Real Time Audio-Based Distress Signal Detection as Vital Signs of Myocardial Infarction Using Convolutional Neural Networks

H. M. Mohan and S. Anitha
Department of ECE, ACS College of Engineering, Bangalore, India
Email: {mohanhm, dranithasammilan}@gmail.com

***Abstract*—In recent years, with rapid advancement in Artificial Intelligence technology, several intelligent systems have been developed for human emergency prediction under ambient intelligence. Automatic pain recognition through state-of-the-art deep learning algorithms has attracted much attention recently in smart healthcare informatics. This research presents a Convolutional Neural Network (CNN) approach for detecting audio-based emergency identification during Myocardial Infarction. For evaluation, simulated emergency distress audio signals are recorded during possible myocardial infarction as a private dataset. This work demonstrates an approach to train the deep learning CNN model for multiclass audio samples and deploy it on an edge embedded Artificial Intelligence device Jetson Nano for real-time recognition.**

***Index Terms*—ambient intelligence, myocardial infarction, convolutional neural network, emergency distress signal**

I. Introduction

Ambient intelligence is an emerging healthcare application domain that intends to create a smart environment for improved wellness and quality of life through the state-of-the-art technologies such as Artificial Intelligence (AI), IoT, 5G, etc. The concept of Ambient Assisted Living (AAL) delivers intelligent services for remote health monitoring and medical automation at various living spaces such as homes, offices, hospitals, public transport by deploying smart sensors, edge devices, wireless networks and software applications [1]. Data is acquired by various sensory modalities like embedded sensors, wearable devices, mobile phones, surveillance cameras and processed through advanced AI algorithms allows to analyze the emergencies or critical health issues amidst daily routine activities by offering immediate health diagnostics and treatment. Recently healthcare domain is empowered with Edge Intelligence in which Edge/Fog computing architectures and powerful AI techniques are collaboratively utilized to develop efficient, responsive and context-aware edge computing platform aiming at promoting a healthier society [2]. Researchers have implemented fall detection [3], voice disorder detection [4], skin cancer detection [5] and automatic triage screening system [6] as smart health care applications.

Myocardial Infarction (MI) is among the highest causes of human fatalities across the globe [7]. A recent report of the American Heart Association suggests that the heart disease rate may increase to 23.3% worldwide by the year 2030 [8]. Continuous monitoring of cardiac patients and instant treatment during emergency health situations can avoid fatal consequences of life risk. Chest pain is the prime presenting symptom of a heart attack occurring in almost 80% of the patients. The chest pain in patients during MI experiences a radiated pain in the jaw, shoulder, arm, or back [9]. A person encountering intense or stabbing chest pain places a clenched fist in the center of the chest called Levine's sign is an indicative nature of possible MI [10], [11]. This responsive symptom of heart attack is sudden and intense which makes the patient easy to recognize and get immediate care and treatment.

Single-phrase speech identification systems are categorized into three main segments based on application areas such as i) Keyword Spotting (KWS), ii) Wake-up-Word (WUW) detection, and iii) Spoken Content Retrieval (SCR) [12]. Keyword identification technology is an automatic technique to spot particular predefined keywords in a continuous stream of speech and voice. Keyword spotting systems are designed as cost-effective, low power consumption, small memory footprint, highly accurate application deployed on low-resource devices like embedded edge devices, mobile phones, and tablets.

Ambient intelligence offers a more secure and safer living environment for the elderly and cardiac patients through emergency response systems, video surveillance, and fall detections systems. Research related to automatic emergency surveillance systems is predominantly carried out through video and audio streams. When compared to video-based approaches, audio-based surveillance offers several advantages such as smaller data size, acoustic sensory devices can acquire quality data in both day and night conditions and acoustic information presents the spherical field of view [13]. The dynamic audio signals carry vital information regarding human emergency situations in living spaces such as offices, homes, elevators, public transport vehicles, etc. Lately, considerable research works have been developed based

Manuscript received September 26, 2021; revised January 11, 2022.

doi: 10.12720/jait.13.2.106-116

on audio emergency detection systems such as audio distress recognition [14], human screams in urban areas [15], scream and gunshot in metropolitan cities [16].

This paper extends our previous work on non-invasive technique for real-time myocardial infarction detection using faster R-CNN [17]. The present work employs state-of-the-art CNN lightweight architecture mainly designed to meet the real-time implementation on GPU edge devices such as lesser model size, efficient energy consumption, faster computation time, and considerably good accuracy. The automatic detection of vital signs of MI through audio distress signals is an arduous task since the devised intelligent system has to discriminate between speech and different pain emotions to reduce false alarms. The main aim of this research work is to devise an audio-based intelligent system to recognize distress signals during MI in the indoor environment with a low-power embedded GPU edge device commingled with the power of deep learning.

The main contributions of this research work are summarized as follows:

i) Our study presents a novel framework for audio distress signal identification of a person during MI as vital signs of MI.
ii) We created a synthetic private dataset of audio streams simulated for a cardiac arrest situation.
iii) We propose a lightweight deep learning neural network classification model and deploy it on an Edge GPU device for real-time inference evaluation.

Remainder of the paper is structured as follows: section II highlights the recent works carried out. Section III elaborates the proposed algorithm and the system architecture. Section IV delivers the experimental results, section V explains the embedded edge implementation, section VI brings out the discussion and finally, section VII highlights the future scope and conclusion of our work.

II. Related Work

This section provides basic information on various audio-based keyword detection approaches and pain detection researches in recent years.

Traditional approaches like Gaussian hidden Markov models and Hidden Markov Models (HMM) was popularly used during the previous decade in keyword spotting systems. Rohlicek *et al.* proposed a word prediction system based on Gaussian hidden Markov models and estimated via linear prediction and cepstrum spectral envelopes. The authors observed that choice of features, normalization, specific techniques for modelling and scoring metrics can have crucial implications on the performance of the speech model [18]. The authors advocated an HMM-based keyword recognition system depending upon a continuous speech recognition model. A detailed investigation is performed based on non-vocabulary speech and identified that difficulty exists in the size of keyword and characterization of the non-keyword in a conversational speech. The problem has been addressed explicitly using filler models trained from transcribed continuous speech [19]. Another work on HMM-based keyword spotting based on speech signals in telephone networks was performed to recognize keywords in unconstrained speech from a pre-defined vocabulary list. Evaluations on three benchmark databases were carried out: Stonehenge Road Rally database, a five-word vocabulary used to automate operator-assisted calls, and a three-word Spanish vocabulary that is currently being trialed in Spain's telephone network [20]. In the work [21] the authors provide an alternative solution to detect keywords from unconstrained speech using the Viterbi decoding algorithm. The approach uses a Viterbi matching technique and time normalization for the keywords from BREF database to arrive at an effective solution. The authors [22] identify the complexity in Viterbi matching and time normalization procedure and uniquely proposed a twofold approach for keyword spotting: i) Optimal segmentation technique for maximizing the average observation probability wherein the scoring for begin/endpoint is eliminated. ii) Classification step to categorize as keyword or not based on a predefined threshold. David Grangier *et al.* uniquely proposed a large margin and kernel approach for keyword spotting to overcome the problems of traditional HMM methodologies. The aim of the technique is to map the input acoustic characterization speech utterance with the target keyword into a vector space. The experiments were performed using the TIMIT corpus speech database and achieved an improved performance over HMM-based systems [23]. This discriminative approach concept is later extended by the authors [24] and an evolutionary algorithm is utilized to classify sentences that contain keywords or not. A probabilistic approach of identifying phenomes in speech frames based on acoustic, spectral, and statistical features is achieved at low computational complexity.

Lately, deep learning techniques have emerged as a predominant approach for wake-up word spotting and distress recognition. The authors [25] implement an automatic classification of speech pronunciations using Extreme Learning Machine (ELM), a neural network approach. Power Normalized Cepstral Coefficients is used as feature vector representation and length normalization is performed to express as fixed-length sequences. Length normalization using Dynamic Time Warping (DTW) and classification through ELM neural network classifier achieved superior performance. Fengpei Ge *et al.* presented a novel approach of Automatic Wakeup-Word Speech Recognition (AWUWSR). A two-level classification scheme is adopted to combine phonetic knowledge and DNN model-based classifier for wake-up word identification. An accuracy of 90% was achieved with low false alarm rate [26]. The authors propose a novel anthropogenic disaster recognition framework for classifying 10 emergency sounds such as gunshot, explosion, scream cry etc. Initially, acoustic feature representations are performed, later probabilities distances between extracted sounds are exploited and in the final stage, a hierarchical

classification algorithm is employed for classification [13]. Quang Nguyen *et al.* worked on human screams emergency situation audio samples utilizing a perception sensor network. The audio samples were acquired through Kinect microphone and an audio-visual integration technique is adopted to identify a single speaking person among multiple ones. Experiments were performed with a robot under lab conditions for a rescue operation under emergency scream situations [27]. The authors devised a novel bag of words procedure for emergency situations spotting through audio surveillance. The methodology is tested for low and high noise levels in diverse conditions of real environments with background noise. This method successfully achieved low SNR with high robustness [28]. Emanuele Principi *et al.* developed ambient assistance for elderly people based on an acoustic signal emergency system. Day-to-day human monitoring of voice recognition and emergency situation analysis is performed and implemented using low power embedded platform [29]. The authors built a speech command detection model that can effectively detect predefined keywords. Google's Tensor Flow speech dataset was modeled using three neural network architectures: Vanilla Network, Deep Neural Network, and Convolution Neural Network. The experimental results highlight that CNN architecture was more efficient in keyword recognition and classification with a cost of the increased number of parameters [30]. The authors DNN based Syllable recognition and keyword detection framework for a dataset of Hindi regional language. The methodology is twofold: In the initial step neural network model is trained to extract hierarchical non-linear transformation features and later is fused with LSTM network and subsequently fine-tuned with auto-encoder network [31]. The authors designed AI-based audio distress signal recognition in metropolitan city areas as a real-time emergency alarm system. A private dataset collected under real-world scenarios is trained up with a deep Convolution Neural network algorithm and later deployed to a low-cost Raspberry Pi portable device [14]. Zehetner *et al.* devised WUW spotting in smart mobile devices as an alarm system for detecting personalized keywords under real-time audio streaming [12]. Jivitesh Sharma *et al.* adopt a unique approach for the Environment Sound Classification Task (ESC) for three benchmark environment sound classification datasets namely the UrbanSound8K, ESC-10, and ESC-50. The aim was to address the issues concerning feature extraction, computational complexity, and complex CNN architectures. The challenges are overcome by using multiple features, data augmentation, depth-wise convolutions in the CNN network, and attention mechanism [32]. Improvisation of the work was demonstrated to achieve a state-of-the-art performance by employing multiple audio feature extraction approaches like the mel-frequency cepstral coefficients, gamma tone frequency cepstral coefficients, constant Q-transform and chromagram. The work successfully reduced the size of the model thereby achieving an emergency sound detection accuracy of an impressive 99.56% [33].

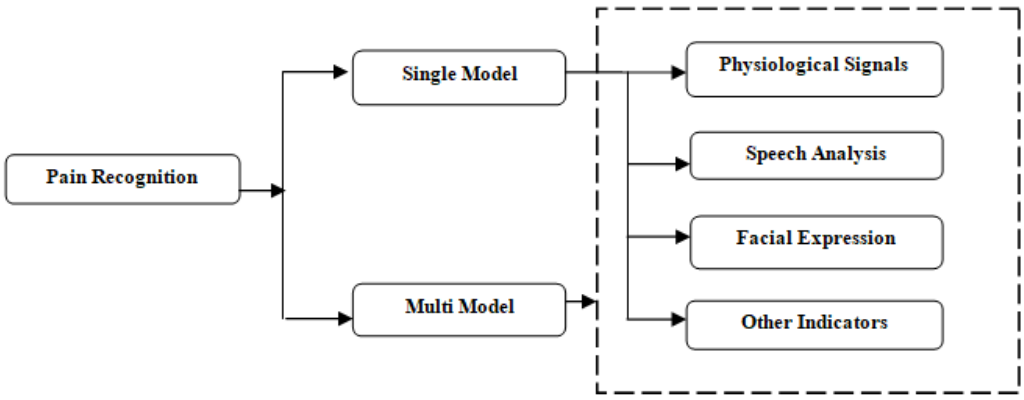


Figure 1. Pain recognition categorization in deep learning networks.

In recent years the deep learning approach are utilized to solve the arduous problem of pain recognition in human beings. Fig. 1 shows different approaches to pain identification using deep learning models [34]. Table I highlights the studies carried out lately considering different modalities, pain databases, feature extraction techniques, and deep learning models. Lopez-Martinez *et al.* developed a novel technique of identifying pain intensity levels in patients considering the skin conductance and electrocardiogram (ECG) parameters. A Multitask Neural network framework was designed and trained up using the BioVid heat pain database to evaluate pain from physiological responses [35]. The authors proposed a multimodal approach of evaluating the intensity of pain by combining physiological signals namely: ECG, Electromyography (EMG), and Epidural Analgesia (EDA) from Part A of the BioVid heat pain database. Various Convolutional Neural Networks models based on 1D and 2D convolutional layers are developed for pain classifications to outperform traditional machine learning techniques [36]. The authors uniquely adopted an automatic protective behavior detection using a deep learning-based wearable device. The authors made a study on behavioral moment patterns due to fear of pain at rehabilitation centers based on the kinematic and EMG data. Investigations based on RNN LSTM and dual-stream LSTM were performed and achieved an F1 score of 81.5% [37]. Rodriguez *et al.*

performed an investigation on pain intensity based on facial expression analysis. The authors devised a CNN and LSTM two-stage deep learning framework to automatically learn the features from raw images and to extract the temporal relation from video frames [38]. The authors advocated an improvised CNN-based architecture for automatic feature extraction and regression learning of facial pain features by using cumulative attributes as intermediate features. The architecture was trained using UNBC-McMaster Shoulder Pain Expression Archive database and achieved improved performance in facial Action Unit (AU) intensity estimation [39]. Boyi Hu *et al.* proposed a neural network approach of evaluating the Lower back pain problems by exploiting Kinematic and Motion sensor data. The authors proved that through the data fusion approach classification performance of the DNN could be improvised [40].

Through the extensive literature survey the authors find out that considerably less work has been carried out in pain recognition through the human speech approach. Tsai *et al.* carried out a novel work in estimating the pain intensity level of patients from emergency department patients during triage. A Long-Short Term Memory neural networks (LSTMs) framework was developed to classify the vocal features of real patients [41].

TABLE I. PAIN-DETECTION RESEARCHES BASED ON DEEP LEARNING

Study	Features	Deep Learning models	Modalities	Dataset
LopezMartinez [35]	Classification	Multitask Neural network	Skin conductance and ECG	BioVid Heat Pain database
Thiam [36]	• Feature Extraction • Classification	CNN	EDA, ECG, EMG	BioVid Heat Pain database
Wang [37]	Classification	LSTM	Kinematic data, EMG	BioVid Heat Pain database
Rodriguez [38]	• Feature Extraction • Classification	• CNN • LSTM	Face	UNBC-McMaster Shoulder Pain Expression Archive database
Jaiswal [39]	Classification	CNN	Face	UNBC-McMaster Shoulder Pain Expression Archive database
Boyi Hu [40]	Classification	LSTM	Kinematic data: Motion sensors	Private database- healthy people:22; LBP patients:22
Tsai [41]	Feature Extraction	LSTM	Audio and video from handy camera	Triage Pain-Level multimodal database and Chinese corpus: The DaAi database

III. PROPOSED METHODOLOGY

A. Overview of Proposed Model

In early 1950s David Hubel and Torsten Wiesel performed classic experiment to understand neuronal activity in the visual cortex of a cat [42]. The research revealed that the visual part of the frontal cortex of a Cat's central nervous system is formed by simple and complex cells alternatively and in control of detecting light in receptive fields. This famous cat experiment inspired Kunihiko Fukushima to design a multi-layered neural network model in 1979 named as Neocognitron which is regarded as the predecessor of Convolutional Neural Network [43]. In 1989, LeCun *et al.* developed a groundbreaking modern framework for CNN for the application of Handwritten Zip Code Recognition [44].

B. Convolutional Neural Network

Convolutional neural network belongs to a class of feed-forward deep learning architecture which are predominantly used to solve complex problems in various fields relevant to the computer vision domain. During recent years various adaptations of CNN architectures are developed including AlexNet, VGGNet, ResNet, GoogleNet, MobileNet, DenseNet, etc. However, the common structure of CNN remains the same consisting of convolutional layers, pooling (subsampling) layers followed by fully connected layers. Fig. 2 shows a generalized architecture of convolution neural network.

1) Convolution layer

The convolution layer is a prominent section of the CNN architecture that plays a vital role in feature extraction. In the Initial stage of computation, a kernel matrix known as a filter or a feature detector is used to extract useful features from the image. A convolution operation between kernel matrix values and each pixel value in the input image matrix is performed to produce a resulting output of a feature map or 2D activation map. The architecture of convolutional neural network model designed in our work is shown in Fig. 3(a). The filter is slid across the input image pixel values and its scalar product is obtained as shown in Fig. 3(b).

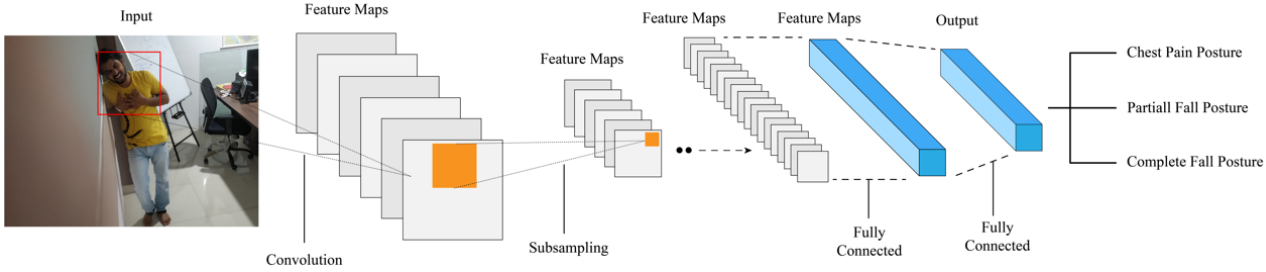
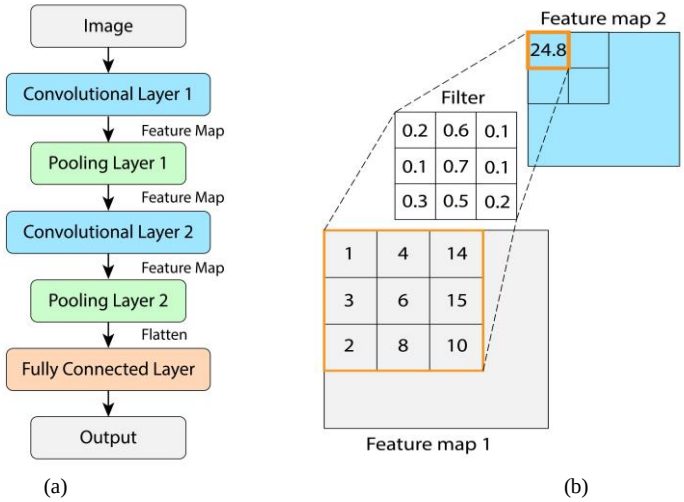


Figure 2. Convolution Neural network architecture.



(a) (b)

Figure 3. (a) Convolution neural network model (b) Convolution operation.

The dimension of input to a convolution layer is given as $H \times W \times C$. Where, H is the height and W is the width of the image and C indicated the number of channels. Value of C=1 for a greyscale image and C=3 for an RGB image. If K is considered as the number of filters F, stride as S and total zero padding as P, the output dimension of convolution layer is estimated as, $W2 \times H2 \times C2$ where,

$$W2 = \left[\frac{W1-F+2P}{S}\right] + 1 \quad (1)$$

$$H2 = \left[\frac{H1-F+2P}{S}\right] + 1 \quad (2)$$

$$C2 = K \quad (3)$$

2) Zero-Padding

Padding is the technique of padding the border of the input matrix symmetrically with zeros. This process is performed to maintain the same spatial dimensionality of both input and output size. The spatial size of the output volume of the next layer is a given as in (4):

$$O = (W - F + 2P)/\mathrm{S} + 1 \quad (4)$$

where, W implies volume size of input image, F indicates receptive field size of the convolutional layer, S is the stride, P is the amount of zero padding. The value of zero padding to obtain same spatial size of input output volume is calculated as in (5):

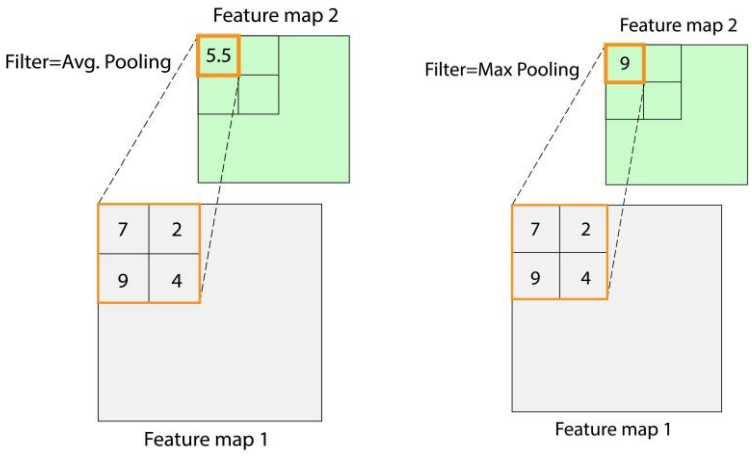
$$P = (F - 1)/2 \quad (5)$$

3) Hidden layer activation function

The result of convolution layer operation is passed through a nonlinear activation function. Non-linear activation introduces nonlinear properties to the neural network model which are vital for pattern extraction from non-linear complex data such as images, audio, video datasets. A precise choice of activation function in hidden layers is required for efficient feature extraction in a neural network model. Various choices of activation functions are made such as Rectified Linear Activation (ReLU), Hyperbolic Tangent (Tanh), Logistic (Sigmoid), etc. ReLu is mostly preferred as a default function in CNNs for its increased speed and efficiency. The ReLU activation function is determined as,

$$a_{i,j,k} = max(z_{i,j,k}, 0) \quad (6)$$

where, $z_{i,j,k}$ is the input to the function at location (i, j) on k-th channel. If the value of z is negative, then a value 0 is returned else the same positive value is retained.



(a) (b)

Figure 4. (a) Average pooling (b) Max pooling.

4) Pooling layers

The pooling layer or subsampling layer is placed after convolution layers to perform a down sampling operation to lower the feature map dimensionality. The main aim of pooling layers is to reduce the number of parameters and to improve computational efficiency while preserving the features. Different types of pooling layers are implemented such as distance weighted average pooling max pooling, average pooling and L2 norm pooling. Average pooling and max pooling are popularly used to retain the maximum feature response within a given sample size in a feature map. The concept of average pooling and max pooling are indicated in Fig. 4(a) and (b) respectively.

5) Fully connected layer

The output feature maps from the pooling layer is converted into a data format of 1-dimensional array or vector through a technique of flattening. The feature generated is connected to a classification model known as fully connected layers or dense layers. A fully connected layer is a Neural Network layer formed by numerous neurons wherein neurons in adjacent layers share connections for the final classification process. The total neuron in the output layer equals the number of classes to be identified by the multi-class classification neural network model.

6) Last layer activation function

The activation function at the output layer will decide the prediction of the convolutional neural network model. A choice of appropriate activation function depends on the type of prediction problem. For a classification task, commonly used activation functions are logistic (sigmoid) and softmax. For binary classification problems, a logistic function is used and for a multi-classification, the softmax activation function is utilized. In the present work, the softmax function is used for the multiclass classification problem.

C. The Proposed Method

The block diagram of the proposed methodology is shown in Fig. 5 [45]. The entire process is divided into three stages: i) Stage 1: Input stage consisting of data collection and preprocessing the audio data samples ii) Stage 2: Training stage incorporates Convolution Neural Network model for training the preprocessed dataset. iii) Stage 3: Detection stage wherein the vital signs of MI model deployed in Jetson Nano detects the distress sound and classifies the results. The block diagram of convolutional neural network model for training and classification is as shown if Fig. 6.

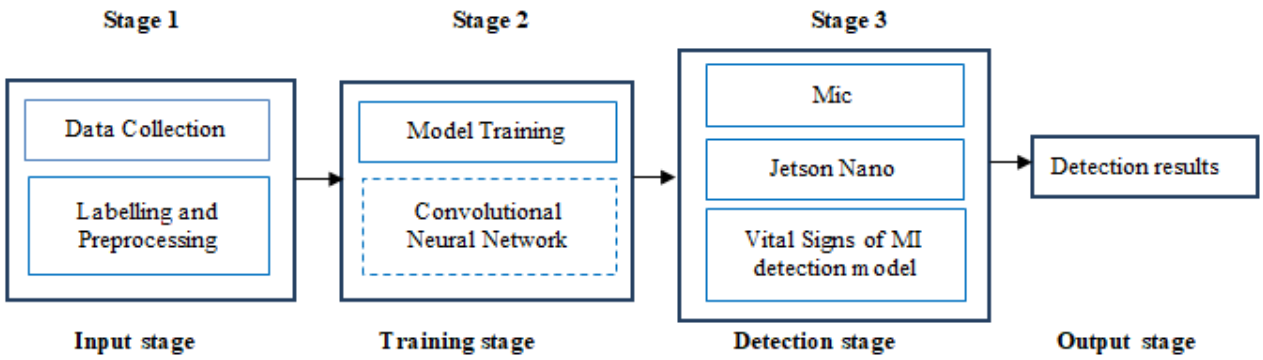


Figure 5. Method for training and detecting vital signs of MI.

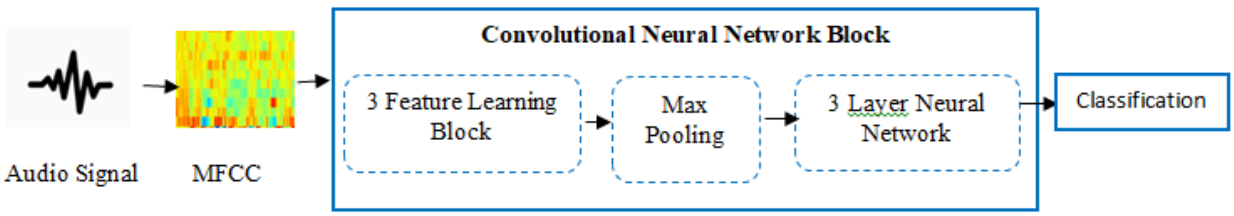


Figure 6. Block diagram of training and classification of audio samples using CNN.

1) Data set collection and preprocessing

One of the most significant challenges researchers face in automatic pain recognition is collecting the quality dataset adhering to the benchmarking standard. Simulating the pain screaming voices crying for help during the possible MI are hard to collect. The authors have made prominent efforts in gathering the audio dataset from the subjects of Kannada and English-speaking persons. An almost equal ratio of male and female subjects has been gathered and variations in the voice tone are contemplated. The main objective in collecting datasets from participants is to develop a deep learning model to generalize for all age groups and voice tones. Initially, we trained 60 people to propose expressions in the Kannada language that they would scream in a situation of heart attack situation where they required help in an indoor environment. The present research work takes into consideration keywords from Kannada and English language consisting of four classes namely: i) Ammaa ii) Kapadi (synonym of help) iii)

Aaaah (Painful scream) iv) Help. Various background sounds were included in the dataset categorized as: low, medium, and high. The type of pain simulated is pain experienced during MI. A total of 360 total audio samples are collected and to maintain uniformity in the dataset 80 samples are taken from each class.

2) Preprocessing

Audio features are mainly categorized into two classes as time-domain features and frequency-domain features. The feature representation of time-domain signals involves short-term energy of the signal, zero-crossing rate, maximum amplitude, minimum energy, and the entropy of energy. Such parameter analysis helps in better understanding audio and speech signals. Frequency domain parameters comprised of spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral roll-off, spectral entropy, and chroma coefficients help in better unveil of embedded features in the audio signals under limited data.

MFCCs was first implemented in 1980 by Davis and Mermelstein for parametric representations for monosyllabic word recognition and have been used by most widely used by researchers ever since. MFCCs have increasingly applied to audio and speech-based applications such as human emotion recognition, secure voice applications, speaker biometric applications, automatic recognition of numbers into a telephone and audio similarity measures, etc., MFCC is an effective pre-processing technique for extracting quality features from an audio signal which is represented by a set of cepstrum coefficients. MFCCs feature vectors are favored in speech recognition applications because of its ease in computation, compact representation of the spectral structure, uncorrelated parameterization, and being computationally effective. The audio features of each signal are expressed as MFCC parameters which are representations of images. These images act as an input to the CNN deep learning algorithm for the classification process. Fig. 7 shows the Mel spectrogram of the five different input audio classes considered in the present work.

A frequency-to-mel transform function for a frequency *f* is determined using (7),

$$\mathrm{m} = 2595 \log_{10}(1 + \frac{\mathrm{f}}{700}) \quad (7)$$

The inverse transform can be represented as in (8),

$$f = 700(10^{\frac{m}{2595}} - 1) \quad (8)$$

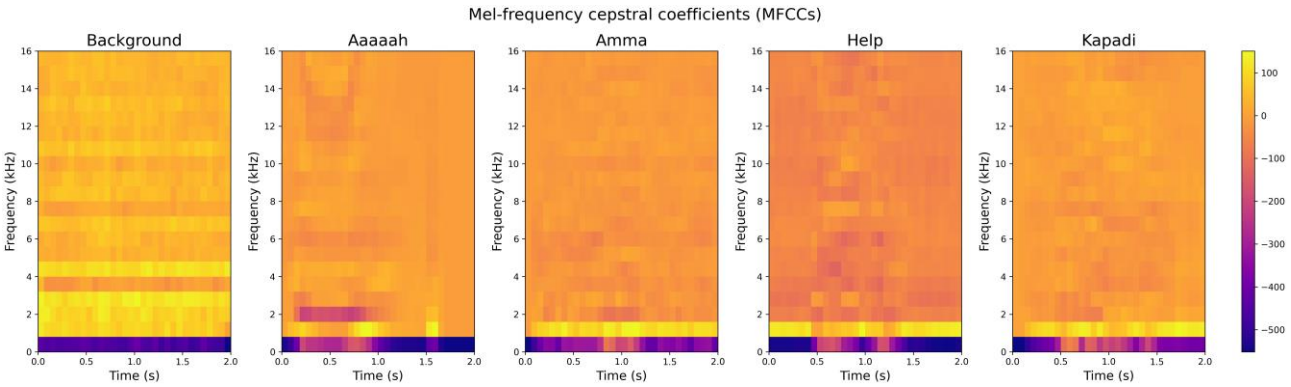


Figure 7. Mel spectrogram in a window of 2 seconds for each class.

3) Performance evaluation metrics

a) Confusion matrix

A) confusion matrix is commonly used to determine the quality of performance of a classifier model. There are four main cases indicated in the confusion matrix table i) True Positives (TP): results when the classification model predicted true and correct class was true. ii) True Negatives (TN): results when the model predicted false and correct class was false. iii) False Positives (FP) (Type I error): indicates cases when the model predicted true but correct class was false. iv) False Negatives (FN) (Type II error): represents classifier prediction as false, but the correct class was true. Fig. 8 illustrates the confusion matrix. The performance metrics classification accuracy, Precision and recall are calculated using the formula as in (9), (10), and (11).

$$\text{B) Classification Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$\text{C) Precision} = \frac{TP}{Total\ true\ predictions} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{D) Recall} = \frac{TP}{Actual\ True} = \frac{TP}{TP+FN} \quad (11)$$

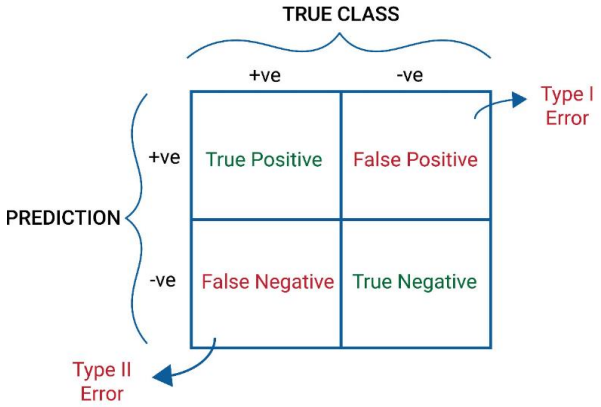


Figure 8. Confusion matrix.

4) Training

A lightweight CNN architecture is designed with python programming paradigm, Google TensorFlow 2.4.1, and Keras API 2.4.3 framework. The learning rate is tested for several values for the adam optimization algorithm and is selected as 0.0001 to obtain high efficiency and good convergence. Batch size is randomly chosen and fixed as 128 training samples for adam and made sure to avoid overfitting and also for the network to learn effectively. Also, a simple and efficient regularization approach of dropout has been adopted to reduce overfitting of the neural model, and drop out probability is taken as 0.2.

The dataset is organized into 2402 total training samples consisting of four classes: i) Ammaa ii) Aaaah iii) Help iv) Kapadii. Training a complex architecture neural network on a large dataset can be an extremely computational-consuming process. Our research audio dataset of vital signs of MI consists of considerably fewer samples and the CNN model designed is a lightweight model. Henceforth there was no necessity for choosing a complex GPU for the training process. The configuration of the computer system used for training is AMD Ryzen 5 3400G Quad-core Processor and RAM size of 14 GB. Table II shows various parameter settings for the training process carried out.

TABLE II. PARAMETER SETTING FOR TRAINING

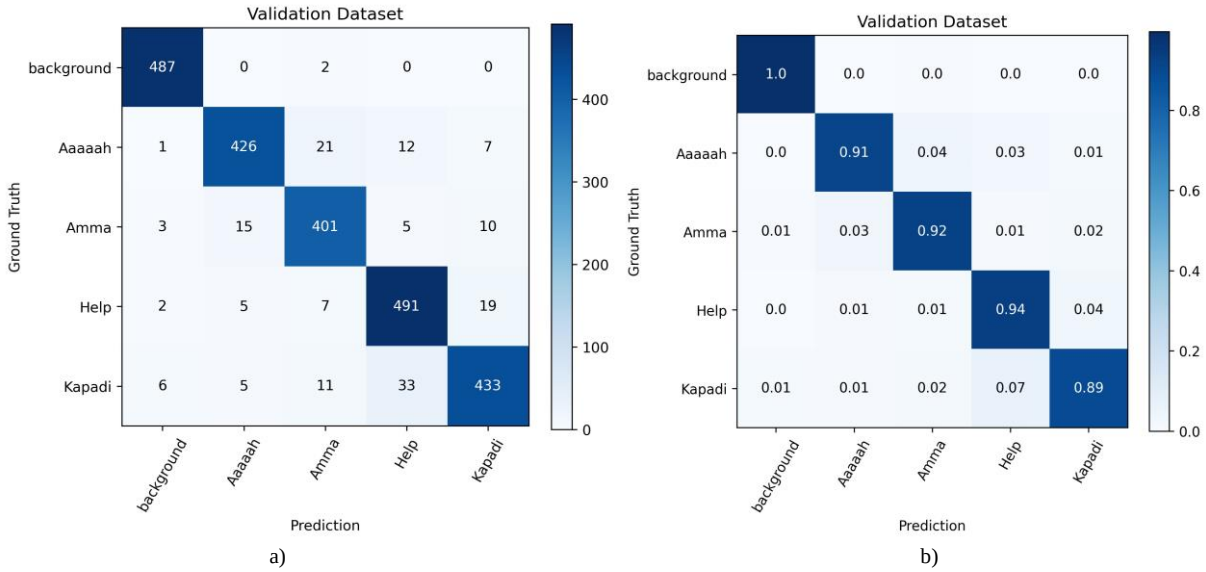
Parameter	Value
Epoch	120
Optimizer	Adam
Batch Size	128
Learning Rate	0.001

IV. EXPERIMENTS AND RESULTS

The general prototype design consists of NVIDIA Jetson Nano and a Logitech HD Webcam integrated with the noise-reducing microphone. The choice of embedded edge device is made mainly based on its compact size and powerful enough for advanced AI IoT applications with low power consumption. The development kit is integrated with 128-core Maxwell GPU, quad-core ARM A57 64-bit CPU, 4GB LPDDR4 memory, along with support for MIPI CSI-2 and PCIe Gen2 high-speed I/O. One of the other main considerations is to design a portable device with a low-cost parameter. The total cost of the implementation is around $300 United States dollars including Jetson Nano, microphone, and additional peripherals.

The results highlight the demonstration of the supervised audio classification scheme to detect the distress sounds at an indoor ambient smart space. The present research employs validation for the intelligent audio surveillance with four classes of events as Aaah, Ammaa, Kapadii, Help. Efforts are being made to obtain an extension of audio dataset samples collected through data augmentation technique and the addition of background noises to improve the detection accuracy even in challenging environments. Data augmentation of audio input samples such as variation in pitch, speed, and shift is performed to improve the training performance.

The performance of the proposed CNN network is tested with the private audio dataset. For the CNN architecture using MFCCs as inputs, the confusion matrix plot of the 2D CNN model is indicated in Fig. 9 which provides the quality of output of the classifier model. A true positive recognition rate of events is indicated in the diagonal elements can be estimated as a percentage of 100%, 91%, 92%, 94%, 89% for background noise, Aaah, Ammaa, Help, Kapadii, respectively. Table III provides different score values of precision, recall, F-1 score for multiclass classification for audio distress problems. The accuracy of the neural network model is calculated over the training and validation set illustrated in Fig. 10. The training and validation plot obtained is a smooth curve and indicates 97.91% overall accuracy in Fig. 9(a) and proves to have a good convergence speed and Fig. 9(b) shows a decreasing training as validation loss curves.

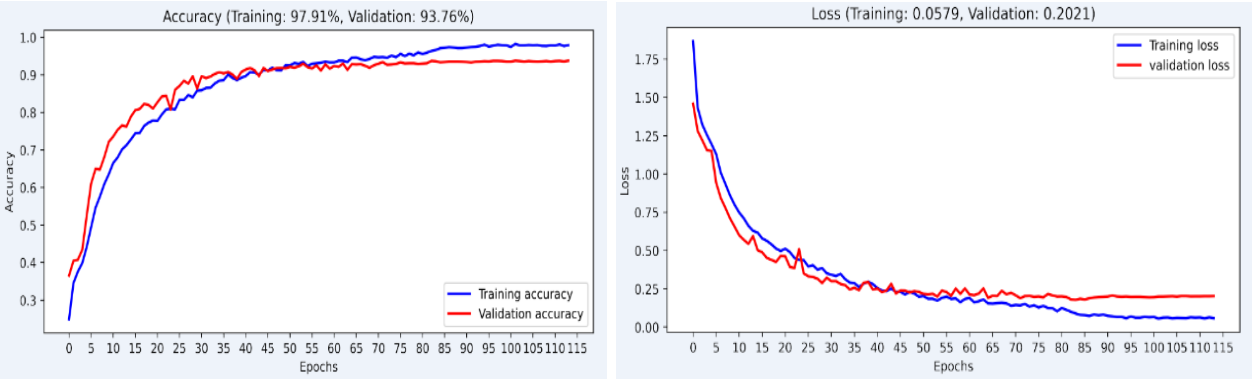


a) b)

Figure 9. a) Training and validation loss curves b) Training and validation accuracy curves.

TABLE III. PRECISION, RECALL AND F1 SCORE VALUES OF THE PROPOSED CNN MODEL

Class	Precision	Recall	F1-Score
Aaaaah	0.94	0.91	0.93
Ammaa	0.91	0.92	0.92
Help	0.91	0.94	0.94
Kapadi	0.92	0.89	0.90
Background sounds	0.98	1.0	0.99



(a) (b)

Figure 10. Confusion matrix of proposed CNN model for audio distress signals.

V. EMBEDDED EDGE IMPLEMENTATION

We examined the performance of the trained vital signs MI recognition model in real-time using Jetson Nano and a low-cost microphone. The neural network model that was trained on a computer system was deployed on NVIDIA Jetson Nano using Tensorflow RT, a deep learning framework for edge devices [14]. A chunk of audio samples is taken every 100 milliseconds and a preprocessing stage is performed to obtain MFCCs that are further passed to the convolution neural network for classification. Table IV illustrates the performance of our proposed CNN model in terms of power consumption and inference time during the real-time implementation of Jetson Nano. The results obtained highlight that the implementation of the lightweight model of CNN in the Jetson Nano is well suited to provide a quick response for vital signs of heart attack as real-time processing. The results obtained from our deep learning model on the edge device are comparable with the existing works that are implemented as real-time applications of audio processing. Table V shows the comparison of performance metrics of this proposed work with the results of current works.

TABLE IV. PERFORMANCE ANALYSIS OF PROPOSED CNN MODEL ON JETSON NANO

Embedded GPU	Power (Watt)	Proposed CNN Model	
		Power Consumption milli Watt (mW)	Inference Time milli Seconds (ms)
Jetson Nano	Idle	1400	-
	5	2643	300
	10	4281	100

TABLE V. PERFORMANCE COMPARISON WITH RELATED WORKS

Paper	Algorithm	Hardware	Inference time milli seconds (ms)
JF Gaviria *et al.* [14]	Multi-headed CNN	Raspberry Pi	550
Y Arslan [46]	Base, Linear Predictive Coding Warped Linear Prediction algorithms	Laptop -Intel(R) Core(TM) i5 CPU	104 to 1360
W Huang *et al.* [47]	Support Vector Machine	Pentium mobile CPU	1000
The proposed work	CNN	Jetson Nano	100

VI. DISCUSSION

Artificial intelligence techniques have garnered tremendous interest in speech recognition and sound event classification owing to their near-human level performance. The present work incorporates a Convolution Neural Network model for a distress sound classification for four possible sound classes. Audio recorded with high-quality microphone is collected as a private synthetic dataset which is sampled at 44,100 kHz. An audio dataset is added with an augmented data and an additional noise which enhanced the performance of the classifier model. An effective preprocessing technique of MFCC is carried out with 40ms time windows. The window frame length was chosen long enough to obtain a high performance and recognition rate in the classification task. In summary, the observation from the experimental results shows that our proposed multilayer CNN framework provides promising recognition accuracy for our vital signs of MI audio dataset. However,

a large collection of the validated data sets can enhance the detailed understanding of the presented medical emergency application. One of the effective approaches of audio-based classification is the usage of trained transfer learning neural network models which can enhance the performance which will be implemented in our future work. The performance of our vital signs of MI neural network model can be improvised by scaling it up with more data.

VII. Conclusion and Future Work

Healthcare automated pain recognition systems are complex systems that need to be solved by effective AI-based methods. This paper presented a novel approach of pain voice recognition through distress keywords during emergency situations of MI using a state-of-the-art convolution neural network algorithm. Audio signals consisting of four classes for the distress the situation was collected and the CNN model was developed. Our approach was validated by the synthetic private audio dataset for the selected emergency keywords of Kannada and English language. Experiment results highlight the effectiveness of the distress sound recognitions of humans during simulated scenarios of MI. Our future work focus is threefold: Firstly, on expanding the database to obtain better classification results that generalize for complex real-world scenarios. Secondly, to develop a much lighter neural network architecture to reduce the inference time and finally, to incorporate IoT technology for remote monitoring of heart attack elderly patients.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

H. M. Mohan conducted the research, collected data, and wrote the paper. Dr. S. Anitha supervised the work and approved the final version.

References

[1] A. Pazienza, G. Mallardi, and C. Fasciano, "Artificial intelligence on edge computing: A healthcare scenario in ambient assisted living," in *Proc. the Fifth Italian Workshop on Artificial Intelligence for Ambient Assisted Living*, November 2019, pp. 22-37.

[2] L. Greco, G. Percannella, and P. Ritrovato, "Trends in IoT based solutions for health care moving ai to the edge," *Pattern Recognition Letters*, vol. 135, pp. 346-353, July 2020.

[3] J. P. Queralta, T. N. Gia, and H. Tenhunen, "Edge-AI in LoRa-based health monitoring: Fall detection system with for computing and LSTM recurrent neural networks," in *Proc. 42nd International Conference on Telecommunications and Signal Processing*, July 2019.

[4] G. Muhammed and M. F. Alhamid, "Edge computing with cloud for voice disorder assessment and treatment," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 60-65, April 2018.

[5] X. Dai, I. Spacic, and B. Meyer, "Machine learning on mobile: An on device inference app for skin cancer detection," in *Proc. 4th Int Conf. on Fog and Mobile Edge Computing*, June 2019.

[6] C. Hegde, Z. Jiang, and P. B. Suresha. AutoTriage-An open source edge computing raspberry Pi-based clinical screening system. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.04.09.20059840v2

[7] World Health Organization. World Health Statistics Overview 2019. [Online]. Available: www.who.int/gho/publications/world_health_statistics/2019/en.

[8] E. J. Benjamin, *et al.*, "Heart disease and stroke statistics—2019 update: A report from the American heart association," *Circulation*, vol. 139, no. 10, 2019.

[9] K. L. Smith, P. Cameron, and A. Meyer, "Knowledge of heart attack symptoms in a community survey of Victoria," *Emergency Medicine*, vol. 14, pp. 255-260, 2002.

[10] A. Leviton, "Further comments on the Levine sign," *N. Engl. J. Med.*, pp. 273-282, 1965.

[11] W. M. Edmondstone, "Cardiac chest pain: Does body language help the diagnosis?" *Bmj*, vol. 311, pp. 1660-1661, 1995.

[12] A. Zehetner, M. Hagmuller, and F. Pernkopf, "Wake-up-word spotting for mobile systems," in *Proc. 22nd European Signal Processing Conference*, 2014.

[13] J. Ye, T. Kobayashi, and X. Wang, "Audio data mining for anthropogenic disaster identification: An automatic taxonomy approach," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, pp. 126-136, March 2020.

[14] J. F. Gaviria, A. Escalante-Perez, and J. C. Castiblanco, "Deep learning-based portable device for audio distress signal recognition in urban areas," *Appl. Sci.*, vol. 10, no. 21, 2020.

[15] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 161-165.

[16] G. Valenzise, M. Tagliasacchi, and F. Antonacci, "Scream and gunshot detection in noisy environments," in *Proc. 15th European Signal Processing Conference*, Poznan, Poland, September 2007.

[17] H. M. Mohan, *et al.*, "Non-invasive technique for real-time myocardial infarction detection using faster R-CNN," *Multimed. Tools Appl.*, vol. 80, pp. 26939-26967, 2021.

[18] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden Markov modeling for speaker-independent wordspotting," in *Proc. the International Conference on Acoustics, Speech and Signal Processing*, 1990, pp. 627-630.

[19] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in *Proc. the International Conference on Acoustics, Speech and Signal Processing*, 1990, pp. 129-132.

[20] J. G. Wilpon, L. G. Miller, and P. Modi, "Improvements and applications for key word recognition using hidden Markov modeling techniques," in *Proc. the International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. 309-312.

[21] S. Marius-Calin and H. Bourlard, "Iterative posterior based keyword spotting without filler models," in *Proc. the Automatic Speech Recognition and Understanding Workshop*, 1999, pp. 213-216.

[22] S. Marius-Calin, "Spotting subsequences matching an HMM using the average observation probability criteria with application to keyword spotting," in *Proc. the National Conference on Artificial Intelligence*, 2005, p. 1118.

[23] D. Grangier, J. Keshet, and S. Bengio, "Discriminative keyword spotting," *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, pp. 175-194, 2009.

[24] S. Tabibian, A. Akbari, and B. Nasersharif, "An evolutionary based discriminative system for keyword spotting," in *Proc. International Symposium on Artificial Intelligence and Signal Processing*, 2011, pp. 83-88.

[25] E. Principi, S. Squartini, and E. Cambria, "Acoustic template-matching for automatic emergency state detection: An ELM based algorithm," *Neurocomputing*, vol. 149, part A, pp. 426-434, 2015.

[26] F. Ge and Y. Yan, "Deep neural network based wake-up-word speech recognition with Two-stage detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[27] Q. Nguyen, S. Yun, and J. Choi, "Detection of audio-based emergency situations using perception sensor network," in *Proc. 13th International Conference on Ubiquitous Robots and Ambient Intelligence*, 2016.

[28] A. Saggese and N. Strisciuglio, "Reliable detection of audio events in highly noisy environment," *Pattern Recognition Letters*, vol. 65, pp. 22-28, November 2015.

[29] E. Principi, S. Squartini, and R. Bonfigli, "An integrated system for voicecommand recognition and emergency detection based on audio signals," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5668-5683, August 2015.
[30] X. Li and Z. Zhou, "Speech command recognition with convolutional neural network," *CS229 Stanford Education*, 2017.
[31] L. Pandey and K. Nathwani, "LSTM based attentive fusion of spectral and prosodic information for keyword spotting in Hindi Language," in *Proc. Interspeech Conference*, September 2018.
[32] J. Sharma, O. C. Granmo, and M. Goodwin, "Environment sound classification using multiple feature channels and attention based deep convolutional neural network," in *Proc. Interspeech*, 2020.
[33] J. Sharma, O. C. Granmo, and M. Goodwin, "Emergency detection with environment sound using deep convolutional neural networks," in *Proc. Fifth International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing*, 2021.
[34] R. M. Al-Eidan, H. Al-Khalifa, and A. Al-Salman, "Deep-Learning-Based models for pain recognition: A systematic review," *Appl. Sci.*, vol. 10, no. 17, 2020.
[35] P. Lopez-Martinez, "Multi-task neural networks for personalized pain recognition from physiological signals," in *Proc. Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, 2017, pp. 181-184.
[36] P. Thiam, P. Bellmann, H. Kestler, and F. Schwenker, "Exploring deep physiological models for nociceptive pain recognition," *Sensors*, vol. 19, no. 20, 2019.
[37] C. Wang, *et al.*, "Recurrent network based automatic detection of chronic pain protective behavior using MoCap and sEMG data," in *Proc. the 23rd International Symposium on Wearable Computers*, September 2019, pp. 225-230.
[38] P. Rodriguez, *et al.*, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, pp. 1-11, 2017.
[39] S. Jaiswal, J. Egede, and M. Valstar, "Deep learned cumulative attribute regression," in *Proc. 13th IEEE International Conference on Automatic Face & Gesture Recognition*, Xi'an, China, May 2018, pp. 715-722.
[40] B. Hu, C. Kim, and X. Xu, "Using a deep learning network to recognize low back pain in static standing," *Ergonomics*, vol. 61, pp. 1374-1381, 2018.
[41] F. Tsai and Y. Weng, "Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage," in *Proc. Seventh International Conference on Affective Computing and Intelligent Interaction*, San Antonio, USA, October 2017, pp. 313-318.
[42] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *Physiol.*, vol. 148, no. 3, pp. 574-591, 1959.
[43] A. Waibel, T. Hanazawa, and G. Hinton, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328-339, 1989.
[44] Y. Lecun, *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541-551, Dec. 1989.
[45] H. M. Mohan, S. Anitha, R. Chai, and S. H. Ling, "Edge artificial intelligence: Real-Time noninvasive technique for vital signs of myocardial infarction recognition using jetson nano," *Advances in Human-Computer Interaction*, vol. 2021, article ID 6483003, 2021.
[46] Y. Arslan, "A new approach to real time impulsive sound detection for surveillance applications," arXiv:1906.06586, 2019.
[47] W. Huang, *et al.*, "Scream detection for home applications," in *Proc. 5th IEEE Conference on Industrial Electronics and Applications*, Taichung, Taiwan, June 2010, pp. 2115-2120.

H. M. Mohan was born in Bangalore, Karnataka, India. He received his Bachelor of Engineering degree in electrical engineering in 2009 from BNMIT, Visvesvaraya Technological University, Bangalore, India. He went on to earn his Master of Technology degree in VLSI and Embedded Systems in 2011, from RNSIT, Visvesvaraya Technological University, Bangalore, India. His academic research interests include embedded systems, Internet of things (IoT), BioMedical signal processing. He is now currently working as a Research Engineer, AI at Digital Shark Technology, Bangalore.

Dr. S. Anitha was born in Bangalore, India. She received her Bachelor of Engineering in Medical Electronics from BMSCE, Bangalore University, Bangalore in 1999, and her Master in Technology from Biomedical Instrumentation from SJCE, Visveswaraya Technological University, Mysore in 2002 and her Ph.D. degree in Electronics & Instrumentation from Dr. M.G.R. University, Chennai, in 2013. She has authored more than 40 technical papers at the national and international level. She is a Member of Board of Studies, ML board, DSCE, Bangalore and Member, Board of examiners. IT/ML/EI, VTU Belagavi. She has been a Technical program committee/reviewer at various international conferences. She is an award recipient of GRABS educational charitable trust as "Best Young Teachers" 2017 and "Best Young Researcher Award" 2018. She has a teaching experience of around 23 years and is currently working as Prof. & Head, Dept. of Biomedical Engineering, ACS college of engineering. Her research interests are in the areas of digital signal and image processing, computer vision.