



A Ranked-Aware GA with HoG Features for Infant Cry Classification

Suhaila N. Mohammed¹ Adnan J. Jabir^{1*}

¹*Department of Computer Science, College of Science, University of Baghdad, Iraq*

* Corresponding author's Email: adnan.jabir@sc.uobaghdad.edu.iq

Abstract: Infants typically cry to get their parents' attention. Through their cries, infants express their basic needs like hunger, tiredness, pain, and discomfort. Unfortunately, it is difficult to interpret cries to comprehend the demands of an infant. The only way to solve this problem is to analyze the infant's acoustic speech pattern and determine the cause of the crying. In this study, the cry signal is converted to a spectrogram image to take advantage of the wide spectral range of image-based features. Before generating the represented features, the watershed segmentation algorithm is used to remove distracting areas of the image. Then, histogram of gradients (HoG) features are generated. Because the feature vector has high dimensionality, two stages of dimensionality reduction are presented. First, the feature pool is decreased using the fisher score feature selection approach. The ideal feature set is then chosen using a combination of transfer learning, genetic algorithm (GA), and neural networks. To motivate GA to pick characteristics that will operate successfully with the neural network, a ranked aware mutation operator is suggested. As system evaluation material, the donateacry-copus public dataset is employed. Experiments reveal that when 80 HoG features are generated and the best 37 Fisher scores are chosen, the model has the best accuracy of 92% when applying transfer learning to 11 hidden layers of the neural network. The study's findings support the use of image-based features to identify the cause of a baby's crying.

Keywords: Infant cry analysis, Signal spectrogram, Watershed segmentation, Histogram of gradients (HoG), Genetic algorithm (GA), Neural network (NN), Ranked-aware mutation.

1. Introduction

Communication is essential in life. Adults have a variety of communication options. However, for infants, crying is their only mode of communication. Infants convey their requirements, such as hunger, tummy ache, the need for burping, discomfort, and tiredness, by crying in their early stages of life [1].

Infant cry analysis is an interdisciplinary field of research involving physiology, anatomy, and phonetics. The findings of infant cry classification studies are significant for many healthcare professions, such as nurses, midwives, and speech and language therapists, as well as medical professions such as pediatricians, by assisting in the interpretation of the newborn cry to recognize an infant's needs or health state [2].

Acoustic analysis is essential for understanding the genesis of paralinguistic vocalizations. Infant

vocalizations are high-dimensional time series signals, presenting a data reduction challenge. The signal must be represented in terms of the most important traits, those around which development is fundamentally oriented. Acoustic measures used to evaluate infant vocalizations include duration, f_0 , means, peaks, standard deviations, contours, formant frequencies, spectral concentration, and tremor [3]. There have been several research works attempted to address the infant cry problem, for example, Alaie et al. [4] used dynamic mel-frequency cepstral coefficients (MFCCs) and static MFCCs to create a unique cry pattern for each cry vocalization type and pathological condition using the boosting mixture learning (BML) method. Maghfira et al. [5] used a combination of convolutional neural networks (CNN) and recurrent neural networks (RNN) that acts as a feature extraction and classifier method at once. Evaluation in the dunstan baby language dataset shows that the

system gives an average classification accuracy of up to 94.97%. Pathirana and Sumathipala [6] applied the voice activity detection (VAD) method to predict the presence of the human voice. An MFCCs feature vector was extracted, and a K-nearest neighbour (KNN) classifier was trained with infant cry signals labelled "hunger", "belly pain" and "tired". The overall accuracy for frame-wise and event-wise classifications on the donatecry-corpus dataset was 72.51% and 77.46%, respectively. Rani et al. [7] employed the MFCCs as an element extraction strategy and the KNN for classification. The most noticeable precision was 76.16% and it was achieved when the K worth is 2 on a dataset containing 8 classes of sound records, i.e., awake, belly torment, burping, discomfort, hugging, hungry, sleepy, and tired. Sutanto et al. [8] used MFCC and CNN for the recognition of infant crying. In this work, around 85% accuracy was achieved using five classes of the donatecry-corpus dataset. Cha and Bae [9] used MFCC and short-time fourier transform (STFT) with deep neural networks (DNN) for classifying four types of crying (pain, stomachache, discomfort, and fatigue). They achieved an accuracy of 85.59% for DNN and 84.49% for STFT using the donatecry-corpus dataset.

However, because automatic infant cry classification is a pattern recognition problem, it frequently deals with vast amounts of redundant and irrelevant data. The redundant characteristics and irrelevant features have no bearing on the underlying structure of the data and neither provide any new information about it. This circumstance may reduce classifier prediction performance while increasing computational processing time. Most of the research in this field [4-9] is aimed at solving the infant cry classification problem without addressing the problem of the higher dimensionality of the data, which can affect the accuracy of the suggested model. For example, 39 MFCC features, 40 cepstral coefficients, and 16 MFCC features are extracted from each sound frame in [4, 7, 8] respectively. In [5], a two-dimensional space of 64×64 is extracted from the CNN. While in [6, 9] the whole MFCC and STFT characteristics are acquired from each sound frame. The simplest solution to this problem is to select the relevant features and discard the rest. This is known as feature selection, and it is divided into two categories: filter techniques and wrapper approaches. Filter approaches work independently of a classifier, whereas wrapper strategies use the classification process as part of the function evaluation to find relevant feature subsets [10].

By concentrating on the filter and wrapper

methods for the selection of the relevant subset of features, we attempt both to generate a more effective set of acoustic features and to reduce the large dimensionality of the data in this study. The primary contributions of the presented work are:

- Transforming the cry signal into a spectrogram image, which is then segmented using the watershed technique to remove noise.
- Applying histogram of gradients (HoG) features to the backdrop segment to provide image-based features.
- Reducing the dimensionality of the features first using the fisher score measure and then picking the optimal set using the wrapper technique.
- Proposing a mutation-aware genetic algorithm with a neural network as the fitness function and a weighted rank for the mutation operator.
- Using the transfer learning concept during the search for the optimal set of attributes in different genetic algorithm generations.

The remainder of the paper is structured as follows: Section 2 provides the theoretical basis for the approaches used. We described the proposed method for infant cry classification in section 3. Section 4 includes the findings and discussions of the experimental tests. Section 5 discusses the future scope of the investigation and concludes the paper.

2. Theoretical background

2.1 Watershed algorithm

Image segmentation is a basic image processing technique that is primarily used for finding segments that form the entire image [11]. Watershed segmentation is one of the image segmentation algorithms that combines geomorphology and regional growth concepts. It perceives a grayscale image as a "topographic map", with high-pixel areas indicating mountains and low-pixel areas denoting low-lying areas. If it rains, water will flow down mountainsides and into valleys, forming "lakes". Catchment basins are areas where water levels rise, potentially spilling over into surrounding catchments. There won't be an overflow of water if a dam is constructed at the intersection of each catchment basin. The watershed line (i.e., the desired image segmentation outcome) is located at these dam positions [12]. Fig. 1 demonstrates the idea of the watershed algorithm.

Let $f: D \rightarrow N$ be a grayscale image, and the max and min values of f are denoted by h_{min} and h_{max} , respectively. Construct a recursion with the gray

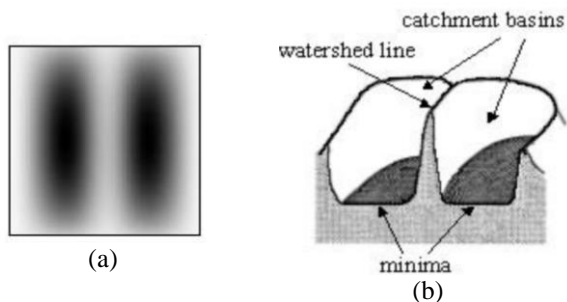


Figure. 1 The idea of the watershed algorithm: (a) an example of a grayscale image and (b) the resulted topographic [13]

level h rising from h_{min} to h_{max} , where the basins related to the minima of f are increased successively. Let X_h be the fusion of the basins be estimated at level h . An updated value for X_{h+1} is obtained by computing the geodesic influence zone of X_h within $Th+1$. A connected component of the threshold set $Th+1$ at level $h+1$ might represent a new minimum or an expansion of an X_h basin. Let Min_h represents the union of all regional minima at altitude h [14], then:

$$X_{h_{min}} = \{p \in D | f(p) = h_{min}\} = Th_{min} \quad (1)$$

$$X_{h+1} = Min_{h+1} \cup IZ_{h+1}(X_h), h \in \{h_{min}, h_{max}\} \quad (2)$$

The watershed $Wshed(f)$ is the complement of $X_{h_{max}}$ in D :

$$Wshed(f) = D \setminus X_{h_{max}} \quad (3)$$

2.2 Histogram of gradients

Dalal and Triggs in [15] introduced the histogram of oriented gradients (HoG), a feature descriptor used in computer vision and image processing for object recognition and image classification tasks. HoG is based on the idea that the distribution of local intensity gradients or edge directions, which are by definition perpendicular to the gradient's direction, may accurately describe the local appearance and shape of an object in an image [16]. In general, the HoG algorithm is composed of four phases [17]:

Phase 1 (Gradient Computation): Gaussian smoothing is used to generate gradients, which are then followed by discrete derivative masks.

Phase 2 (Orientation Binning): The second phase is responsible for partitioning the gradient image into small, connected areas called "cells" and quantizing edge orientations into q bins.

Phase 3 (Descriptor Blocks): In the third phase,

groupings of neighboring cells are viewed as spatial regions known as blocks. The grouping of cells into blocks is the foundation for histogram grouping and normalization.

Phase 4 (Normalization Blocks): Lastly, the previously defined blocks are normalized. Let v be the un-normalized block, and $\|v\|^k$ be its k -norm for $k=1, 2$, then the normalized v can be obtained as:

$$v = \frac{v}{\sqrt{\|v\|_k^k + \epsilon^k}} \quad (4)$$

Where ϵ be a small normalization constant to prevent division by zero. The final descriptor is then the vector of all components of the normalized cell responses from all of the blocks in the image.

2.3 Genetic algorithm

Evolutionary algorithms (EA) are stochastic optimization algorithms inspired by biology and the processes that enable organismal populations to adapt to their surroundings. They keep a population of potential solutions (chromosomes) and make them evolve by applying stochastic operators iteratively. A genetic algorithm (GA) is part of an EA and can be used to solve optimization problems. The basic elements of GA are [18]:

1) *Chromosome representation:* Before a genetic algorithm can be utilized to solve any problem, a mechanism for encoding potential solutions in a form that the computer can comprehend is required. Some of the encoding methods are binary encoding, value encoding, permutation encoding, and tree encoding.

2) *The Evolutionary Cycle:* An evolutionary algorithm enters a loop after generating an initial population. Each iteration (also known as a generation) will result in the creation of a new population by applying a given number of stochastic operators to the previous population. The appropriate operator is selected based on how solutions are encoded. GA operators include:

- *Selection:* It aims to simulate the Darwinian law of "survival of the fittest". It selects a subset of the population of individuals based on their fitness values. Many selection operators exist; some of them are roulette wheel selection, tournament selection, rank selection, etc.

- *Crossover:* combining two chromosomes (parents) to produce new chromosomes (offspring) that are better than their parents. Crossover occurs according to a user-defined crossover probability.

There are different types of crossover operators, such as one-point, two-point, and arithmetic.

- *Mutation*: The purpose of mutation is to simulate the effect of transcription errors that can happen with a very low mutation probability. The mutation operators are of many types, such as flip bits, uniform, non-uniform, etc.

2.4 Artificial neural network

A neural network (NN) is a computer-simulated depiction of the human brain that attempts to mimic its learning process. The single neuron is the most basic type of NN, yet artificial neurons in isolation are not particularly spectacular and cannot tackle complicated problems. Because of their relative simplicity and computing capability, feed-forward multilayer networks have gotten a lot of interest. A neural network's learning implies an adaptive technique in which the network's weights are progressively updated to improve a pre-specified criterion. The most often used approach for doing supervised learning on feed-forward networks is the backpropagation algorithm. The weights (w_{ij}) are updated in the backproportion algorithm as in Eq. (5) [19].

$$w_{ij}(t + 1) = w_{ij}(t) - \alpha y_i y_j (d_j - y_j) \delta_j \quad (5)$$

Where, d_j is the desired output, y_j is the actual output of the neuron j , δ_j is the error between y_j and d_j , and α ($0 < \alpha < 1$) is a parameter called the learning rate that determines learning speed.

3. Proposed framework

The overall framework of the proposed approach is depicted in Fig. 2. First, the input cry signal is converted to a spectrogram image by dividing it into short frames. To create the spectral representation of the signal, the fast fourier transform (FFT) is utilized together with a windowing function to process these frames. After that, feature extraction is carried out. The features to be utilized are derived from HoG features. The generated features are then fed to the filter-based and wrapper-based feature selection methods in the feature dimension reduction stage. Fisher score is used for the former and ranked-aware GA for the latter. To perform the classification task, neural networks are adopted. The outcome of the classification task will be the causes of the infant's cries (belly pain cry, burping cry, discomfort cry, hungry cry, tired cry).

The significance of our proposed work is that it combines HoG image-based features with GA-NN

feature selection to improve accuracy. The next section provides a thorough description of each stage of this proposed method.

3.1 Dataset description

The donateacry-corpus dataset is used as system evaluation material. It is an infant cry audio corpus that is part of the final project in the speech technology course at the royal institute of technology in Sweden. It contains samples for five classes (belly pain, burping, discomfort, hungry, and tired) categorized based on DBL (Dunstan Baby Language) as a reference. The babies are males and females, 0-2 years old. The data samples are all in WAV format with a sampling rate of 8 kHz. Data tagged with lonely, scared, and unknown are removed. Additionally, data tagged with cold or hot was merged into discomfort. All non-cry data has been manually removed by listening to it. These include white noise, baby chat, adults mimicking baby cries, etc. [20].

3.2 Spectrogram image generation

A spectrogram is a visual depiction of the spectrum of sound frequencies as they change over time. It is a two-dimensional graph with a third dimension expressing the amplitude of a certain frequency at a specific time. To benefit from both acoustic and prosodic information held in the spectrogram image and then represent them as HoG features, the cry signal is converted to a spectrogram representation. The following five steps are applied to build the spectrogram image for the speech signal [21]:

- 1) *Framing and Windowing*: The speech signal is first partitioned into overlapped frames concerning the time. After that, the windowing process is applied to reduce the effect of disconnectivity at the ends of each frame.
- 2) *Frame Transformation*: Each frame is then transformed from the time domain into the frequency domain by incorporating the short term fourier transform (STFT).
- 3) *Intensity Computation*: After the intensity degree of each value in the frame is computed, the frame will be represented as a column in the constructed spectrogram image.

The resulting image has dimensions equal to the number of $blocks \times frame_size$. The image is then rescaled using bilinear interpolation to smaller dimensions to speed up the computations of the remaining stages.

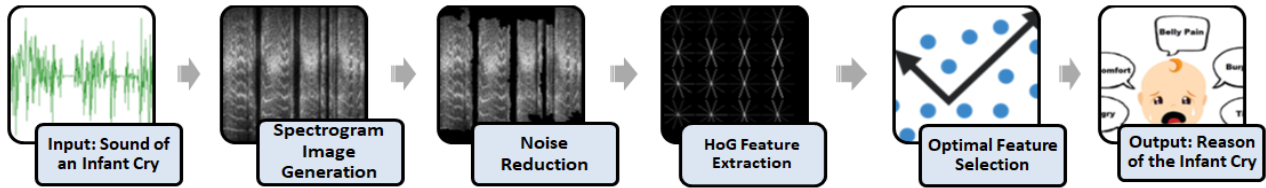


Figure. 2 General framework of the proposed system

3.3 Noise reduction

Apart from the crying, a real-world cry signal usually includes background noise, speech, the sound of medical equipment, and silence [22]. VAD is often used to conduct audio segmentation tasks [23]. The VAD approach is commonly used in speech recognition to recognize human speech in audio sources. It is also used by researchers to detect newborn cries and erase silence periods in sample recordings. VAD also confronts the issue of distinguishing between crying and noise [24].

VAD is not suitable to be applied to the spectrogram representation of the signal because it can only work with the time representation of the sound signal. The watershed algorithm is instead used to reduce the noisy parts of the image by extracting only the background segment and marking the pixels of other segments with black.

3.4 HoG feature extraction

HoG features are generated from the spectrogram image as representational features for classifying the queried sound recording. HoG is employed due to its effectiveness in capturing texture features in each block and representing them as bins of the histogram of pixel gradients. With HoG features, the acoustic and prosodic information of the cry signal is captured and represented in compact bins.

3.5 Optimal feature selection

The HoG features might include some duplicate information that makes it difficult to distinguish between the various cry signal types. This stage's goal is to minimize the dimensionality of the HoG features while maintaining classification accuracy. Feature selection strategies eliminate features that are unrelated to the job at hand, reducing the feature space, saving computing time, and improving classification accuracy.

Two feature selection methods are presented in this work: the filter-based and the wrapper-based methods. The generated HoG features are first fed to the Fisher score algorithm, and the best N_1 features

are retrieved. In turn, these features are given to the GA-NN algorithm, which selects the best N_2 ones. These two methods are described in detail below.

3.5.1. Fisher-based feature reduction

The Fisher score algorithm is one of the filter-based feature selection methods that depends on statistics to compute feature scores. The Fisher score algorithm selects each feature independently based on their scores. The score of the i^{th} feature S_i can be calculated by Fisher's score as in Eq. (6) [25]:

$$S_i = \frac{\sum_{j=1}^C n_j (\mu_{ij} - \mu_i)^2}{\sum_{j=1}^C n_j \times \rho_{ij}^2} \quad (6)$$

Where C is the total number of classes, n_j is the number of samples in the j^{th} class. μ_{ij} and ρ_{ij} are the mean and the variance of the i^{th} feature in the j^{th} class, respectively, and μ_i is the mean of the i^{th} feature in all classes.

HoG features are then sorted in ascending order concerning the Fisher score, and the first N_1 features are selected as a candidate set of features to contribute to the selection of the N_2 optimal features by the GA-NN feature selection method.

3.5.2. GA-NN feature selection

The GA-NN feature selection method is described with the following steps (as shown in Fig. 3):

1) The basic hyper-parameters of GA are first defined, which include population size, number of generations, chromosome length, cross-over rate, and mutation rate. The chromosomes with length (L) equal to the number of features selected by Fisher's score (N_1) are then initialized with binary encoding using Eq. (7).

$$Gene(i) = \begin{cases} 0, & r < 0.5 \\ 1, & r \geq 0.5 \end{cases} \quad (7)$$

Let $G=1$, steps (2-5) are repeated till G is greater than the maximum number of generations.

2) The population is evaluated using neural network

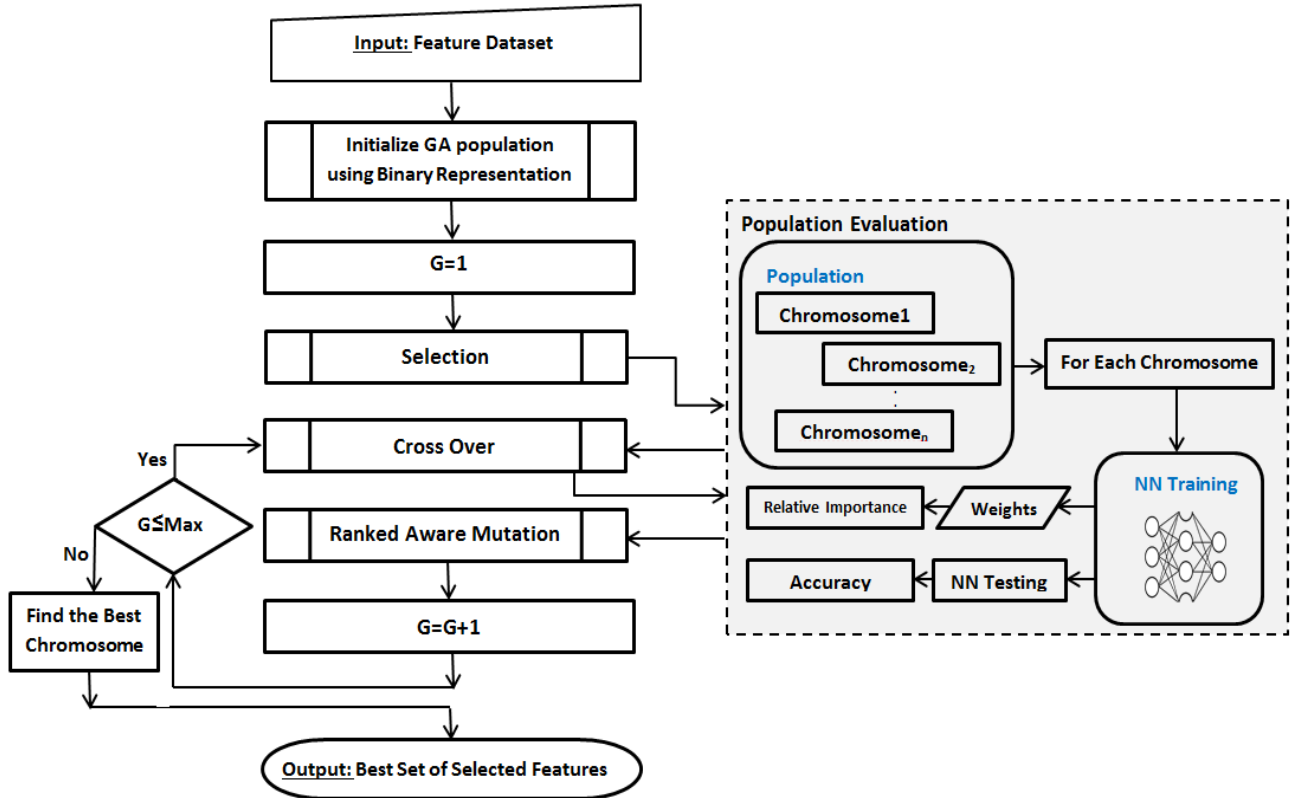


Figure. 3 A flow chart for the GA-ANN feature selection stage

accuracy as an objective function. The relative rank is then computed for each feature according to the weights resulting from the NN training. For each gene in the chromosome (feature) $j, j=1, 2, \dots, L$, the relative importance (RI_j) is calculated using Eq. (8) [26].

$$RI_j = \frac{\sum_{m=1}^{N_h} \left(\sum_{n=1}^{N_o} \left(\frac{|w_{jm}^{ih}|}{\sum_{k=1}^{N_i} |w_{km}^{ih}|} \right) * |w_{mn}^{ho}| \right)}{\sum_{p=1}^{N_i} \left(\sum_{m=1}^{N_h} \left(\sum_{n=1}^{N_o} \left(\frac{|w_{pm}^{ih}|}{\sum_{k=1}^{N_i} |w_{km}^{ih}|} \right) * |w_{mn}^{ho}| \right) \right)} \quad (8)$$

Where N_i and N_h are the numbers of input and hidden neurons, respectively, and w is the connection weight, the superscripts i, h , and o refer to input, hidden, and output layers, respectively, and the subscripts k, m , and n refer to the input, hidden, and output neurons.

The concept of transfer learning is utilized here, where the weights of trained NN are returned along with the RI and the weights. Transfer learning means taking the weights of the pre-trained NN models and adopting them in the next evaluation of the chromosome. In other words, the knowledge developed from previous training is recycled to help performing the next evaluation of the chromosomes.

3) The selection operator is then applied to select parent chromosomes from the population using the roulette wheel selection method.

4) To ensure exploration during the search for optimal features, a crossover operator is applied. The single-point crossover method is applied. The single point crossover method is utilized with a crossover cut point equal to $(L/2)$.

5) The mutation operator is utilized to achieve the exploitation of the solutions. A new ranked-aware mutation operator is proposed. The key idea of this operator is to mutate the gene (i) based on its $RI(i)$ value, as in Eq. (9).

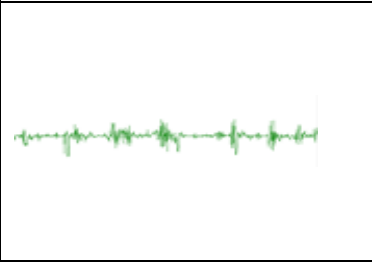
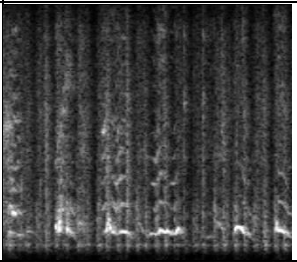
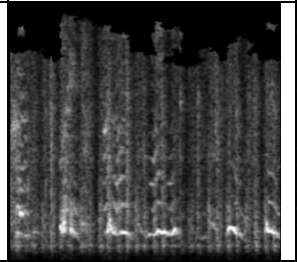
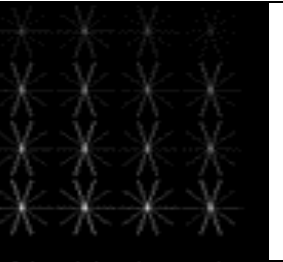
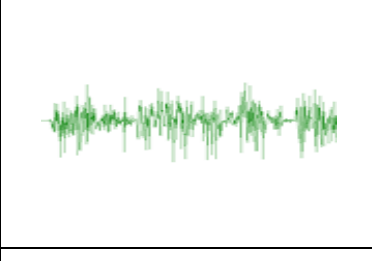
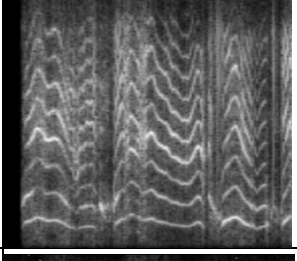
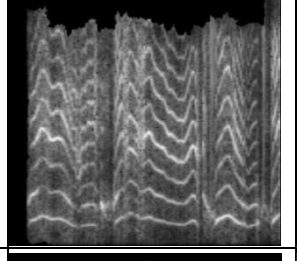
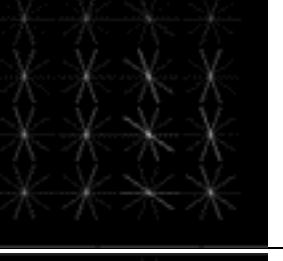
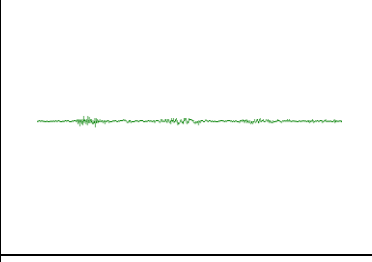
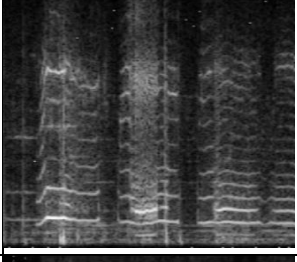
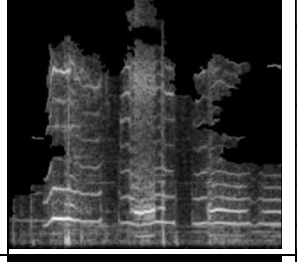
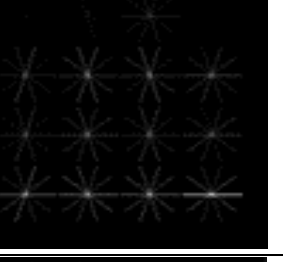
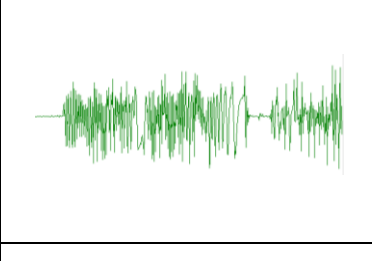
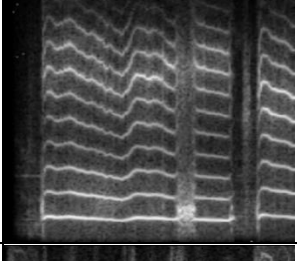
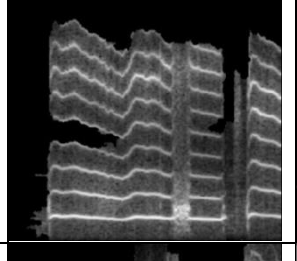
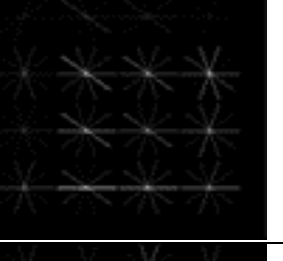
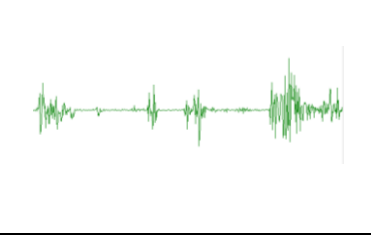
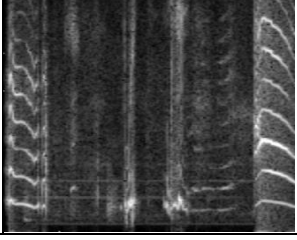
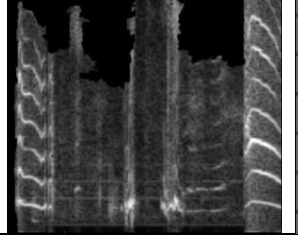
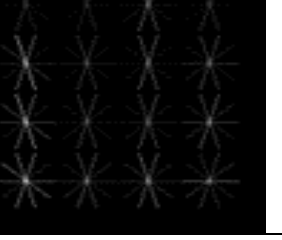
$$Gene(i) = \begin{cases} 1, & \text{if } Gene(i) = 0 \text{ and } RI(i) \geq T \\ 0, & \text{if } Gene(i) = 1 \text{ and } RI(i) < T \end{cases} \quad (9)$$

Where T is a threshold that is computed as:

Table 1. Setup of the fixed parameters of the proposed system

Parameter Name	Parameter Value
Frame Size	256
Frames Overlap	0.4
Spectrogram Image Width	400
Spectrogram Image Height	400
Crossover Rate	0.6
Mutation Rate	0.2
Population Size	20
Number of Generations	10
Number of Hidden Layers	11
Learn Rate	0.5
Momentum	0.5

Table 2. Results of spectrogram image generation, noise reduction and HoG features extraction stages

Cry Reason	Original Cry Signal	Spectrogram Image	segmented Image	HoG Features
Belly Pain				
Burping				
Discomfort				
Hungry				
Tired				

$$T = \frac{\sum_{k=1}^L RI(k)}{L} \tag{10}$$

L is the length of the chromosome, RI is the relative importance of the chromosome genes.

4. Experimental results and analysis

The main focus of the experimental results section is to test the validity of each stage of the

proposed system. First, the experimental setup of the important system parameters is presented. Then, the visual results of the first three stages of the proposed system are listed. Finally, the results achieved by the feature selection stage are given.

4.1 Experimental setup

The proposed system is implemented using the Skimage and OpenCV Python libraries for

Table 3. The different combinations of parameters' setup

Feature set	pixels_per_cell	cells_per_block	orientations
Set1	20	3	5
Set2	20	3	7
Set3	20	4	5
Set4	20	4	7
Set5	30	3	5
Set6	30	3	7
Set7	30	4	5
Set8	30	4	7

watershed segmentation and HoG feature extraction stages. The remaining stages of the system are implemented using Microsoft C# programming language.

Different experiments were conducted using different system parameters. Table 1 shows the values of parameters whose values are fixed during experiments. The effects of remaining parameters such as HoG parameters (*pixels_per_cell*, *cells_per_block*, *orientations*), optimal feature set using the Fisher score (N_1) and GA-NN-based feature selection are discussed in the next subsections.

4.2 Visual results

Table 2 shows the results obtained after generating the spectrogram image, applying watershed segmentation, and extracting HoG features. These results are generated using different signals for the five different crying reason classes (belly pain, burping, discomfort, hunger, and tiredness). HoG features are represented as a

grayscale image for easier understanding.

As shown by the images in Table 2, the spectrogram image differs from one cry class to another, and this promises the extraction of effective features. However, the spectrogram image contains some noisy and irrelative parts, which are reduced in the segmented image as presented in the fourth column of the table. On the other hand, the images in the final column of the table show different HoG feature patterns but with many redundant features (zero regions and some similar patterns in the HoG images).

4.3 Feature selection results

Two different values for each HoG parameter (*pixels_per_cell*, *cells_per_block*, and *orientations*) are attempted during the validation of the feature selection stage. The different permutations of the parameters values are grouped into eight sets, as listed in Table 3. First, we present in Fig. 4 the Fisher score values for the HoG features generated using the different combinations of these parameters.

On the other hand, Fig. 5 shows results for the GA-NN feature selection stage. Each figure depicts the accuracy of each chromosome within each GA generation. These chromosomes are populated within different feature sets, which are generated using different HoG parameters. The number of original feature sets fed to GA-NN, the number of selected features for the best chromosome within each feature set, and the obtained accuracy are listed in Table 4.

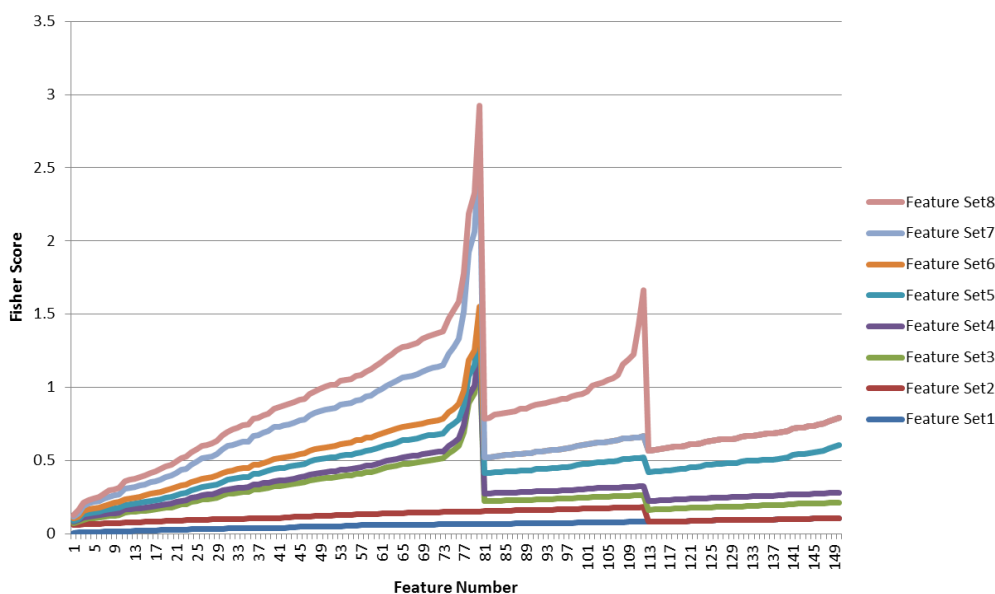


Figure 4. Fisher scores of features generated using different HoG parameters

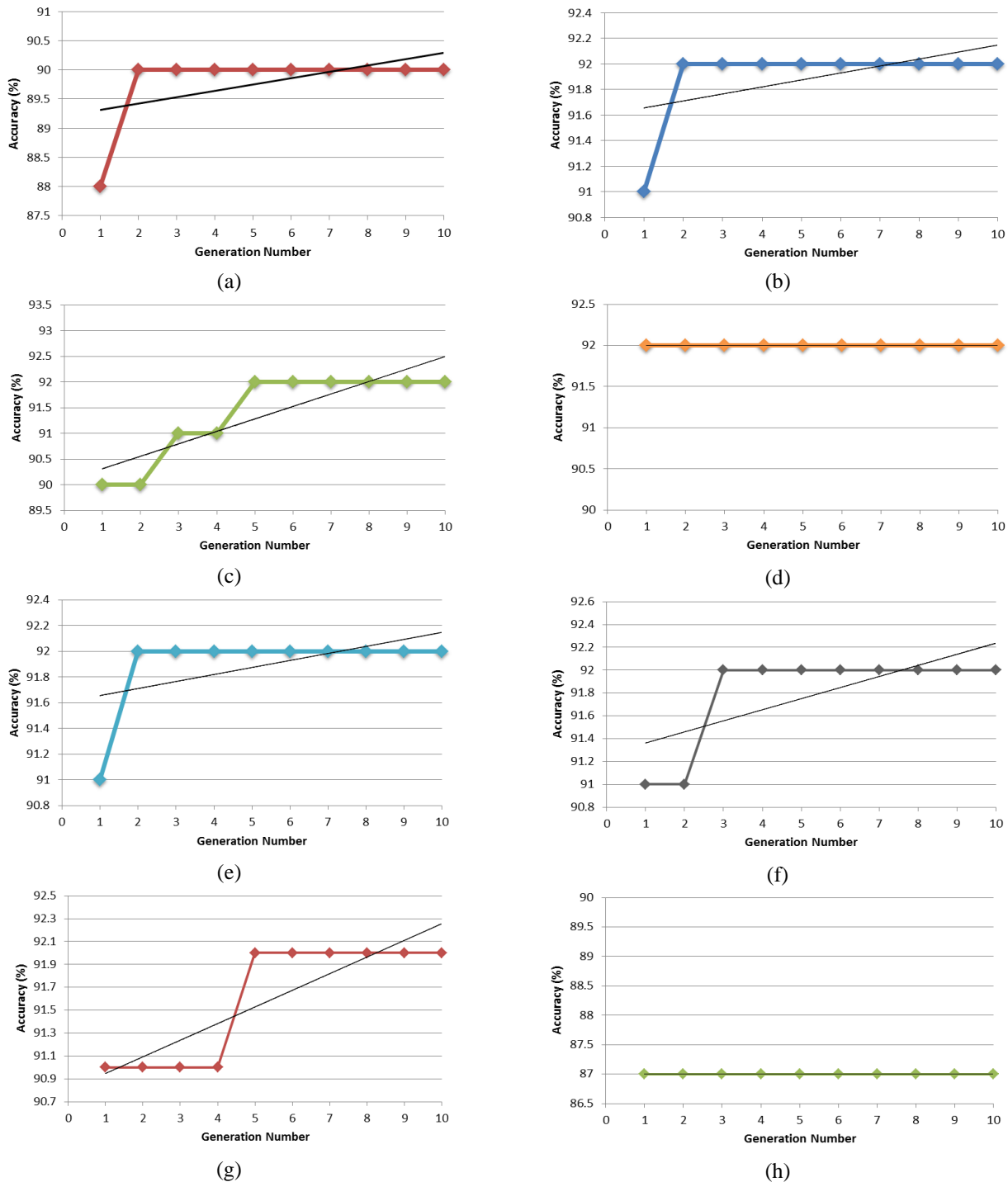


Figure 5 Accuracy achieved by the best chromosome during GA-NN feature selection: (a) Using feature Set1, (b) Using feature Set2, (c) Using feature Set3, (d) Using feature Set4, (e) Using feature Set5, (f) Using feature Set6, (g) Using feature Set7, and (h) Using feature Set8

The final experiment has been conducted by applying the proposed method without using a ranked-aware mutation operator. The goal of this experiment is to measure the effectiveness of the proposed mutation operator in comparison with the original GA operator. The results reached are illustrated in Fig. 6. The figure supports the effectiveness of the proposed mutation operator

because it can enhance accuracy by about 3%.

4.4 Comparison with previous works

In Table 5, we compare the results of the proposed method with state-of-the-art works that utilized the Donateacry-corpus dataset as a test

Table 4. Results achieved using GA-NN

Feature Set	Original Features	N_1	N_2	Best Accuracy
Set1	720	13	10	90%
Set2	504	51	23	92%
Set3	80	37	22	92%
Set4	1008	111	60	92%
Set5	180	55	30	92%
Set6	252	33	25	92%
Set7	80	37	24	92%
Set8	112	13	9	87%

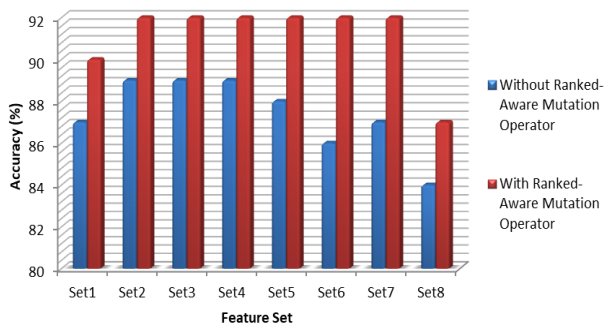


Figure. 6 Comparison between the best accuracy achieved by ranked aware GA-NN and original GA-NN for the eight different feature sets

Table 5. Comparison with other studies utilizing Donateacry-corpus dataset as a test material

Study	Method	No. of classes	Accuracy
[8]	MFCC and CNN	5	85%
[9]	MFCC, STFT and DNN	4	84.49% for MFCC and 85.59% for STFT
Proposed	HoG of Spectrogram image, fisher score and ranked-aware GA	5	92 %

material. The number of classes used by each work is also listed in the table to ensure a fair comparison. Although the proposed system is tested using all classes of the Donateacry-corpus dataset (i.e., the five classes), the results demonstrated in this table show the superiority of the proposed method in terms of classification accuracy by an enhancement of about 7% over the accuracy achieved by [8].

5. Conclusions and future work

Identifying the cause of an infant's crying is one of the challenges that parents and babysitters face. The purpose of this study is to assist such people in

determining the reason. Treating the information provided in the cry sound signal as a spectrogram image allows for the extraction of image-relevant features. The increased dimensionality of the resulting HoG features was decreased using the Fisher score and GA-NN. The suggested ranked-aware mutation operator effectively flips the bit based on the feature rank in the training network. Inheriting the weights from the previous training phase within the same GA generation allowed the model to complete the feature selection based on his background information. The experiments conducted yielded promising results for determining the cause of infant cries utilizing image features, the GA algorithm, and the transfer learning concept. Future work will make use of deep learning techniques in conjunction with the methods discussed in the paper.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Suhaila N. Mohammed and Adnan J. Jabir; methodology, software, and validation, Suhaila N. Mohammed and Adnan J. Jabir; formal analysis, Adnan J. Jabir; investigation, Adnan J. Jabir; resources, Suhaila N. Mohammed; data curation, Adnan J. Jabir; writing—original draft preparation, Suhaila N. Mohammed; writing—review and editing, Adnan J. Jabir; visualization, Suhaila N. Mohammed; supervision, project administration and funding acquisition, Adnan J. Jabir.

References

- [1] C. lakshmi, B. Aravinda, and D. Sadhana, "Predicting the Reason for the Baby Cry Using Machine Learning", *Journal of Artificial Intelligence, Machine Learning and Soft Computing*, Vol. 4, No. 1, pp. 11-25, 2019.
- [2] T. Fuhr, H. Reetz, and C. Wegener, "Comparison of Supervised-learning Models for Infant Cry Classification", *International Journal of Health Professions*, Vol. 2, No. 1, pp. 4-15, 2015.
- [3] A. Warlaumont, D. Oller, and E. Buder, "Data-Driven Automated Acoustic Analysis of Human Infant Vocalizations Using Neural Network Tools", *J. Acoust. Soc. Am.*, Vol. 127, No. 4, pp. 2563-2577, 2010.
- [4] H. Alaie, L. A. Abbas, and C. Tadj, "Cry based infant pathology classification using GMMs",

- Speech Communication*, Vol. 77, pp. 28–52, 2016.
- [5] T. Maghfira, T. Basaruddin, and A. Krisnadhi, “Infant Cry Classification Using CNN – RNN”, In: *Proc. of 4th International Seminar on Sensors, Instrumentation, Measurement and Metrology, Journal of Physics: Conference Series*, Vol. 1528, pp. 1–6, 2020.
- [6] P. Pathirana and S. Sumathipala, “A Low-Cost Intelligent Hardware System for Real-Time Infant Cry Detection and Classification”, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Vol. 8, No. 12S2, pp. 18–22, 2019.
- [7] P. Rani, P. Kumar, V. Immanuel, P. Tharun, and P. Rajesh, “Baby Cry Classification Using Machine Learning”, *International Journal of Innovative Science and Research Technology*, Vol. 7, No. 3, pp. 677–681, 2022.
- [8] E. Sutanto, F. Fahmi, W. Shalannanda, and A. Aridarma, “Cry Recognition for Infant Incubator Monitoring System Based on Internet of Things using Machine Learning”, *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 1, pp. 444–452, 2021, doi: 10.22266/ijies2021.0228.41.
- [9] J. Cha and G. Bae, “Deep Learning Based Infant Cry Analysis Utilizing Computer Vision”, *International Journal of Applied Engineering Research*, Vol. 17, No. 1, pp. 30–35, 2022.
- [10] N. Wahid, P. Saad, and M. Hariharan, “Automatic Infant Cry Classification Using Radial Basis Function Network”, *Journal of Advanced Research in Applied Sciences and Engineering Technology*, Vol. 4, No. 1, pp. 12–28, 2016.
- [11] N. Salman and S. Mohammed, “Image Segmentation Using PSO-Enhanced K-Means Clustering and Region Growing Algorithms”, *Iraqi Journal of Science*, Vol. 62, No. 12, pp. 4988–4998, 2021.
- [12] Y. Wu and Q. Li, “The Algorithm of Watershed Color Image Segmentation Based on Morphological Gradient”, *Sensors*, Vol. 22, pp. 1–23, 2022.
- [13] A. Kaur and Aayushi, “Image Segmentation using Watershed Transform”, *International Journal of Soft Computing and Engineering*, Vol. 4, No. 1, pp. 5–8, 2014.
- [14] J. Roerdink and A. Meijster, “The Watershed Transform: Definitions, Algorithms and Parallelization Strategies”, *Fundamenta Informaticae*, Vol. 4, pp. 187–228, 2001.
- [15] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection”, In: *Proc of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, United States, pp. 886–893, 2005.
- [16] J. Soler, H. Beuther, M. Rugel, Y. Wang, P. Clark, S. Glover, P. Goldsmith, M. Heyer, L. Anderson, A. Goodman, and P. Schilke, “Histogram of Oriented Gradients: A Technique for the Study of Molecular Cloud Formation”, *Astronomy & Astrophysics*, Vol. 622, No. A166, pp. 1–31, 2019.
- [17] E. G. Ramírez, A. García, E. G. Ramírez, A. O. Molina, O. Ramírez, and I. Arroyo, “Multi-object Recognition Using a Feature Descriptor and Neural Classifier”, *Book Chapter, Vision Sensors - Recent Advances*, intechopen, 2022.
- [18] S. Katoch, S. Chauhan, and V. Kumar, “A review on genetic algorithm: past, present, and future”, *Multimedia Tools and Applications*, Vol. 80, pp. 8091–8126, 2021.
- [19] A. Tettamanzi and M. Tomassini, *Soft Computing Integrating Evolutionary, Neural, and Fuzzy Systems*, 1st edition, Springer-Verlag Berlin Heidelberg, 2001.
- [20] https://github.com/gveres/donateacry-corpus/tree/master/donateacry_corpus_cleaned_and_updated_data, visited on: 5/12/2022.
- [21] S. Mohammed and A. Hassan, “Speech Emotion Recognition Using MELBP Variants of Spectrogram Image”, *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 5, pp. 257–266, 2020, doi: 10.22266/ijies2020.1031.23.
- [22] G. Naithani, J. Kivinummi, T. Virtanen, O. Tammela, M. Peltola, and J. Leppänen, “Automatic Segmentation of Infant Cry Signals Using Hidden Markov Models”, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 1, pp. 1–14, 2018.
- [23] S. Mohammed and A. Hassan, “Automatic Voice Activity Detection Using Fuzzy-Neuro Classifier”, *Journal of Engineering Science and Technology*, Vol. 15, No. 5, pp. 2854 – 2870, 2020.
- [24] C. Ji, T. Mudiyansele, Y. Gao, and Y. Pan, “A Review of Infant Cry Analysis and Classification”, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 8, pp. 1–17, 2021.
- [25] C. Aggarwal, “Data Classification Algorithms and Applications”, *CRC Press*, 2015.
- [26] S. Andreas, P. Harris, and B. Max, *Artificial Intelligence Applications and Innovations*, Springer, 2018.