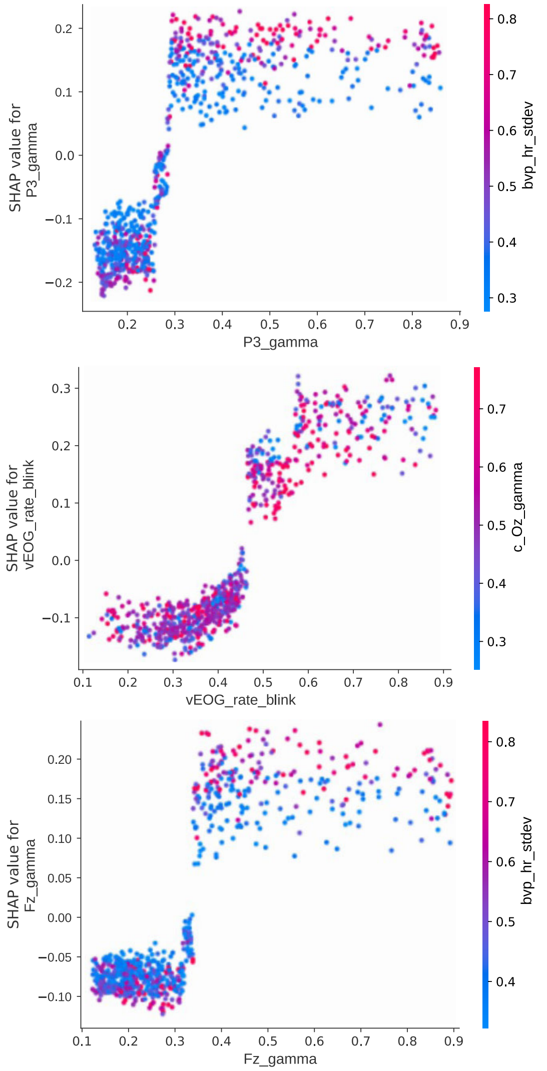

FIGURE 5. Significant feature dependencies for DEAP Arousal classification.

The bottom diagram shows the distribution of clusters of the best-performing GAEBA, indicating significantly higher dimensionality reduction. However, the separation between classes was not as distinct and delineated as compared to FAs.

TABLE 3. Performance comparison of GAEFA and GAEFA-XGB before and after feature selection.

Dataset	Exp.	Aro	Val
DEAP	[10]	0.616	0.647
	[77]	0.541	0.512
	GAEFA	0.615	0.620
	GAEFA-XGB	0.630	0.750
	GAEFA-XGB w/ feat. sel.	**0.654**	**0.774**
DREAMER	[11]	0.575	0.521
	GAEFA	0.644	0.608
	GAEFA-XGB	0.820	0.725
	GAEFA-XGB w/ feat. sel.	**0.833**	**0.807**
AMIGOS	[12]	0.564	0.560
	[78]	0.644	0.671
	GAEFA	0.659	0.636
	GAEFA-XGB	0.730	0.671
	GAEFA-XGB w/ feat. sel.	**0.749**	**0.680**

TABLE 4. Impact of feature selection on predictions.

Dataset	Exp.	Impact	Δ Confidence
DEAP	GAEFA-XGB Aro.	0.118	0.235
	GAEFA-XGB Val.	0.066	0.176
DREAMER	GAEFA-XGB Aro.	0.095	0.320
	GAEFA-XGB Val.	0.263	0.353
AMIGOS	GAEFA-XGB Aro.	0.171	0.123
	GAEFA-XGB Val.	0.470	0.400

B. FEATURE SELECTION

Table 4 shows the outcome of computing the impact of excluding low-SHAP features on the predictive ability and confidence of the models. The Impact Score is a measure of how the predictions of the model changes in response to a subset of critical factors being excluded, ranging from 0 (no impact) to 1 (maximum effect), and is computed as a function of the decisions before and after feature selection:

$$I = \frac{1}{n}\Sigma_{i=1}^{n}(y_i' \neq y_i) \tag{14}$$

The Δ Confidence metric measures the shift in confidence of the predictions after a subset of features was excluded. A negative score indicates that excluding the features has a detrimental effect on testing performance, while a positive value shows an improvement:

$$I = \frac{1}{n}\Sigma_{i=1}^{n}(y_i' == y_i) \vee (z_i' - z_i) \tag{15}$$

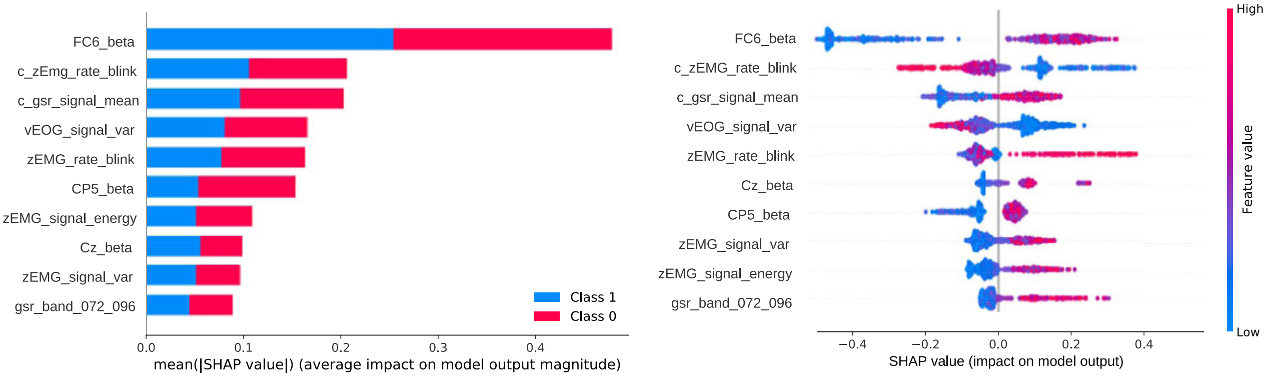

FIGURE 6. SHAP summary plots for Valence classes (left) and Positive Valence class (right) for DEAP dataset.

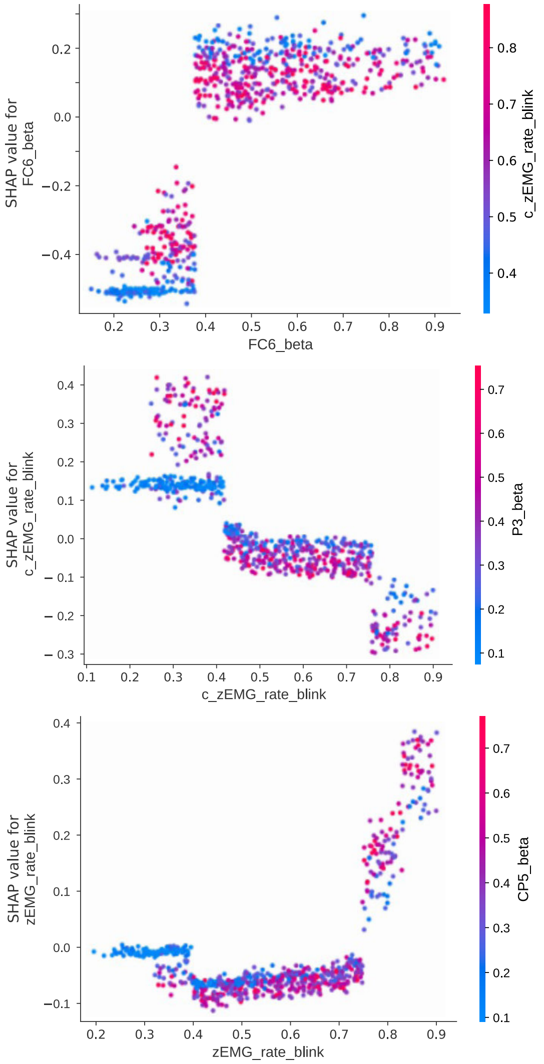

FIGURE 7. Significant feature dependencies for DEAP Valence classification.

In general, feature selection has a positive impact on model generalization. Large Impact Scores such as in the Valence classification of DREAMER and AMIGOS datasets indicate a significant negative effect of irrelevant features towards classification. Small Impact Scores such as in the Valence classification of DEAP dataset show that the removed features have less impact on classification. The positive Δ Confidence metrics indicated an overall improvement in generalization.

SHAP values were calculated from GAEFA-XGB, representing the overall impact of each feature over the entire model. Features with above-average SHAP values were then selected for retraining. Significant feature interactions were selected for analysis. F1-scores of GAEFA-XGB before and after feature selection are reported in Table 3, benchmarked against contemporary studies using the same datasets, features, and cross-validation strategies. Results show an improvement in generalization after each methodology was applied.

For Arousal classification of the DEAP dataset, 28.8% of features were found to have above-average SHAP values, totaling 73.0% of the combined SHAP values. EEG features make up 36.5% of the most significant features, while peripheral features constitute 63.5%, with BVP making up 39.7% followed by EMG/EOG (14.0%), Respiration (6.3%) and GSR (3.5%). For Valence classification, 26.3% of features were significant, totaling 78.2% of the combined SHAP values. EEG makes up 59.7% of the significant SHAP values, followed by EMG/EOG (18.4%), GSR (11.2%), with Rsp, BVP, and Tmp features making up 10.7% combined.

For Arousal classification of the AMIGOS dataset, 22.3% of features have above-average SHAP, totaling 82.6% of the combined SHAP values. EEG features make up 61.9% of the significant SHAP values, followed by ECG (33%) and GSR (5.1%). For the Valence classification, 31.9% features were considered significant, totaling 72.3% of the combined SHAP values. 66.9% of the significant features were EEG, followed by ECG (26.7%) and GSR (6.4%).

For the DREAMER dataset, EEG features were not as prominent as compared to the other datasets. In the Arousal classification task, ECG features make up 77.3% of the combined SHAP scores of significant features, while EEG makes up 22.7%. In total, 28.1% of features were significant, representing 74.9% of the total SHAP values. For Valence classification, ECG features make up the majority, consisting

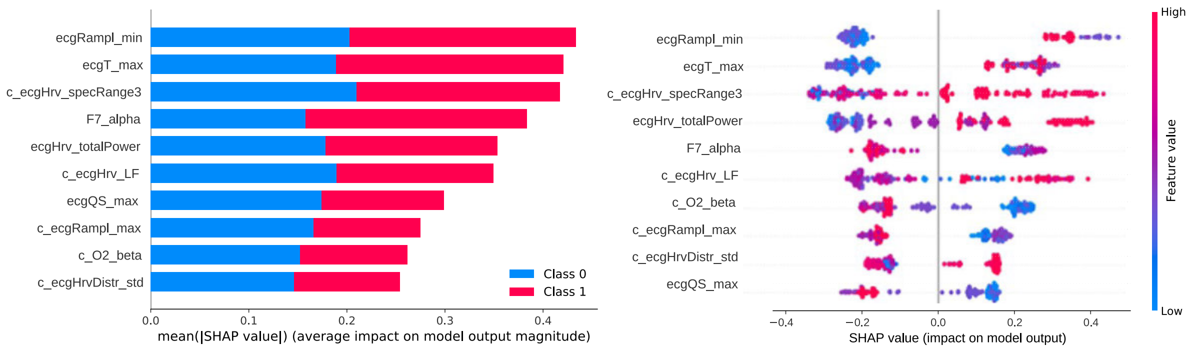

FIGURE 8. SHAP summary plots for Arousal classes (left) and High Arousal class (right) for DREAMER dataset.

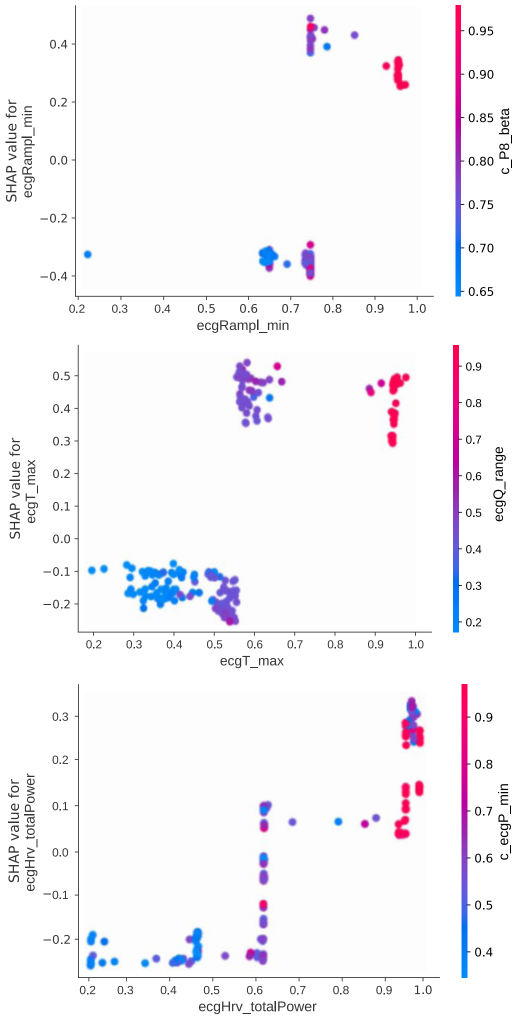

FIGURE 9. Significant feature dependencies for DREAMER Arousal classification.

of 63.9% of the combined SHAP values for significant features and the remaining 36.1% for EEG features. Only 18.8% of the features have above-average SHAP, totaling 88.6% of all SHAP values.

Each dataset's significant features were selected for retraining the GAEFA-XGB, with the results being presented in the next section.

1) DEAP EXPLANATIONS

Fig. 4 shows the SHAP summary plots for the High Arousal class for the DEAP dataset. From the SHAP values, High Arousal was characterized by a mixture of medium-to-high feature values from a multitude of channels, including the gamma-band power from the electrodes P3, Fz, Cp2, and Oz, eye-blink rate, and the standard deviation of heart rate and heart rate variability.

Fig. 5 shows the most significant feature dependencies. Normalized gamma-band power at 0.3 or higher showed good contribution towards Arousal classification, likewise with above-average vEOG eye-blink rate. Significant SHAP was observed for normalized gamma-band power 0.3 or higher at the P3 and Fz electrodes intersected with medium-to-high standard deviation of heart rate from the top and bottom subfigures amplitude. From the center subfigure, above-average eye-blink rate was indicative of High Arousal. Simultaneously, the gamma-band power distribution at the Oz electrode was not concentrated at any specific regions.

Fig. 6 shows the SHAP summary plots for the positive Valence class for the DEAP dataset. Significant contributions included EEG beta-band power from FC6, Cz, and CP5 and EMG features mainly from the zygomaticus muscle, such as blink rate, amplitude variability, and energy.

The feature dependency plots in Fig. 7 showed some interesting trends. From the top subfigure, medium-to-high beta-band power at the FC6 electrode was indicative of Positive Valence. In addition, low values of complement-coded eye-blink rate measured from the zygomaticus muscle contributed more towards Positive Valence compared to medium-to-high values of the same. Likewise, high SHAP was observed in the center subfigure for low c_zEMG_rate_blink combined with medium-to-high values beta-band power at the P3 electrode. The bottom subfigure instead displayed the regular measurements of zEMG_rate_blink, where high SHAP was observed for high eye-blink rate.

2) DREAMER EXPLANATIONS

Fig. 8 shows the SHAP summary plots for Arousal classification of the DREAMER dataset. Significant features consisted mainly of ECG features, including amplitude and waveform features (ecgRampl_min), and HRV spectral

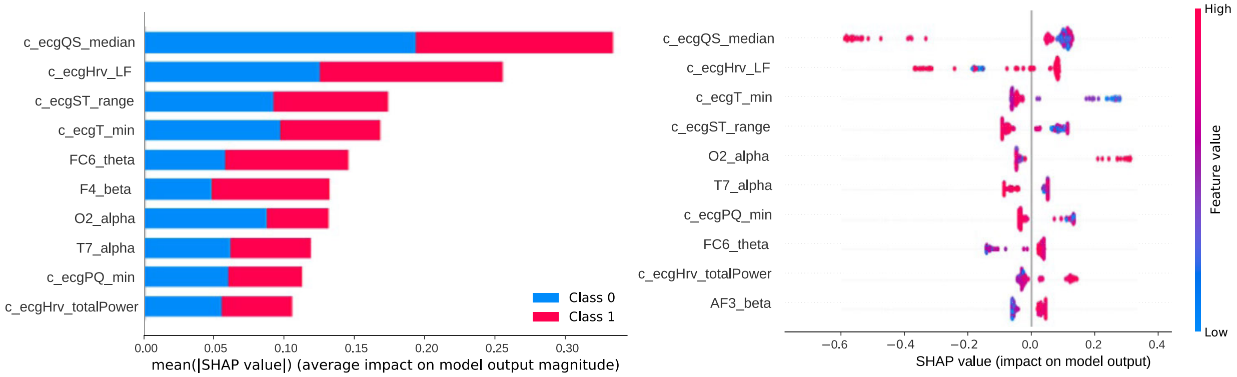

FIGURE 10. SHAP summary plots for Valence classes (left) and Positive Valence class (right) for DREAMER dataset.

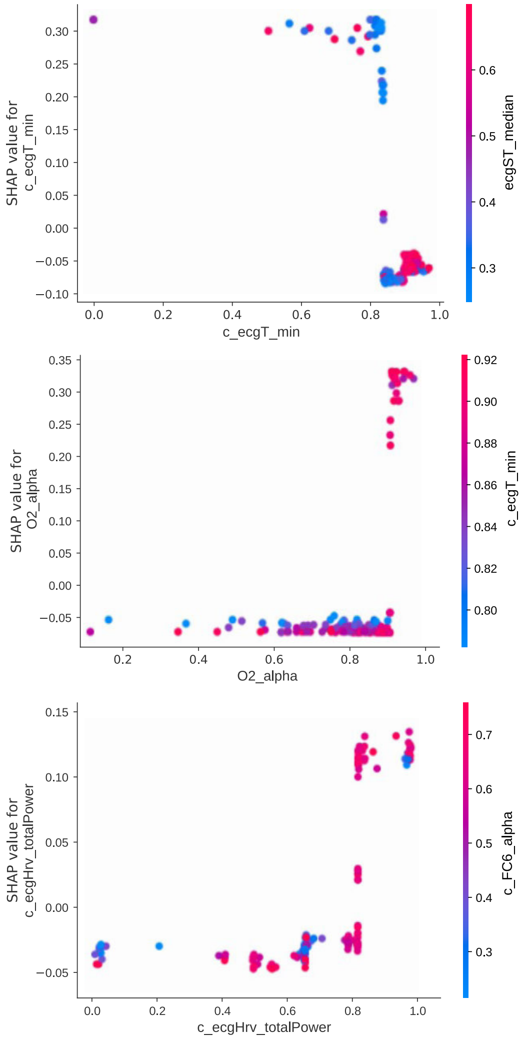

FIGURE 11. Significant feature dependencies for DREAMER Valence classification.

power features (c_ecgHrv_specRange3, ecgHrv_totalPower, c_ecgHRV_LF).

Fig. 9 shows the three most significant feature interactions for Arousal classification of the DREAMER dataset. The top subfigure shows the interaction between the minimum amplitude of the ECG R-waveform and the complement-coded EEG beta-band power at the P8 electrode. High SHAP for High Arousal affect was observed for clusters with high ecgRampl_min and medium-to-high c_P8_beta. The center subfigure shows the interaction between the ECG T-waveform's maximum amplitude and the amplitude range for the ECG Q-waveform. High SHAP was observed for the intersection between average-to-high values for both features. The bottom subfigure shows the interaction between the signal power of the HRV signal and the complement-coded minimum amplitude of the ECG P-waveform. High SHAP occurred for high values of hrv_totalPower and extremely high values of c_ecgP_min.

From Fig. 10, Valence classification for the DREAMER dataset showed few features with significantly large contributions for classifying Positive Valence, mainly from c_ecgT_min and O2_alpha. Significant features consisted mainly of complement-coded ECG features and a few EEG band power features.

Fig. 11 shows the three most significant feature interactions for Valence classification of the DREAMER dataset. The top subfigure shows the interactions between the complement-coded minimum amplitude of the ECG T-waveform and the ECG S-and-T-waveform median amplitude. High SHAP values were observed mainly for low-to-medium values of c_ecgT_min and mostly low values and some high values of ecgST_-median. The center subfigure shows the interaction between the EEG alpha-band power at the O2 electrode and the complement-coded minimum amplitude of the ECG T-waveform. The clusters' positioning showed that Positive Valence was characterized by extremely high values of both O2_alpha and c_ecgT_min. The bottom subfigure shows the interaction between the complement-coded features for both HRV total band power and the EEG alpha-band power at the Fc6 electrode, where significant contribution towards Valence classification was observed for high values of both features.

3) AMIGOS EXPLANATIONS

Fig. 12 shows the SHAP summary plots for Arousal classification of the AMIGOS dataset. Significant contributions

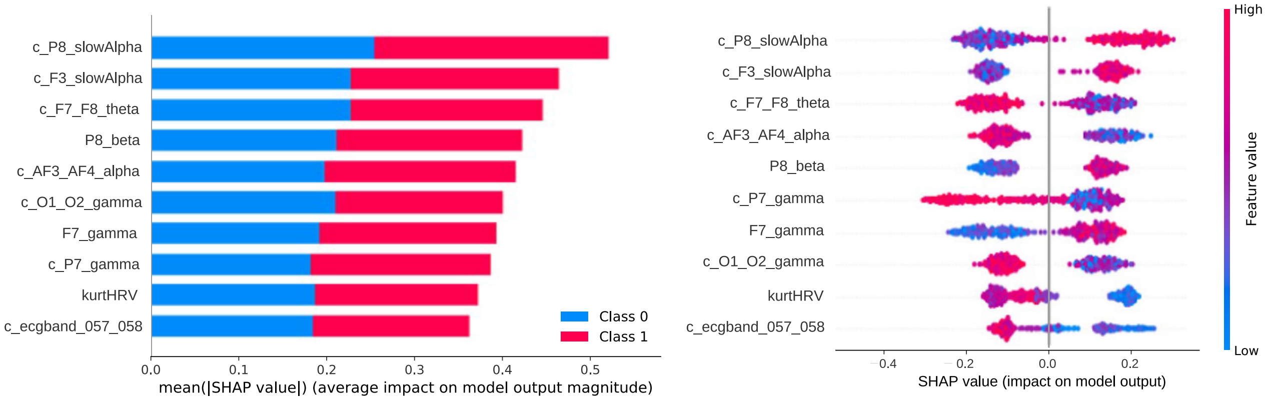

FIGURE 12. SHAP summary plots for Arousal classes (left) and High Arousal class (right) for AMIGOS dataset.

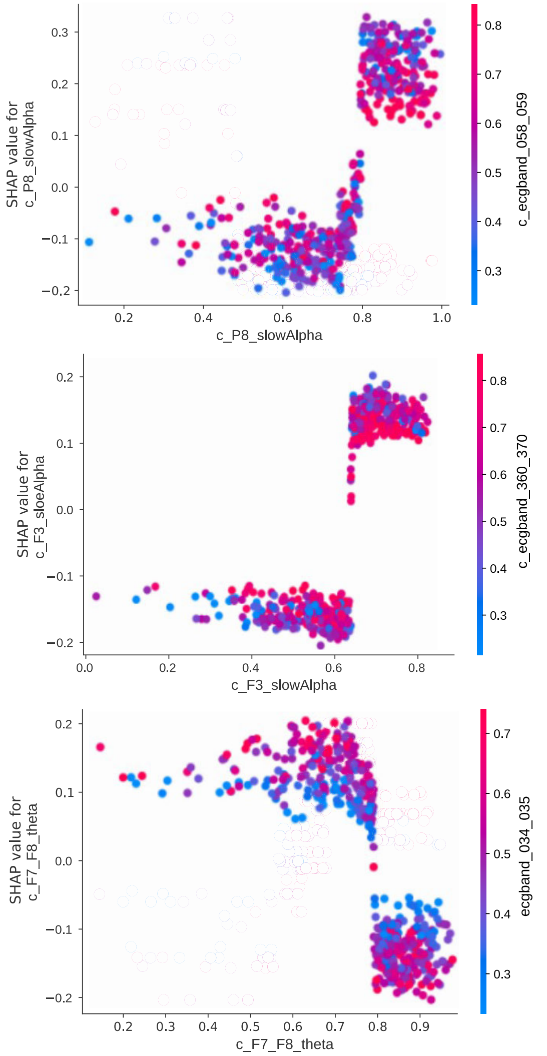

FIGURE 13. Significant feature dependencies for AMIGOS Arousal classification.

were observed mainly from complement-coded features, particularly from EEG gamma-band and slow-alpha-band power, and from EEG band power asymmetry between the left and right brain hemispheres (F7_F8, AF3_AF4, O1_O2). Low values of the complement-coded ECG band power between the frequencies 5.7Hz and 5.8Hz and the kurtosis measure of the HRV signal were both indicators of High Arousal.

Fig. 13 shows the three most significant feature interactions for Arousal classification. The top subfigure shows that a significant contributor comes from the complement-coded EEG slow-alpha-band power from the P8 electrode with high normalized value (x-axis, > 0.7), in conjunction with primarily medium-to-high values of the complement-coded ECG band power between 5.8Hz to 5.9Hz. The center subfigure showed a similar pattern with c_f3_slowAlpha (> 0.7) and c_ecgBand_3.6_3.7. From the bottom subfigure, however, the low complement-coded asymmetry between EEG band power at F7 and F8 (x-axis < 0.8) was a contributor towards Arousal classification.

Fig. 14 shows the SHAP summary plots for Valence classification of the AMIGOS dataset. Individual contributions from features were relatively low (< 0.3), mainly centering around EEG band power in the frontal regions of the brain (AF and F electrode positions) and asymmetry between the left and right hemispheres (T7_T8, Fc5_Fc6, F7_F8). The complement-coded mean heart rate amplitude showed the largest SHAP contribution towards Valence classification.

Fig. 15 shows the three most significant feature interactions for Valence classification. In the top subfigure, the EEG gamma-band power at the AF4 position showed a clear delineation, where values above 0.4 were significant contributions for Valence classification in conjunction with the EEG alpha-band power at the T8 position. Similarly, for the center subfigure, the EEG beta-band power asymmetry between T7 and T8 electrodes showed a clear contribution to Valence classification for normalized values 0.25 or higher, intersecting with the ECG band power between 3.6Hz and 3.7Hz. In the bottom subfigure, EEG theta-band power at AF3 displayed a slightly lower distinction. Values above 0.45 showed good contribution towards Valence classification, in conjunction with the complement-coded EEG beta-band power at O2.

The results from the three datasets showed some commonalities. Significant contributors for Arousal classification in the DEAP dataset consisted mostly of EEG gamma-band power from electrode positions in a distributed area and variability of the heart rate features. Arousal

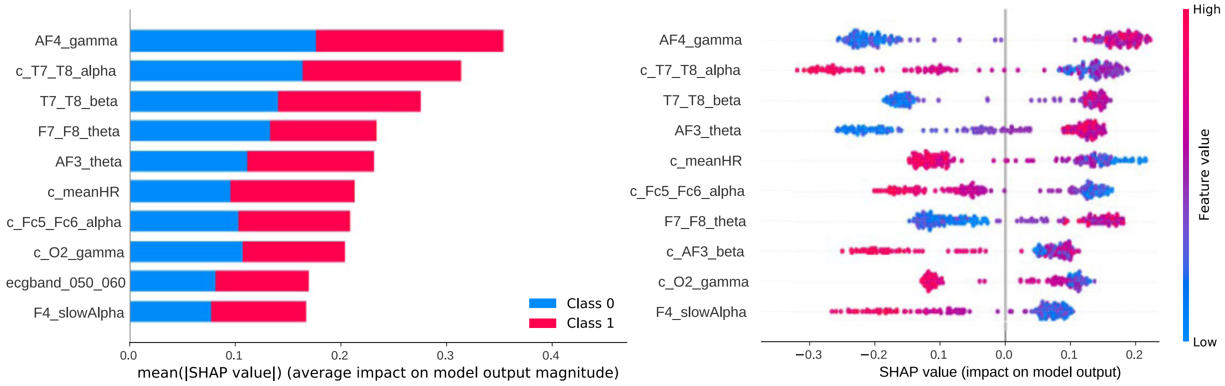

FIGURE 14. SHAP summary plots for Valence classes (left) and Positive Valence class (right) for AMIGOS dataset.

classification for the AMIGOS dataset also displayed significant contributions from EEG gamma and slow-alpha-band power features and EEG band power asymmetry. However, the DREAMER dataset showed fewer significant EEG contributors, the majority being ECG waveform and band power features.

For the Valence classification, the main contributors were typically EEG beta-band power features, eye-blink-related features, and some GSR and EMG signal energy from the DEAP dataset. For the AMIGOS dataset, EEG played a larger role with significant contributions from the frontal electrode positions' band power and from the EEG band power asymmetry between the left and right hemispheres. However, in the DREAMER dataset, SHAP contributions were low (< 0.2), where only one ECG and one EEG feature yielded high SHAP values.

4) DECISION TREES

Smaller subsets of the samples were generated by selecting features with higher-than-average SHAP scores, which were then used for training XGBoost models. The figures below represent a small sample of the generated XGBoost decision trees. Due to space constraints, not all DTs can be shown here. DTs were visualized using the *dtreeviz* package in Python. Nodes were represented using scatter plots to visualize the distribution of feature values vs class labels, each dot representing a single data sample. Vertical lines on the plots indicate the split of the feature values to create a decision boundary. Horizontal lines represent the average target value at either side of the feature split value. Leaf nodes at the rightmost column or bottom row of the figures indicate the average target prediction values and the number of samples that arrived at that point. Leaf nodes with target scores closer to 0 represent either Low Arousal or Negative Valence classes, while target scores closer to 1 represent either High Arousal or Positive Valence classes. Leaf nodes with average target values close to 0.5 represent an ambiguous decision. The n-score quantifies the number of data samples represented by the decision branches leading up to that node. A low number relative to the number of data samples indicates overfitting and outlier decisions, while a high number shows good generalization.

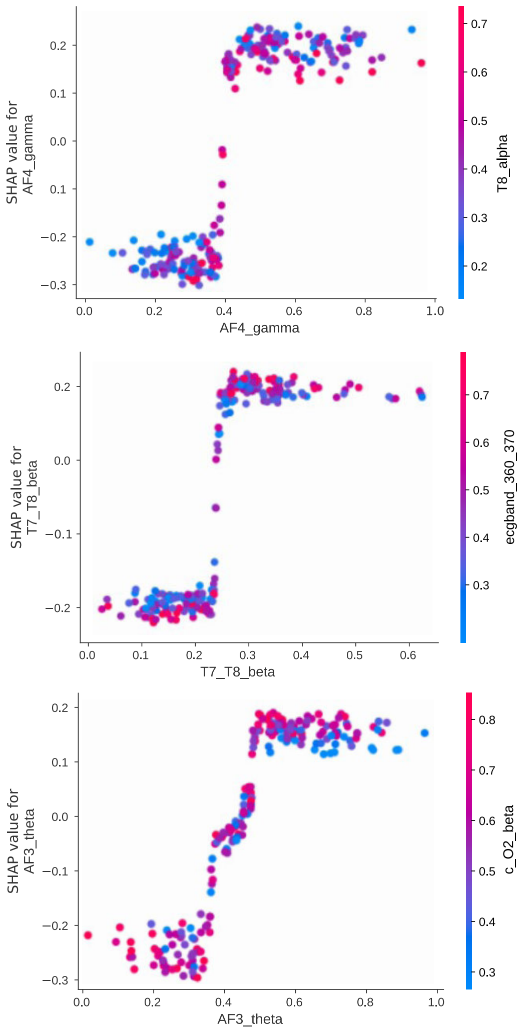

FIGURE 15. Significant feature dependencies for AMIGOS Valence classification.

Fig. 16 shows one of the DTs for Arousal classification of DEAP. The branching nodes consisted of the complement-coded respiration band power between 1.92Hz and 2.16Hz, EEG gamma-band power at the Oz electrode position, and the standard deviation of heart rate variability. Of the four leaf nodes, two nodes were approximated as Low

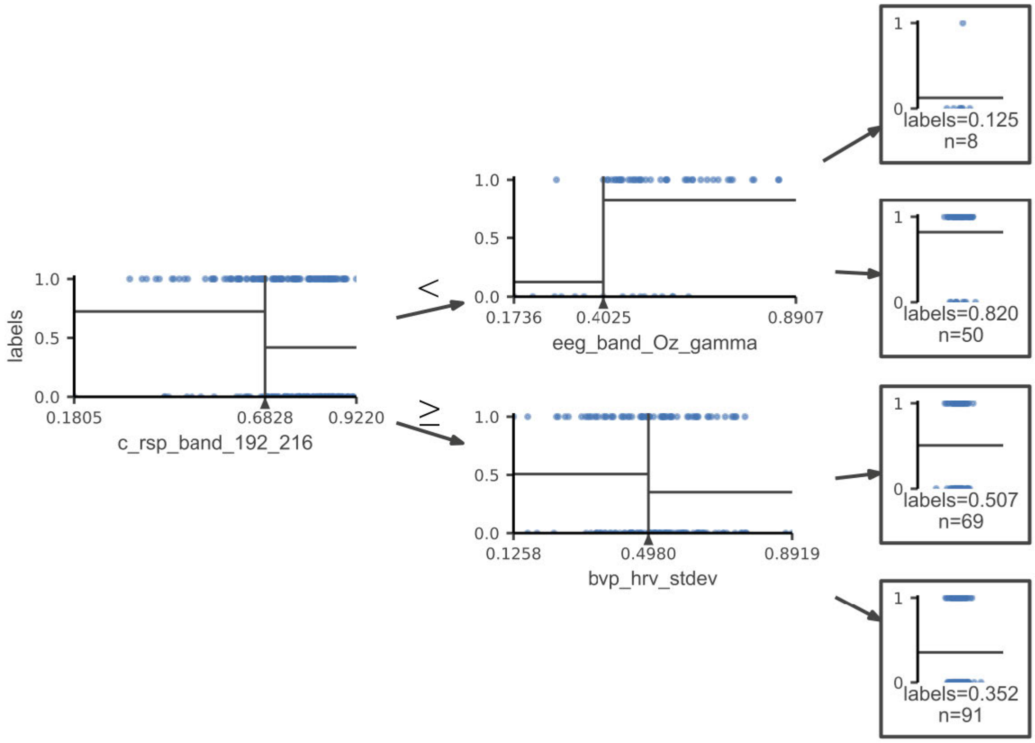

FIGURE 16. One XGBoost DT for Arousal classification of DEAP, after feature selection.

Arousal by having average target scores of 0.125 and 0.352. The first node was defined as having low feature values for the respiration and EEG features, and the second node having high respiration band power and HRV standard deviation. The leaf node with the target score of 0.852 was considered High Arousal, characterized as having low respiration feature value and high EEG band power. The leaf node with the target score of 0.507 was considered ambiguous, neither tending towards Low nor High Arousal.

The DTs provide a good visualization of the decision branches leading toward the predicted affect classes. The quality of the predictions is provided using leaf purity and class distribution metrics. Reliable and consistent predictions (i.e., leaf node scores close to 0 or 1 and high *n*-scores) are considered good candidates to define affect rule antecedents. Taking the second leaf node in Fig. 16 as an example: the distribution of labels is relatively concentrated towards one class label and applies to approximately 29% of the samples across all leaf nodes. Therefore, a good affect rule can be crafted from the decision branches leading up to it. While the individual trees presented here are relatively shallow, the XGBoosted ensembles as a whole showed good accuracy in predicting the affect classes as reported in Table 3.

VI. CONCLUSION

A methodology was proposed to provide an explainable affect recognition model. Noisy datasets were clustered into more representative exemplars using a combination of signal-to-noise amplification, synthetic minority sample generation, and Fuzzy ART clustering. The clustering process was optimized using genetic algorithms for hyper-parameter tuning. The clustered samples were used for training an ensemble of boosted decision trees and subsequently analyzed using SHAP scores. Physiological features with above-average scores were considered important for affect classification and were extracted to enhance the affect predictive ability of the decision tree ensemble. The performance of the final ensemble was on par with state of the art methods for classifying affect using the same datasets. The impact of features and their interactions with each other were easily visualized using a combination of SHAP scores and decision trees, providing interpretable and useful information on the physiological features relative to human affect.

REFERENCES

[1] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, ''Mastering chess and shogi by self-play with a general reinforcement learning algorithm,'' 2017, *arXiv:1712.01815*. [Online]. Available: http://arxiv.org/abs/1712.01815

[2] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, ''A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis,'' *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019.

[3] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, ''One-shot learning with memory-augmented neural networks,'' 2016, *arXiv:1605.06065*. [Online]. Available: http://arxiv.org/abs/1605.06065

[4] B. Goodman and S. Flaxman, ''European union regulations on algorithmic decision-making and a 'right to explanation,''' *AI Mag.*, vol. 38, no. 3, pp. 50–57, Oct. 2017.

[5] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, ''Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation,'' in *Proc. IEEE Conf. Comput. Intell. Games (CIG)*, Aug. 2018, pp. 1–8.

[6] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 0210–0215.

[7] O. Cordón, "A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems," *Int. J. Approx. Reasoning*, vol. 52, no. 6, pp. 894–913, Sep. 2011.

[8] A. C. Weidman, C. M. Steckler, and J. L. Tracy, "The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research," *Emotion*, vol. 17, no. 2, p. 267, 2017.

[9] V. Kurbalija, M. Ivanović, M. Radovanović, Z. Geler, W. Dai, and W. Zhao, "Emotion perception and recognition: An exploration of cultural differences and similarities," *Cognit. Syst. Res.*, vol. 52, pp. 103–116, Dec. 2018.

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[11] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost Off-the-Shelf devices," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[12] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, early access, Nov. 30, 2018, doi: 10.1109/TAFFC.2018.2884461.

[13] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.

[14] E. Arce, A. N. Simmons, M. B. Stein, P. Winkielman, C. Hitchcock, and M. P. Paulus, "Association between individual differences in self-reported emotional resilience and the affective perception of neutral faces," *J. Affect. Disorders*, vol. 114, nos. 1–3, pp. 286–293, Apr. 2009.

[15] J. A. Coan and J. J. Allen, *Handbook of Emotion Elicitation and Assessment*. Oxford, U.K.: Oxford Univ. Press, 2007.

[16] J. Preethi, M. Sreeshakthy, and A. Dhilipan, "A survey on EEG based emotion analysis using various feature extraction techniques," *Int. J. Sci., Eng. Technol. Res.*, vol. 3, no. 11, pp. 3113–3120, 2014.

[17] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014.

[18] Q. Wei, Y. Zhao, Q. Xu, L. Li, J. He, L. Yu, and B. Sun, "A new deep-learning framework for group emotion recognition," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 587–592.

[19] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2, pp. 1–23, Jun. 2019.

[20] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien, "FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 17–25.

[21] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals," in *Proc. Interspeech*, Sep. 2018, pp. 162–166.

[22] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[23] S. Vaid, P. Singh, and C. Kaur, "EEG signal analysis for BCI interface: A review," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Technol.*, Feb. 2015, pp. 143–147.

[24] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion classification based on biophysical signals and machine learning techniques," *Symmetry*, vol. 12, no. 1, p. 21, Dec. 2019.

[25] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2069–2072.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[27] H.-C. Yang and C.-C. Lee, "Annotation matters: A comprehensive study on recognizing intended, self-reported, and observed emotion labels using physiology," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 1–7.

[28] F. Mokdad, D. Bouchaffra, N. Zerrouki, and A. Touazi, "Determination of an optimal feature selection method based on maximum shapley value," in *Proc. 15th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, Dec. 2015, pp. 116–121.

[29] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," 2020, *arXiv:2002.03898*. [Online]. Available: http://arxiv.org/abs/2002.03898

[30] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based human emotion recognition and visualization," in *Proc. Int. Conf. Cyberworlds*, Oct. 2010, pp. 262–269.

[31] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 1, pp. 275–281, Jan. 2015.

[32] P. Masulli, F. Masulli, S. Rovetta, A. Lintas, and A. E. P. Villa, "Fuzzy clustering for exploratory analysis of EEG event-related potentials," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 1, pp. 28–38, Jan. 2020.

[33] R. M. Ghoniem, A. D. Algarni, and K. Shaalan, "Multi-modal emotion aware system based on fusion of speech and brain information," *Information*, vol. 10, no. 7, p. 239, Jul. 2019.

[34] I. Škrjanc, "Cluster-volume-based merging approach for incrementally evolving fuzzy Gaussian clustering—eGAUSS+," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 9, pp. 2222–2231, Sep. 2020.

[35] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 698–713, Sep. 1992.

[36] R. Palaniappan, P. Raveendran, S. Nishida, and N. Saiwaki, "Fuzzy ATRMAP classification of mental tasks using segmented and overlapped EEG signals," in *Proc. TENCON Intell. Syst. Technol. New Millennium*, vol. 2, Sep. 2000, pp. 388–391.

[37] A. Jafarifarmand, M. A. Badamchizadeh, S. Khanmohammadi, M. A. Nazari, and B. M. Tazehkand, "A new self-regulated neuro-fuzzy framework for classification of EEG signals in motor imagery BCI," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1485–1497, Jun. 2018.

[38] C. M. Vineyard, S. J. Verzi, M. L. Bernard, S. E. Taylor, I. Dubicka, and T. P. Caudell, "A multi-modal network architecture for knowledge discovery," *Secur. Informat.*, vol. 1, no. 1, pp. 1–12, Dec. 2012.

[39] C. K. Loo, W. S. Liew, and M. S. Sayeed, "Genetic ensemble biased ARTMAP method of ECG-based emotion classification," in *Intelligent Interactive Multimedia: Systems and Services* (Smart Innovation, Systems and Technologies), vol. 14, T. Watanabe, J. Watada, N. Takahash, R. Howlett, and L. Jain, Eds. Berlin, Germany: Springer, 2012, pp. 299–306, doi: 10.1007/978-3-642-29934-6_29.

[40] M. Yaghini and M. A. Shadmani, "GOFAM: A hybrid neural network classifier combining fuzzy ARTMAP and genetic algorithm," *Artif. Intell. Rev.*, vol. 39, no. 3, pp. 183–193, Mar. 2013.

[41] W. S. Liew, M. Seera, C. K. Loo, and E. Lim, "Affect classification using genetic-optimized ensembles of fuzzy ARTMAPs," *Appl. Soft Comput.*, vol. 27, pp. 53–63, Feb. 2015.

[42] P. Hänggi, "Stochastic resonance in biology how noise can enhance detection of weak signals and help improve biological information processing," *Chem. Phys. Chem.*, vol. 3, no. 3, pp. 285–290, Mar. 2002.

[43] Y. Lin, B. Liu, Z. Liu, and X. Gao, "EEG gamma-band activity during audiovisual speech comprehension in different noise environments," *Cognit. Neurodynamics*, vol. 9, no. 4, pp. 389–398, Aug. 2015.

[44] L.-F. Rebolledo-Herrera and F. G. Espinosa, "Novel parameter tuned methodology for under-damped stochastic resonance applied to EEG signal enhancement," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, pp. 002128–002132.

[45] I. Mendez-Balbuena, P. Arrieta, N. Huidobro, A. Flores, R. Lemuz-Lopez, C. Trenado, and E. Manjarrez, "Augmenting EEG-global-coherence with auditory and visual noise: Multisensory internal stochastic resonance," *Medicine*, vol. 97, no. 35, 2018, Art. no. e12008.

[46] O. Osoba and B. Kosko, "Noise-enhanced clustering and competitive learning algorithms," *Neural Netw.*, vol. 37, pp. 132–140, Jan. 2013.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[48] J. Vanhoeyveld and D. Martens, "Imbalanced classification in sparse and large behaviour datasets," *Data Mining Knowl. Discovery*, vol. 32, no. 1, pp. 25–82, Jan. 2018.

[49] P. Kaur and A. Gosain, "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise," in *ICT Based Innovations* (Advances in Intelligent Systems and Computing), vol. 653, A. Saini, A. Nayak, and R. Vyas, Eds. Singapore: Springer, 2018, pp. 23–30, doi: 10.1007/978-981-10-6602-3_3.

[50] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, Jan. 2017.

[51] J. Bang, T. Hur, D. Kim, T. Huynh-The, J. Lee, Y. Han, O. Banos, J.-I. Kim, and S. Lee, "Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments," *Sensors*, vol. 18, no. 11, p. 3744, Nov. 2018.

[52] N. Henderson, J. Rowe, L. Paquette, R. S. Baker, and J. Lester, "Improving affect detection in game-based learning with multimodal data fusion," in *Artificial Intelligence in Education* (Lecture Notes in Computer Science), vol. 12163, I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Cham, Switzerland: Springer, 2020, pp. 228–239, doi: 10.1007/978-3-030-52237-7_19.

[53] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci.*, vol. 505, pp. 32–64, Dec. 2019.

[54] Q. Gu, X.-M. Wang, Z. Wu, B. Ning, and C.-S. Xin, "An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification," *J. Digit. Inf. Manage.*, vol. 14, no. 2, pp. 92–103, 2016.

[55] K. Jiang, J. Lu, and K. Xia, "A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE," *Arabian J. Sci. Eng.*, vol. 41, no. 8, pp. 3255–3266, Aug. 2016.

[56] J. Wang, Q. Zhang, and G. Xu, "Genetic stochastic resonance: A new fault diagnosis method to detect weak signals in mechanical systems," *Adv. Sci. Lett.*, vol. 4, no. 6, pp. 2508–2512, Jul. 2011.

[57] Y. Zheng, M. Huang, Y. Lu, and W. Li, "Fractional stochastic resonance multi-parameter adaptive optimization algorithm based on genetic algorithm," *Neural Comput. Appl.*, vol. 32, pp. 16807–16818, Nov. 2020, doi: 10.1007/s00521-018-3910-6.

[58] P. Vasuki, "Speech emotion recognition using adaptive ensemble of class specific classifiers," *Res. J. Appl. Sci., Eng. Technol.*, vol. 9, no. 12, pp. 1105–1114, Apr. 2015.

[59] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[60] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. Int. Conf. Artif. Intell.*, vol. 56, 2000, pp. 1–7.

[61] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization," *Frontiers Neurorobotics*, vol. 12, p. 78, Nov. 2018.

[62] M. D. J. Tran, C. P. Lim, C. Abeynayake, and L. C. Jain, "Feature extraction and classification of metal detector signals using the wavelet transform and the fuzzy ARTMAP neural network," *J. Intell. Fuzzy Syst.*, vol. 21, no. 1, 2, pp. 89–99, 2010.

[63] C. K. Loo, W. S. Liew, M. Seera, and E. Lim, "Probabilistic ensemble fuzzy ARTMAP optimization using hierarchical parallel genetic algorithms," *Neural Comput. Appl.*, vol. 26, no. 2, pp. 263–276, Feb. 2015.

[64] B. McGinley, J. Maher, C. O'Riordan, and F. Morgan, "Maintaining healthy population diversity using adaptive crossover, mutation, and selection," *IEEE Trans. Evol. Comput.*, vol. 15, no. 5, pp. 692–714, Oct. 2011.

[65] M. N. Haque, N. Noman, R. Berretta, and P. Moscato, "Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0146116.

[66] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 2522–5839, 2020.

[67] B. Vigdor and B. Lerner, "The Bayesian ARTMAP," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1628–1644, Nov. 2007.

[68] H. A. Kakudi, C. K. Loo, F. M. Moy, N. Masuyama, and K. Pasupa, "Diagnosing metabolic syndrome using genetically optimised Bayesian ARTMAP," *IEEE Access*, vol. 7, pp. 8437–8453, 2019.

[69] P. Nooralishahi, C. K. Loo, and M. Seera, "Semi-supervised topo-Bayesian ARTMAP for noisy data," *Appl. Soft Comput.*, vol. 62, pp. 134–147, Jan. 2018.

[70] E. A. Lisangan, A. Musdholifah, and S. Hartati, "Two level clustering for quality improvement using fuzzy subtractive clustering and self-organizing map," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 15, no. 2, pp. 373–380, 2015.

[71] N. Masuyama, C. K. Loo, H. Ishibuchi, N. Kubota, Y. Nojima, and Y. Liu, "Topological clustering via adaptive resonance theory with information theoretic learning," *IEEE Access*, vol. 7, pp. 76920–76936, 2019.

[72] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, Jan. 1973.

[73] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

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

[75] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[76] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[77] Y. Shu and S. Wang, "Emotion recognition through integrating EEG and peripheral signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2871–2875.

[78] H.-C. Yang and C.-C. Lee, "An attribute-invariant variational learning for emotion recognition using physiology," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1184–1188.

WEI SHIUNG LIEW was born in Kuala Lumpur, Malaysia, in 1986. He received the B.S. degree in electronics engineering, majoring in robotics and automation from Multimedia University, Malaysia, in 2010, and the M.S. degree in biomedical engineering from the University of Malaya, in 2015. From 2009 to 2011, he was a Research Assistant with the Faculty of Information Science and Technology, Multimedia University. Since then, he has been a Postgraduate Research Assistant with the Department of Artificial Intelligence, University of Malaya. His research interests include evolutionary algorithms and biologically-inspired artificial intelligence methods.

CHU KIONG LOO (Member, IEEE) received the B.Eng. degree (Hons.) in mechanical engineering from the University of Malaya, in 1996, and the Ph.D. degree from Universiti Sains Malaysia, in 2004, specializing in neurorobotics. He is currently a Full Professor with the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya. His research interest includes neuroscience-inspired machine intelligence. He was the IEEE Systems, Man and Cybernetics (SMC) Society Vice-Chairman for Malaysia Chapter, from 2013 to 2014. He was also the President of the Asia Pacific Neural Network Assembly (APNNA), in 2014. He was a recipient of the Georg Forster Research Fellowship for Experienced Researchers from the Alexander von Humboldt-Foundation, Germany.

STEFAN WERMTER (Member, IEEE) is currently a Full Professor with the University of Hamburg, Germany, and the Director of the Department of Informatics, Knowledge Technology Institute. He has previously held positions at the University of Dortmund, the University of Massachusetts, the International Computer Science Institute in Berkeley, and the University of Sunderland. His research interests include neural networks, hybrid knowledge technology, neuroscience-inspired computing, cognitive robotics, and human–robot interaction. He is the Co-Coordinator of the International Collaborative Research Center on Crossmodal Learning (TRR-169) and the Coordinator of the European Training Network SECURE on safety for cognitive robots. In 2019, he was elected and also serves as the President for the European Neural Network Society. He has been an Associate Editor of the journal IEEE Transactions on Neural Networks and Learning Systems. He is also an Associate Editor of *Connection Science* and *International Journal for Hybrid Intelligent Systems*. He is on the editorial board of the journals *Cognitive Computation* and *Journal of Computational Intelligence*.