IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 July 2023, accepted 15 September 2023, date of publication 25 September 2023, date of current version 9 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318751

RESEARCH ARTICLE

FF-BTP Model for Novel Sound-Based Community Emotion Detection

ARIF METEHAN YILDIZ[1], MASAYUKI TANABE[2], (Member, IEEE), MAKIKO KOBAYASHI[2,3], (Senior Member, IEEE), ILKNUR TUNCER[4], PRABAL DATTA BARUA[5], SENGUL DOGAN[1], TURKER TUNCER[1], RU-SAN TAN[6,7], AND U. RAJENDRA ACHARYA[3,8]

[1]Department of Digital Forensics Engineering, Technology Faculty, Firat University, 23119 Elazig, Turkey
[2]Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto 860-8555, Japan
[3]International Research Organization for Advanced Science and Technology (IROAST), Kumamoto University, Kumamoto 860-8555, Japan
[4]Elazig Governorship, Interior Ministry, 23119 Elazig, Turkey
[5]School of Business (Information System), University of Southern Queensland, Springfield, QLD 4350, Australia
[6]Department of Cardiology, National Heart Centre Singapore, Singapore 169609
[7]Duke-NUS Medical School, Singapore 169857
[8]School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD 4350, Australia

Corresponding author: Sengul Dogan (sdogan@firat.edu.tr)

ABSTRACT Most emotion classification schemes to date have concentrated on individual inputs rather than crowd-level signals. In addressing this gap, we introduce Sound-based Community Emotion Recognition (SCED) as a fresh challenge in the machine learning domain. In this pursuit, we crafted the FF-BTP-based feature engineering model inspired by deep learning principles, specifically designed for discerning crowd sentiments. Our unique dataset was derived from 187 YouTube videos, summing up to 2733 segments each of 3 seconds (sampled at 44.1 KHz). These segments, capturing overlapping speech, ambient sounds, and more, were meticulously categorized into negative, neutral, and positive emotional content. Our architectural design fuses the BTP, a textural feature extractor, and an innovative handcrafted feature selector inspired by Hinton's FF algorithm. This combination identifies the most salient feature vector using calculated mean square error. Further enhancements include the incorporation of a multilevel discrete wavelet transform for spatial and frequency domain feature extraction, and a sophisticated iterative neighborhood component analysis for feature selection, eventually employing a support vector machine for classification. On testing, our FF-BTP model showcased an impressive 97.22% classification accuracy across three categories using the SCED dataset. This handcrafted approach, although inspired by deep learning's feature analysis depth, requires significantly lower computational resources and still delivers outstanding results. It holds promise for future SCED-centric applications.

INDEX TERMS FF-BTP, sound community emotion classification, sound processing, textural feature extraction.

I. INTRODUCTION

A. BACKGROUND

When humans engage in communication, verbal expressions are accompanied by emotional expressions [1], [2]. Emotion, a complex amalgamation of biological and psychological processes, is a conscious experience that fundamentally characterizes the human persona. To convey our internal states effectively, we employ a range of tools, including facial expressions, body language, and speech [3]. Our communication skills, decision-making abilities, and social interactions are profoundly influenced by our emotional states [4]. The individual's internal emotional state becomes more discernible through personal experiences, physiological conditions, and environmental stimuli. Hence, valuable insights into a person's emotional state can be gleaned

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

through the utilization of emotion recognition techniques [5], [6]. Particularly within the realm of human-computer interaction, the detection of individuals' emotional states assumes tremendous significance [7]. Consequently, the classification of emotional states using physiological signals is a hot research topic for the 2020s [8].

In the context of human-computer interaction, emotion recognition systems play a crucial role in comprehending users' emotional states and providing enhanced services. Emotion recognition finds diverse applications across various domains [3], [9]. It can be effectively employed in social media analysis, market research, and emotion tracking [10], [11]. Further, it has been used within the healthcare sector to detect emotional disorders, including depression and anxiety [12].

Community emotion analysis is an important topic for law enforcement since they can detect dangerous situations and it can be used to understand the emotional state of a community. Community emotion analysis interprets the emotional state of users from visual and auditory cues such as facial expressions, tone of voice, and body language, especially on social media and other platforms [13], [14]. For this purpose, different methods have been developed using artificial intelligence and machine learning techniques [15], [16], [17], [18]. This study presents a study on sound-based community emotion recognition (SCED).

B. RELATED WORKS

We searched the literature for SCED but there is limited research. Most of the reviewed research studies focused on speech and textual emotion recognition in individuals, which are summarized in Table 1.

C. LITERATURE GAPS

The following literature gaps have been identified:

- Existing speech emotion datasets mostly pertained to specific individuals. There is a dearth of environmental or crowd-sound datasets for SCED.
- Deep learning models often attain high classification performance but are limited by high computational complexity, as they typically require training of millions of parameters.
- To address this challenge, it is imperative to develop a highly accurate and efficient feature engineering model.

The model is structured in three phases: (1) multilevel FF-BTP-based feature extraction (as detailed in Section III); (2) feature selection employing INCA; and (3) classification using a SVM. Our aim, by leveraging lightweight handcrafted feature engineering techniques, was to emulate the notable classification performance of computationally demanding deep learning models. We achieved this by augmenting feature representation through MDWT-based signal decomposition, enabling multilevel feature generation in both spatial and frequency domains, and complementing it with BTP-based multi-textural feature vector extraction.

TABLE 1. State-of-the-art on text- and speech-based emotion recognition.

Study	Aim	Method/Classifier	Highlights (Data, split ratio, results (%), limitation)
Ghosh et al. [19]	Sentiment detection and emotion recognition	CNN, Softmax	- Texts from social media - 70:15:15 - 71.30 accuracy for sentiment - 66.03 accuracy for emotion - High complexity
Zhang et al. [20]	Sarcasm, sentiment and emotion recognition in conversations	Multitask learning model, SVM	- Text, visual, and acoustic data - 5-fold CV Sarcasm - 73.31 accuracy for MUStARD*ext* dataset - 61.39 accuracy for Memotion dataset Sentiment - 53.40 accuracy for MUStARD*ext* dataset - 36.21 accuracy for Memotion dataset - 60.77 accuracy for the CMU-MOSEI dataset - 66.01 accuracies for the MELD dataset Emotion - 35.27 accuracy for MUStARD*ext* dataset - 49.42 accuracy for Memotion dataset - 78.55 accuracy for the CMU-MOSEI dataset - 40.16 accuracy for the MELD dataset - High complexity
Yang et al. [21]	Emotion recognition of posts in the online student community	Adaptive multi-view selection, base classifiers	- Texts from social media - 5-fold CV - 93.99 accuracy - Small data
Atmaja and Sasou [22]	Sentiment analysis and emotion recognition	UniSpeech-SAT models	- Speech data - Unspecified split ratio - 81.36 accuracy - Small data
Kaur et al. [23]	Sentiment detection	Term frequency-inverse document frequency, XGBoost	- Emoticons - 90:10 - 87.84 accuracy
Pragati et al. [24]	Speech emotion recognition	Mel frequency cepstral coefficients, multi-layer perceptron	- Speech data - 80:20 - 82.84 accuracy - Small data
Bashir et al. [25]	Emotion detection	Long short-term memory, count vectorization, deep neural network	- Texts - 80:20 - 85.30 accuracy - High complexity
Widodo et al. [26]	Sentiment analysis	Term frequency-inverse document frequency, naïve Bayes	- Texts from social media - Unspecified split ratio - 85.40 accuracy - Small data

TABLE 1. *(Continued.)* **State-of-the-art on text- and speech-based emotion recognition.**

Ali et al. [27]	Social media content classification, community detection	Girvan-Newman graph algorithm, long short-term memory, gated recurrent units	- Texts from social media - 70:30 - 98.14 accuracy - High complexity
Yao et al. [28]	Emotion recognition	Bidirectional encoder representations from transformers, multi-layer perceptron	- Texts - 80:20 - 98.65 accuracy
Wang et al. [29]	Emotion recognition	Hierarchically stacked graph convolution, softmax	- Speech data - 85:15 - 65.13 accuracy - High complexity, low accuracy
Qin et al. [30]	Emotion recognition	Pretrained language models	- Texts, speech data - Unspecified split ratio - 71.70 accuracy for the IEMOCAP dataset - 67.11 accuracy for the MELD dataset - 61.42 accuracy for DailyDialog - 39.84 accuracy for EmoryNLP - Low accuracy
Gu et al. [31]	Emotion recognition	CNN, softmax	- Texts - 10-fold CV - 84.38 accuracy - High complexity
Utku et al. [32]	Emotion recognition	Graph CNN, softmax	- Texts from social media - 80:20 - 93.00 accuracy - High complexity
Arbane et al. [33]	Sentiment classification	Long short-term memory	- Texts from social media - Unspecified split ratio - 97.52 accuracy - High complexity

D. MOTIVATION AND THE PROPOSED MODEL

We were motivated to develop a computationally lightweight engineering model for SCED that could match the acknowledged high performance of deep learning models. In deep learning models, the input signals are dynamically trained using layers of hidden feedforward as well as backpropagation networks. Hinton's forward-forward (FF) algorithm [34], which only contains feedforward deep layers, arguably more realistically simulate human neural network and is computationally less expensive. Inspired by this, we proposed a new handcrafted feature extraction function that incorporated an FF feature vector selection algorithm based on identifying the vector with the maximum calculated mean square error (MSE) value. We coupled this FF algorithm with upstream (1) multilevel discrete wavelet transformation (MDWT) [35], which enabled multilevel feature extraction from the one-dimensional raw sound signal and wavelet bands in both spatial and frequency domains; and (2) a binary ternary pattern (BTP) [36] multi-textural feature generator using three mathematical operations. We termed our novel feature extraction function FF-BTP. Other elements of our model–feature selection and classification–were kept simple to minimize the time complexity.

We proposed to use our model to solve the problem of accurate and efficient classification of the prevailing emotion within a crowd or community, leveraging on environmental sounds. Hitherto, there has been no systematic approach for environmental SCED, and also no relevant publicly available research training dataset. Toward the latter end, we have assembled a tailored soundscape dataset specifically designed for SCED. In constructing this dataset, we excluded samples with prominent foreground sounds. Instead, we used overlapping speech or noise to discern emotional states. It is crucial to emphasize that the inclusion of speech from individuals or songs would render it counter to the primary purpose of the SCED paradigm.

E. NOVELTIES AND CONTRIBUTIONS

Innovations and contributions of this research are listed below.

Innovations:

- This study introduces the inaugural SCED dataset. Distinguishing this dataset is its distinctive assembly of overlapped auditory signals, facilitating an enriched spectrum of auditory analyses.
- Through methodological investigation, the FF algorithm, primarily employed for feature vector selection, was integrated with the BTP. This integration culminated in the novel FF-BTP feature extraction function, heralding an innovative paradigm in feature extraction methodologies.
- The proposed FF-BTP is a self-organized feature extraction function.
- This work elucidates the first feature engineering model specifically optimized for automated SCED. The proposed model provides potential pathways for refined emotion detection methodologies within community auditory signals.

Contributions:

- Emotion classification, where the real moods of the subject(s) are automatically classified using machine learning techniques [37], [38], [39], is a growing research field. Many emotion classification models in the literature use various input data: facial images, speech, biophysical signals (e.g., electroencephalography), or functional magnetic resonance images [17], [40], [41], [42]. To our knowledge, there has been little research on environmental SCED, which may be useful for assessing crowd sentiments in diverse applications, e.g., media, security, behavioral science, etc. We proposed SCED as a challenge and proceeded to

solve it using a new deep learning-inspired FF-BTP-based feature engineering model that was trained and developed on a unique new crowd sound dataset.

F. ORGANIZATION

The structure of the remainder of this paper is as follows: Section II details the collected SCED dataset. Section III introduces the proposed FF-BTP feature extraction. In Section IV, we present the FF-BTP sound classification model, outlining its phases and steps. Section V reports the classification performance of the proposed model on the SCED dataset. Discussions and analyses of the results are provided in Section VI. Conclusions are drawn in Section VII, followed by potential future works in Section VIII.

II. DATASET

In this study, we conducted a prospective search on YouTube for videos featuring crowds or communities to acquire sound recordings. Each recording was characterized by a dominant emotion that was categorized into one of three classes: negative, neutral, or positive emotion. To ascertain the accuracy of the emotional classification, we implemented a manual validation process that involved a three-person verification team. Only recordings where all three verifiers reached a unanimous decision on the emotional class were retained. We specifically selected overlapping sounds that lacked discernible individual utterances, singing and the like. Overall, we gathered 187 sound recordings, with a minimum of 60 recordings per emotion class. The total duration of these recordings was 17.2 hours, averaging 5.5 minutes per recording. These recordings were further divided into three-second segments and saved in the .wav format with a high-fidelity 44.1 KHz sampling frequency. The finalized SCED study dataset consisted of 2733 .wav files, distributed among the negative, neutral, and positive classes with 919, 905, and 909 files, respectively. Moreover, this dataset is publicly available on the Kaggle website and users/researchers/developers can download this dataset by using https://www.kaggle.com/datasets/arifmetehanyldz/sced-v1 URL.

III. THE PROPOSED FF-BTP FEATURE EXTRACTION FUNCTION

We designed a novel handcrafted textural feature extraction function that combined the established BTP multi-feature vector generator [36] with a feedforward final feature selection algorithm, which was inspired by Hinton's FF algorithm [34], used in deep learning (Figure 1). Hinton's FF algorithm is a novel machine-learning method that offers various advantages [34]. To our knowledge, there has been no published handcrafted feature extraction function based on Hinton's FF algorithm.

The novel use of Hinton's FF algorithm for final feature vector selection distinguishes FF-BTP from others. This algorithm provides an advanced method to ensure that the selected vector best distinguishes the signal from other classes of signals.

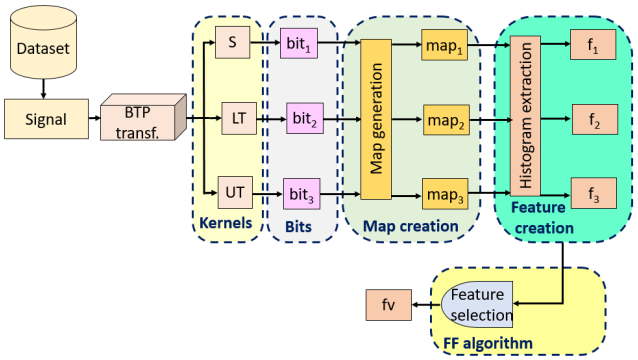

FIGURE 1. Schema of the FF-BTP feature extraction function. *BTP transf.: BTP transformation, S: signum function, LT: lower ternary, UT: upper ternary, f: feature vector, fv: the selected feature vector, FF: forward-forward.

FF-BTP combines the capabilities of BTP and FF, offering a powerful subtraction function that not only provides textural details via BTP but also enables optimal vector selection via FF.

The synthesis of BTP's comprehensive textural feature generation and FF's optimal feature vector selection enables FF-BTP to capture the nuances of the input signal while maintaining a feature representation that optimizes differentiation from other classes. This dual capability potentially increases the accuracy and efficiency of any system using FF-BTP for feature extraction.

BTP is a histogram-based textural feature extraction method that utilizes signum, upper ternary, and lower ternary kernels. These kernels produce three textural feature vectors for each input signal block. The specific steps for feature extraction are outlined below.

1: Load the dataset and read the signal.

2: Calculate primary signals for every class.

3: Create overlapping blocks of length 9.

$$b^t(g) = signal(t+g-1), t \in \{1, 2, \ldots, len-8\}, \quad g \in \{1, 2, \ldots, 9\} \tag{1}$$

where b represents overlapping block; *signal*; input signal; and *len*, signal length.

4: Generate binary features using signum, upper ternary and lower ternary functions.

$$bit_k^t(i) = \delta^k(b^t(i), b^t(5)), k \in \{1, 2, 3\}, \quad i \in \{1, 2, \ldots, 4\} \tag{2}$$

$$bit_k^t(i+4) = \delta^k(b^t(i+5), b^t(5)) \tag{3}$$

$$\delta^1\left(b^t(i), b^t(5)\right) = \begin{cases} 0, & b^t(i) - b^t(5) \geq -t \\ 1, & b^t(i) - b^t(5) < -t \end{cases} \tag{4}$$

$$\delta^2\left(b^t(i), b^t(5)\right) = \begin{cases} 0, & b^t(i) - b^t(5) < 0 \\ 1, & b^t(i) - b^t(5) > 0 \end{cases} \tag{5}$$

$$\delta^3\left(b^t(i), b^t(5)\right) = \begin{cases} 0, & b^t(i) - b^t(5) \leq t \\ 1, & b^t(i) - b^t(5) > t \end{cases} \quad (6)$$

$$t = \frac{SD(signal)}{2} \quad (7)$$

where δ^1 represents lower ternary function; δ^2, signum function; δ^3, upper ternary function; bit, extracted binary features; t, threshold value; and $SD(.)$, standard deviation function. Equations 2 to 7 in this step define the BTP feature extraction function mathematically.

5: Calculate feature maps using binary to decimal transformation.

$$map_k(t) = \sum_{h=1}^{8} bit_k^t(h) \cdot 2^{h-1} \quad (8)$$

where *map* represents generated feature map signals, the histograms of which were used to generate feature vectors.

6: Generate feature vectors by applying histogram extraction.

$$f_k = \sigma(map_k) \quad (9)$$

where f represents a feature vector of length 256 (as maps are coded with eight bits); and $\sigma(.)$, histogram extraction function. Three feature vectors were generated in this step.

7: Apply the FF algorithm to choose the most suitable feature vector.

The FF algorithm (Algorithm 1) was employed to select the feature vector, among those derived from the lower ternary, signum, and upper ternary kernels, with the highest computed mean square error (MSE) value. Consequently, a feature vector of length 256 was produced for each signal.

Below is the pseudocode for Algorithm 1. Initially, primary signals were derived by averaging the values within each signal group. To determine the congruence between the feature vectors of the target and the primary signals of other classes, MSE values were computed. The best feature vector was identified using features from the other classes, and the potency of these vectors was assessed by contrasting them with feature vectors from different classes.

IV. THE FF-BTP-BASED FEATURE ENGINEERING MODEL

The model consists of three phases: (1) multilevel FF-BTP-based feature extraction (refer to Section III); (2) feature selection using INCA; and (3) classification through a SVM. Our objective, utilizing lightweight handcrafted feature engineering techniques, was to mimic the recognized superior classification performance of resource-intensive deep learning models. This was achieved by enhancing feature representation via MDWT-based signal decomposition, facilitating multilevel feature generation across both spatial and frequency domains, coupled with FF-BTP-based multi-textural feature vector extraction. With the latter, an algorithm (Algorithm 1) inspired by FF neural networks [34] was applied to select the most distinctive feature vector based on

Algorithm 1 FF Algorithm for Final Feature Vector Selection

Input: Input signal (S) with length L.
Output: Selected best feature vector (fv).

00: Load signal and feature vectors.
01: Generate average signals (primary signals) for each category.
02: Extract feature vectors of the input signal.
$[f_1, f_2, \ldots, f_n] = FG(signal)$;
03: Extract features of others' classes primary signals per the used signal. We have extracted features of others to g
$[f_1^1, f_2^1, \ldots, f_n^1] = FE(S_1)$; $[f_1^2, f_2^2, \ldots, f_n^2] = FE(S_2)$; $\ldots [f_1^m, f_2^m, \ldots, f_n^m] = FE(S_m)$;
// where f represents feature vector; FE, feature extraction function; and S, primary signal.
04: Calculated MSE of each generated feature vector.
$mse_j = \frac{1}{L}\sum_{i=1}^{m}\left(\left(f_j(i) - f_j^1(i)\right)^2 + \left(f_j(i) - f_j^2(i)\right)^2 + \ldots + \left(f_j(i) - f_j^m(i)\right)^2\right)$,
$j \in \{1, 2, \ldots, n\}z$
05: Find the maximum MSE.
$ind = \max(mse)$; // where ind represents the index of the maximum MSE.
06: Select the feature vector using the calculated index.
$fv = f_{ind}$;

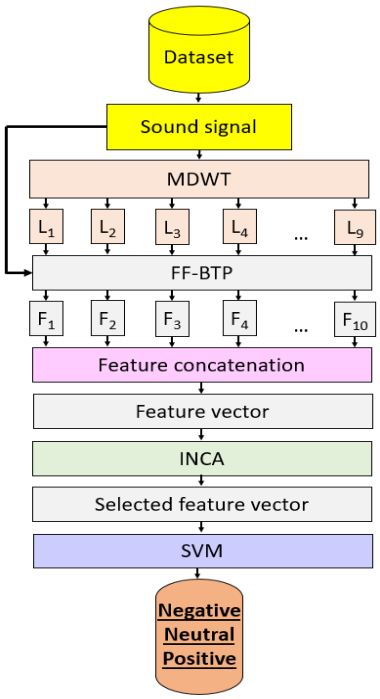

FIGURE 2. Proposed FF-BTP-based feature engineering model. * L: Low pass filter bands, FF-BTP: the proposed forward-forward binary ternary pattern, f: feature vector, INCA: iterative neighborhood component analysis feature selector, SVM: support vector machine classifier.

MSE. Data dimensionality of the multitude of feature vectors generated was reduced using INCA [43], which selected the most discriminative features to feed to downstream standard shallow SVM classifier [44] (Figure 2).

The three phases of the FF-BTP model are explained in the seven steps below.

A. FEATURE EXTRACTION

We combined the MDWT and FF-BTP to extract multilevel features. We used both raw signals and their wavelet band coefficient. The proposed FF-BTP served as the primary feature generation function. We employed a self-organized feature extraction function, making our method inherently self-organized. By analyzing the raw signal, we extracted features in the spatial domain. Concurrently, features in the frequency domain were derived using the proposed FF-BTP in tandem with wavelet bands generated by the MDWT. Given these methodologies, we have proposed the extraction of features that are:

- Multilevel (utilizing MDWT),
- Self-organized (employing FF-BTP),
- Spanning multiple spaces.

To clarify the proposed FF-BTP and MDWT-based feature extraction method, we have given the steps of this method below.

Step 1: Decompose the input signal into wavelet bands using MDWT.

$$[low^1, high^1] = \theta\,(signal) \tag{10}$$

$$\left[low^g, high^g\right] = \theta\left(low^{g-1}\right), g \in \{2, 3, \ldots, 9\} \tag{11}$$

where $\theta(.)$ represents the one-dimensional discrete wavelet function (here, we used the well-known symlet 4 wavelet filter); and *low* and *high*, low and high pass filter bands, respectively. MDWT with nine levels was applied; the generated 9 low-pass filter bands, together with the raw sound signal, were input to the FF-BTP feature extractor.

Step 2: Extract textural features from low-pass filter bands and raw sound signals.

$$fv^1 = \beta(signal) \tag{12}$$

$$fv^t = \beta\left(low^{t-1}\right), t \in \{2, 3, \ldots, 10\} \tag{13}$$

where fv represents the feature vector extracted by FF-BTP; and $\beta(.)$, the feature generation function. A final feature vector of length 256 was generated for each input: the first feature vector corresponding to the raw sound signal; and the 2^{nd} to 10^{th} feature vectors, to the low^1 to low^9 wavelet bands.

Step 3: Merge the generated feature vectors into a final feature vector of length 2560 (=256 × 10).

$$ff\,(a + 256 \cdot (q-1)) = fv^q\,(a), q \in \{1, 2, \ldots, 10\}, \quad a \in \{1, 2, \ldots, 256\} \tag{14}$$

where ff represents the final feature vector.

B. FEATURE SELECTION

We applied INCA [43], which was modified from neighborhood component analysis function, for feature selection; the key parameters were the range of iteration and the misclassification ratio calculator. The steps are detailed below.

Step 4: Calculate qualified indexes of the extracted features.

$$X = \frac{ff - \min(ff)}{\max(ff) - \min(ff)} \tag{15}$$

$$ix = \psi\,(X, out) \tag{16}$$

where X represents normalized feature vector after min-max normalization; ix, qualified index vector; $\psi(.)$, neighborhood component analysis feature selector; and out, real output.

Step 5: Select feature vectors iteratively and calculate their misclassification values.

$$s^z\left(d, b^z\right) = X\left(d, ix\left(b^z\right)\right), d \in \{1, 2, \ldots, ns\}, \quad b \in \{stv, stv+1, \ldots, fnv\}, \quad z \in \{1, 2, \ldots, fnv - stv + 1\} \tag{17}$$

$$mcv\,(z) = \gamma\left(s^z, out\right) \tag{18}$$

where s represents the selected feature vector; ns, number of signals; mcv, misclassification value; and $\gamma(.)$, misclassification rate calculation function (here, we used SVM).

Step 6: Select the feature vector with the minimum misclassification value.

$$il = \min(mcv) \tag{19}$$

$$fv^{final} = s^{il} \tag{20}$$

where il is an index of the minimum misclassification value and fv^{final} is the selected final feature vector.

C. CLASSIFICATION

The classification was performed with SVM classifier [44] using a 10-fold cross-validation strategy. SVM hyperparameters were fine-tuned using a Bayesian optimizer. The specific steps are detailed below:

Step 7: Classify the selected final feature vector.

$$result = \gamma\left(fv^{final}, out\right) \tag{21}$$

The above seven steps defined the proposed FF-BTP-based SCED classification model.

V. EXPERIMENTAL RESULTS

A. SETUP

Our parametric model was implemented in MATLAB 2023a programming environment (see Table 2 for model parameter settings) on a standard personal computer with 64 GB of main memory, a 3.6 GHz processor and Windows 10 Professional operating system. We employed a series of m-files to construct the model, and used the MATLAB Classification Learner toolbox to select the model classifier: SVM outperformed other shallow classifiers in the toolbox and was thus selected.

The feature extraction phase produced a final feature vector of length 2560. INCA feature selection function was then applied to select the most valuable and informative attributes from the original set of features. Across the iteration range of 100 to 1000, the selected final feature vector containing

TABLE 2. Parameters settings for the FF-BTP-based feature engineering model.

Phase	Method	Parameters and explanation
Feature extraction	MDWT	Number of levels: 9; filter: symlet 4.
	FF-BTP	Upper ternary-, lower ternary-, and signum function-based kernels were used to extract three distinct feature vectors, each of length 256, that captured essential sound signal textural characteristics. An FF feature selection function based on MSE was used to select the most distinctive feature vector.
	Feature generation	10 feature vectors generated from the 1 raw sound signal and 9 wavelets bands were merged into a feature vector of length 2560.
Feature selection	INCA	Neighborhood component analysis was used with default settings; the number of iterations was set at 100 to 1000 to capture a wide range of potential feature combinations. SVM with 10-fold cross-validation was used as a misclassification value generator.
Classification	SVM	Kernel: Gaussian; kernel scale: 2.8614; box constraint: 294.6558; standardize: false; multiclass method: one-vs-all; validation: 10-fold cross-validation.
	Bayesian optimizer	Number of iterations: 100; fitness function: classification accuracy.

TABLE 3. Class-wise and overall classification performance of the FF-BTP-based model.

Class	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Negative	96.63	97.05	96.84	-
Neutral	97.68	96.82	97.25	-
Positive	97.36	97.79	97.57	-
Overall	97.22	97.22	97.22	97.22

the top 475 most discriminative features possessed the lowest misclassification rate (Figure 3) and was thus chosen as the optimal input for the downstream classifier.

B. PERFORMANCE EVALUATION METRICS

The task at hand was a multiclass classification problem into three classes: negative, neutral, and positive emotions. Standard metrics were used to assess overall and class-wise performance: accuracy, recall, precision, and F1-score.

C. RESULTS

Across all three classes, there were low rates of misclassification, as shown in the confusion matrix (Figure 4), as well as excellent class-wise and overall performance (Table 3). The overall accuracy attained was 97.22%.

VI. DISCUSSION

In this work, we introduced a new concept of SCED, which broadens the scope of the nascent field of speech emotion detection to crowd situations. A new SCED study database

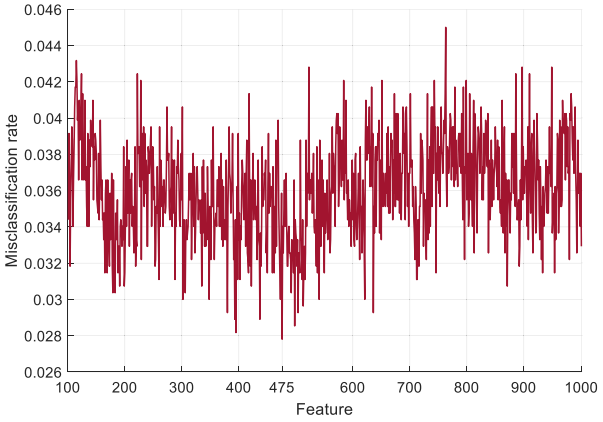

FIGURE 3. Iterative feature selection (range 100 to 1000) using INCA. The feature vector of length 475 has the lowest misclassification rate of 0.0278.

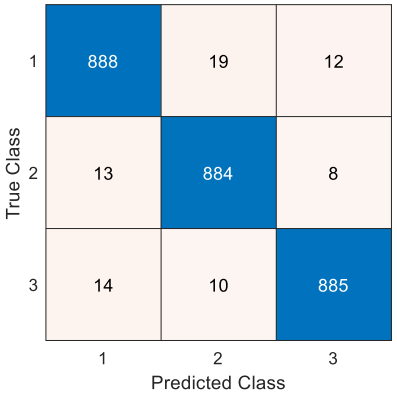

FIGURE 4. Confusion matrix of the FF-BTP-based feature engineering model. *1: Negative; 2: Neutral; 3: Positive.

TABLE 4. Classification accuracies for the ablation studies.

Case	Accuracy (%)
1	95.98
2	96.05
3	95.68
4	97.22

comprising balanced classes of crowd sounds divided into distinct negative, neutral and positive emotion classes was established to test our novel handcrafted FF-BTP-based feature engineering model. The latter accomplished multi-level and input-dependent dynamic textural feature extraction and selection, using BTP and a neural network-inspired FF algorithm, for comprehensive characterization of one-dimensional high-fidelity sound signals that mimicked deep learning. On the SCED study dataset, the model attained excellent overall and class-wise classification performance. Of note, the parametric classification framework confers optionality for its adoption in diverse one-dimensional signal datasets and classification tasks.

In the feature extraction phase, the BTP feature extraction function generated three feature vectors based on lower

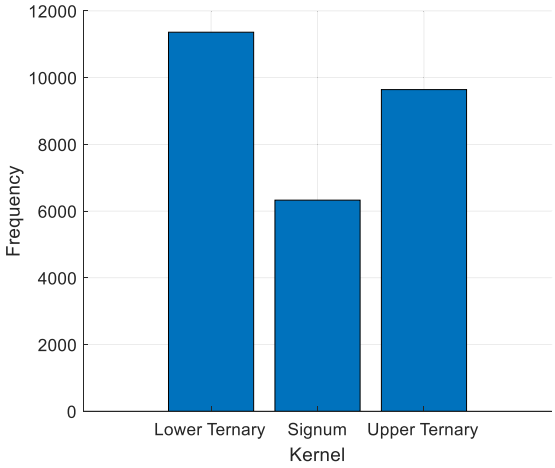

FIGURE 5. FF algorithm-based feature vector selection stratified by kernel function.

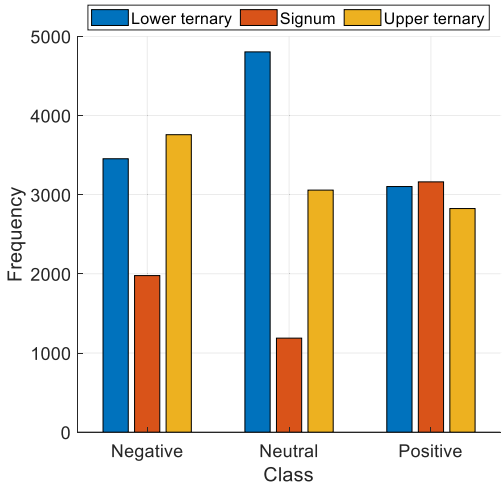

FIGURE 6. FF algorithm-based feature vector selection stratified by kernel function and class. The upper ternary, lower ternary, and signum kernels dominated in the negative, neutral, and positive classes, respectively.

ternary, signum, and upper ternary functions for each input signal block; and the FF algorithm selected the one with the maximum MSE in relation to the other classes. The lower ternary function-based feature vector was the most commonly selected by the FF algorithm overall (Figure 5), a result that was driven by its predominance in the neutral class (Figure 6).

In the feature selection phase, INCA selected the 475 most informative features from among the initial pool of 2560 merged features generated from the raw sound signal and nine MDWT-decomposed wavelet bands. Among all ten possible input signals, the L1 and L9 wavelet bands contributed the highest (103 out of 475) and lowest (1 of 475) proportions of selected features (Figure 7).

A. ABLATION STUDIES

To dissect the contributions of individual components of the parametric FF-BTP feature engineering architecture, we performed ablation studies as listed below.

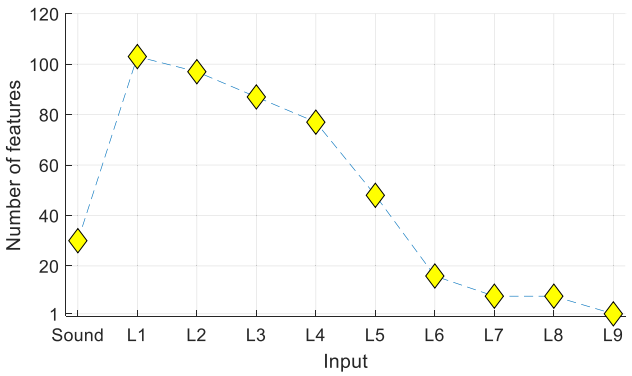

FIGURE 7. Distribution of INCA-selected features stratified by model input (raw signal or wavelet band L1 to L9).

Case 1: Using lower ternary kernelled features only; other model components kept constant.

Case 2: Using signum kernelled features only combined with the well-known one-dimensional local binary pattern-based feature extraction; other model components kept constant.

Case 3: Using upper ternary kernelled features only; other model components kept constant.

Case 4: Our full model, incorporating all components.

The accuracies of the defined cases are listed in Table 4. Case 4 (our full mode) attained the best classification performance, which outperformed by 1.17% the next best case, Case 2, with 96.05% classification accuracy.

B. HIGHLIGHTS

Strengths:

- We have curated a new three-class SCED sound dataset. It was hoped that this contribution would stimulate research interest in developing models for SCED. This dataset was publicly published to contribute to advanced signal processing and SCED. Users can download this dataset using https://www.kaggle.com/datasets/arifmetehanyldz/sced-v1 URL.
- We employed various techniques–MDWT, BTP, FF–for multilevel deep feature vector extraction to emulate deep learning models and attained excellent classification performance.
- Our handcrafted model possessed linear time complexity and is feasible to implement without the need to train millions of parameters as with deep learning models.

Limitations:

- The study dataset was relatively modest in size; collection of larger SCED datasets is labor-intensive.
- Our model used the primary signals to calculate MSEs with little time complexity. However, this lacks generalizability and may not always yield similar high accuracy with other datasets.

VII. CONCLUSION

Our study on emotion recognition has highlighted the transformative capabilities of the SCED model. Designed with a

strong automation-centric approach, this model demonstrates its sueriority in analyzing crowdsourced acoustic data and aims to accurately identify these signals into distinguishable emotional categories. The introduction of the SCED model heralds a significant departure from traditional susceptibility detection methods, offering advanced methodologies characterized by their sensitivity and efficiency.

The design of our model outlines the complex interplay between two important feature extraction methods: MDWT and FF-BTP. These techniques collaboratively reveal multilayered and complex textural nuances from underlying data, paving the way for a comprehensive analysis of auditory signals – a degree of analysis that surpasses many existing models.

An integral component of our model is the FF algorithm inspired by established neural network methodologies. More than just a simple component, this algorithm performs the vital function of identifying and selecting the most relevant features, ensuring that the decisions made by the model are based on the most critical data insights. This naturally increases the likelihood of accurate emotional classification.

While the theoretical framework of our model paints a promising picture, its real-world applicability and effectiveness have been confirmed by empirical studies. The model's performance, demonstrated by its impressive 97.22% accuracy across multiple categories, is a testament to its robust design and the effectiveness of the methods we integrated. This not only reflects the operational excellence of the model but also reaffirms the robustness of the approaches that underpin our research.

VIII. FUTURE WORKS

We intend to accrue a larger dataset that encompasses more classes of emotions. Further, we aim to start a new project that combines feature engineering with deep learning techniques, i.e., handcrafted deep learning networks. For example, using the FF algorithm (Algorithm 1) for selecting the most suitable feature vector algorithm together with Hinton's FF algorithm [34]. These networks, in conjunction with newer larger SCED datasets, will facilitate the development of cutting-edge emotion detection tools for communities.

In this work, we applied the proposed FF-BTP to sound signals. Looking ahead, this feature extraction method has several potential application areas: (i) physiological signals, and (ii) by introducing a 2D version, textural features from images can be generated. This 2D version can then be employed for classifying faces, textural images, and other biometric images, such as those of veins and irises. With these considerations in mind, we propose a versatile feature extraction function. Also, we can apply this feature extraction function.

Declarations

Author contributions: All authors contributed equally to the study.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethics approval was not required for this research.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used dataset can be downloaded using https://www.kaggle.com/datasets/arifmetehanyldz/sced-v1 URL.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] C. Regenbogen, D. A. Schneider, R. E. Gur, F. Schneider, U. Habel, and T. Kellermann, "Multimodal human communication—Targeting facial expressions, speech content and prosody," *NeuroImage*, vol. 60, no. 4, pp. 2346–2356, May 2012.

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, Jun. 2012.

[3] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101847.

[4] T. McElroy, J. Salapska-Gelleri, K. Schuller, and M. Bourgeois, "Thinking about decisions: How human variability influences decision-making," in *Brain, Decision Making and Mental Health*. Cham, Switzerland: Springer, 2023, pp. 487–510.

[5] S. A. Alanazi, M. Shabbir, N. Alshammari, M. Alruwaili, I. Hussain, and F. Ahmad, "Prediction of emotional empathy in intelligent agents to facilitate precise social interaction," *Appl. Sci.*, vol. 13, no. 2, p. 1163, Jan. 2023.

[6] W. Lin and C. Li, "Review of studies on emotion recognition and judgment based on physiological signals," *Appl. Sci.*, vol. 13, no. 4, p. 2573, Feb. 2023.

[7] A. Moin, F. Aadil, Z. Ali, and D. Kang, "Emotion recognition framework using multiple modalities for an effective human–computer interaction," *J. Supercomput.*, vol. 79, pp. 9320–9349, Jan. 2023.

[8] C. Filippini and A. Merla, "Systematic review of affective computing techniques for infant robot interaction," *Int. J. Social Robot.*, vol. 15, no. 3, pp. 393–409, Mar. 2023.

[9] F. Daneshfar and M. B. Jamshidi, "An octonion-based nonlinear echo state network for speech emotion recognition in Metaverse," *Neural Netw.*, vol. 163, pp. 108–121, Jun. 2023.

[10] J. Wen, D. Jiang, G. Tu, C. Liu, and E. Cambria, "Dynamic interactive multiview memory network for emotion recognition in conversation," *Inf. Fusion*, vol. 91, pp. 123–133, Mar. 2023.

[11] J. Li, X. Wang, G. Lv, and Z. Zeng, "GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection," *IEEE Trans. Affect. Comput.*, early access, Mar. 24, 2023, doi: 10.1109/TAFFC.2023.3261279.

[12] Y. Cai, X. Li, and J. Li, "Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review," *Sensors*, vol. 23, no. 5, p. 2455, Feb. 2023.

[13] M. Garg, "Mental health analysis in social media posts: A survey," *Arch. Comput. Methods Eng.*, vol. 30, pp. 1819–1842, Jan. 2023.

[14] J. Zhou, S. Yang, C. Xiao, and F. Chen, "Examination of community sentiment dynamics due to COVID-19 pandemic: A case study from a State in Australia," *Social Netw. Comput. Sci.*, vol. 2, no. 3, pp. 1–11, May 2021.

[15] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106547.

[16] E. Akbal, P. D. Barua, T. Tuncer, S. Dogan, and U. R. Acharya, "Development of novel automated language classification model using pyramid pattern technique with speech signals," *Neural Comput. Appl.*, vol. 34, no. 23, pp. 21319–21333, Dec. 2022.

[17] D. Tanko, S. Dogan, F. B. Demir, M. Baygin, S. E. Sahin, and T. Tuncer, "Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23," *Appl. Acoust.*, vol. 190, Mar. 2022, Art. no. 108637.

[18] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, p. 21552, Nov. 2021.

[19] S. Ghosh, A. Priyankar, A. Ekbal, and P. Bhattacharyya, "Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data," *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110182.
[20] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari, and J. Qin, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Inf. Fusion*, vol. 93, pp. 282–301, May 2023.
[21] Z. Yang, Z. Liu, S. Liu, L. Min, and W. Meng, "Adaptive multi-view selection for semi-supervised emotion recognition of posts in online student community," *Neurocomputing*, vol. 144, pp. 138–150, Nov. 2014.
[22] B. T. Atmaja and A. Sasou, "Sentiment analysis and emotion recognition from speech using universal speech representations," *Sensors*, vol. 22, no. 17, p. 6369, Aug. 2022.
[23] R. Kaur, A. Majumdar, P. Sharma, and B. Tiple, "Analysis of tweets with emoticons for sentiment detection using classification techniques," in *Proc. 19th Int. Conf., Distrib. Comput. Intell. Technol.* Bhubaneswar, India: Springer, Jan. 2023, pp. 208–223.
[24] B. Pragati, C. Kolli, D. Jain, A. Sunethra, and N. Nagarathna, "Evaluation of customer care executives using speech emotion recognition," in *Proc. 3rd Int. Conf. Mach. Learn., Image Process., Netw. Secur. Data Sci.* Cham, Switzerland: Springer, 2023, pp. 187–198.
[25] M. F. Bashir, A. R. Javed, M. U. Arshad, T. R. Gadekallu, W. Shahzad, and M. O. Beg, "Context-aware emotion detection from low-resource Urdu language using deep neural network," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 5, pp. 1–30, May 2023.
[26] D. A. Widodo, N. Iksan, and B. Sunarko, "Sentiment analysis of Twitter media for public reaction identification on COVID-19 monitoring system using hybrid feature extraction method," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 1, pp. 92–99, Mar. 2023.
[27] M. Ali, M. Hassan, K. Kifayat, J. Y. Kim, S. Hakak, and M. K. Khan, "Social media content classification and community detection using deep learning and graph analytics," *Technol. Forecasting Social Change*, vol. 188, Mar. 2023, Art. no. 122252.
[28] X. Yao, S. Chen, and G. Yu, "Effects of members' response styles in an online depression community based on text mining and empirical analysis," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103198.
[29] B. Wang, G. Dong, Y. Zhao, R. Li, Q. Cao, K. Hu, and D. Jiang, "Hierarchically stacked graph convolution for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110285.
[30] X. Qin, Z. Wu, J. Cui, T. Zhang, Y. Li, J. Luan, B. Wang, and L. Wang, "BERT-ERC: Fine-tuning BERT is enough for emotion recognition in conversation," 2023, *arXiv:2301.06745*.
[31] D. Gu, M. Li, X. Yang, Y. Gu, Y. Zhao, C. Liang, and H. Liu, "An analysis of cognitive change in online mental health communities: A textual data analysis based on post replies of support seekers," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103192.
[32] A. Utku, U. Can, and S. Aslan, "Detection of hateful Twitter users with graph convolutional network model," *Earth Sci. Informat.*, vol. 16, pp. 329–343, Jan. 2023.
[33] M. Arbane, R. Benlamri, Y. Brik, and A. D. Alahmar, "Social media-based COVID-19 sentiment classification model using Bi-LSTM," *Expert Syst. Appl.*, vol. 212, Feb. 2023, Art. no. 118710.
[34] G. Hinton, "The forward-forward algorithm: Some preliminary investigations," 2022, *arXiv:2212.13345*.
[35] S.-H. Fang, W.-H. Chang, Y. Tsao, H.-C. Shih, and C. Wang, "Channel state reconstruction using multilevel discrete wavelet transform for improved fingerprinting-based indoor localization," *IEEE Sensors J.*, vol. 16, no. 21, pp. 7784–7791, Nov. 2016.
[36] E. Akbal and T. Tuncer, "A learning model for automated construction site monitoring using ambient sounds," *Autom. Construct.*, vol. 134, Feb. 2022, Art. no. 104094.
[37] A. Dogan, P. D. Barua, M. Baygin, T. Tuncer, S. Dogan, O. Yaman, A. H. Dogru, and R. U. Acharya, "Automated accurate emotion classification using Clefia pattern-based features with EEG signals," *Int. J. Healthcare Manage.*, pp. 1–14, Nov. 2022, doi: 10.1080/20479700.2022.2141694.
[38] T. Tuncer, S. Dogan, M. Baygin, and U. R. Acharya, "Tetromino pattern based accurate EEG emotion classification model," *Artif. Intell. Med.*, vol. 123, Jan. 2022, Art. no. 102210.
[39] T. Tuncer, S. Dogan, and A. Subasi, "LEDPatNet19: Automated emotion recognition model based on nonlinear LED pattern feature extraction function using EEG signals," *Cogn. Neurodynamics*, vol. 16, pp. 779–790, Nov. 2021.
[40] D. Tanko, F. B. Demir, S. Dogan, S. E. Sahin, and T. Tuncer, "Automated speech emotion polarization for a distance education system based on orbital local binary pattern and an appropriate sub-band selection technique," *Multimedia Tools Appl.*, pp. 1–18, Apr. 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-14648-y
[41] M. Baygin, I. Tuncer, S. Dogan, P. D. Barua, T. Tuncer, K. H. Cheong, and U. R. Acharya, "Automated facial expression recognition using exemplar hybrid deep feature generation technique," *Soft Comput.*, vol. 27, pp. 8721–8737, Apr. 2023.
[42] R. Vempati and L. D. Sharma, "A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence," *Results Eng.*, vol. 18, Jun. 2023, Art. no. 101027.
[43] T. Tuncer, S. Dogan, F. Özyurt, S. B. Belhaouari, and H. Bensmail, "Novel multi center and threshold ternary pattern based method for disease detection method using voice," *IEEE Access*, vol. 8, pp. 84532–84540, 2020.
[44] V. Vapnik, "The support vector method of function estimation," in *Nonlinear Modeling*. Cham, Switzerland: Springer, 1998, pp. 55–85.

ARIF METEHAN YILDIZ received the bachelor's and master's degrees in forensic engineering from Firat University, Elazig, Turkey, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Forensic Informatics Engineering. He is a Lecturer with the Distance Education Application and Research Center, Ardahan University. His current research interests include cryptology, malware analysis, artificial intelligence, machine learning, signal processing, computer vision, cyber security, and social media expertise.

MASAYUKI TANABE (Member, IEEE) received the Ph.D. degree in engineering from Tokyo Metropolitan University, in 2011. Following his graduation, he joined the Faculty of Advanced Science and Technology, Kumamoto University, as an Assistant Professor, where he has been serving, since 2011. In addition to his academic career, he has ventured into entrepreneurship. His current research interests include medical ultrasound imaging and biosignal analysis.

MAKIKO KOBAYASHI (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the Department of Electrical and Electronic Engineering, Chiba University, Japan, in 1997 and 1999, respectively, and the Ph.D. degree from McGill University, Montreal, QC, Canada, in 2004. From 2004 to 2011, she contributed her expertise with the Industrial Materials Institute, a division of the National Research Council of Canada (NRCC). Since 2012, she has been an Associate Professor with Kumamoto University, where she was a Professor, in 2022. She has been the Co-Founder and a Technical Advisor with CAST Company Ltd., since 2019. Her current research interest includes the development of porous piezoelectric materials for high-temperature ultrasonic transducers and various applications. Her notable achievements include the Young Scientist Award at the 26th Symposium on Ultrasonic Electronics, in 2005, and being honored with the 2022 NRC's IP Achievement Award.

ILKNUR TUNCER was born in Elazig, in 1988. She received the B.Sc. degree in computer sciences education from the Faculty of Education, Firat University, in 2015, and the master's degree in technology and information management from Firat University, in 2019. Her current research interest includes machine and deep learning in medical applications.

PRABAL DATTA BARUA received the Ph.D. degree in information systems from the University of Southern Queensland. He is currently an Academic and Accredited Research Supervisor with the University of Southern Queensland. He is an Industry Leader of the ICT Entrepreneurship Project, Australia, and sits as an ICT advisory panel member of many organizations. He is also an Adjunct Professor with the University of Southern Queensland and an Honourary Industry Fellow with the University of Technology Sydney. He has 12 years of teaching experience. He received research support from the Queensland Government Innovation Connections under the Entrepreneurs program to research "Cancer Recurrence Using Innovative Machine Learning Approaches." He has published several articles in the Q1 journal. His current research interests include AI technology development in health, education, agriculture, and environmental science.

SENGUL DOGAN received the master's degree in bioengineering and the Ph.D. degree in electrical and electronics engineering from Firat University, Elazig, Turkey, in 2007 and 2011, respectively. She is currently an Associate Professor with the Department of Digital Forensics Engineering, Faculty of Technology, Firat University. Her current research interests include computer forensics, mobile forensics, image processing, and signal processing. She has been working actively on developing algorithms in machine learning for biomedical data.

TURKER TUNCER received the master's degree in electronics and computer sciences and the Ph.D. degree in software engineering from Firat University, Elazig, Turkey, in 2011 and 2016, respectively. He is currently an Associate Professor with the Department of Digital Forensics Engineering, Faculty of Technology, Firat University. His current research interests include feature engineering, image processing, signal processing, information security, and pattern recognition. He has been working actively on developing algorithms in machine learning applied to visual surveillance and biomedical data.

RU-SAN TAN is currently a Senior Consultant with the Department of Cardiology, National Heart Centre Singapore. His specialization is in noninvasive diagnostic cardiac imaging: cardiovascular magnetic resonance imaging, echocardiography, and nuclear cardiology. His current research interests include advanced cardiac imaging, cardiac biomechanics, and computational modeling. He is also a site PI and a member of the steering committees of multinational clinical trials of novel cardiology drugs and notably novel anticoagulants.

U. RAJENDRA ACHARYA received the Ph.D., D.Eng., and DSc. degrees. He is currently a Professor with the University of Southern Queensland, Australia; a Distinguished Professor with the International Research Organization for Advanced Science and Technology, Kumamoto University, Japan; an Adjunct Professor with the University of Malaya, Malaysia; and an Adjunct Professor with Asia University, Taiwan. His funded research has accrued cumulative grants exceeding six million Singapore dollars. He has authored more than 500 publications, including 345 in refereed international journals, 42 in international conference proceedings, and 17 books. He has received more than 73,000 citations on Google Scholar (with an H-index of 134). His current research interests include biomedical imaging and signal processing, data mining, and visualization, as well as applications of biophysics for better healthcare design and delivery. He has been ranked in the Top 1% of the Highly Cited Researchers for the last seven consecutive years (2016–2022) in computer science, according to the Essential Science Indicators of Thomson. He is on the editorial boards of many journals and has served as a guest editor on several AI-related issues.