IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 24, 2021, accepted August 4, 2021, date of publication August 9, 2021, date of current version August 26, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3103501*

# Deep Learning (DL)-Enabled System for Emotional Big Data

**HAOPENG WANG[1], DIANA P. TOBÓN V.[1], M. SHAMIM HOSSAIN[2], (Senior Member, IEEE), AND ABDULMOTALEB EL SADDIK[1], (Fellow, IEEE)**

[1]MCRLAB, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: M. Shamim Hossain (mshossain@ksu.edu.sa)

This work was supported by the Researchers Supporting Project number (RSP-2021/32), King Saud University, Riyadh, Saudi Arabia.

**ABSTRACT** Emotion care for human well-being is important for all ages. In this paper, we propose an emotion care system based on big data analysis for autism disorder patient training, where emotion is detected in terms of facial expression. The expression can be captured through a camera as well as Internet of Things (IoT)-enabled devices. The system works with deep learning techniques on emotional big data to extract emotional features and recognize six kinds of facial expressions in real-time and offline. A convolutional neural network (CNN) model based on MobileNet V1 structure is trained with two emotional datasets, FER-2013 dataset and a new proposed dataset named MCFER. The experiments on three strategies showed that the proposed system with deep learning model obtained an accuracy of 95.89%. The system can also detect and track multiple faces as well as recognize facial expressions with high performance on mobile devices with a speed of up to 12 frames per second.

**INDEX TERMS** Convolutional neural network, emotion, facial expression recognition, mobile application.

## I. INTRODUCTION

With the proliferation of Artificial Intelligence (AI) and Internet of Things (IoT), emotion plays an important role in human life and communication [1] and [2]. In the daily life, emotion is an inextricable part of the interaction of human beings, which can be observed by the changes in physiological features and behaviors. Because emotion recognition has a great potential to improve our quality of life, in the past decades, emotion recognition has aroused a lot of attention of many researchers and has been a popular research topic in various fields such as robotics, human-computer interaction, and entertainment, to name a few [3]. Meanwhile, emotion care can be very useful in medical applications when medical staff need to assess the patient's feeling and behavior during or after the surgery. With the development of big data and deep learning, huge amount of data including emotional data is generated in recent years, which cannot be handled with the traditional techniques. To this end, deep learning techniques has the potential to solve this problem [4]. By using deep learning techniques to analyze the emotional big data, machines can learn and understand emotions to meet human needs, because deep learning techniques can learn and track different physiological features on the body.

Physiological features are closely associated with the generation of emotion and can be used for the recognition of the emotion. However, physiological data are inconvenient to obtain, many researchers pay more attention on other factors, such as facial expressions, gesture, voice [5], [6]. Among these mentioned factors, facial expressions play the most important role, which contribute 55 percent in the emotion analysis, while the vocal part and verbal part contribute approximately 38 percent and 7 percent, respectively [7]. Therefore, facial expressions are the most significant part in the behavior analysis of emotion. In addition, with the development of smartphones and wearable devices, the portable healthcare system becomes more and more important. Many smart devices are developed to help people monitor health, such as heart rate, EEG [8]. However, emotion care system working on smart devices is paid less attention, while there are so many emotion care systems based on computer system, which are not user-friendly and portable. In this paper, we propose an emotion care system based on automatic facial expression recognition (FER) system working on an Android smartphone.

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Khurram Khan.

Digital twin proposed by El Saddik [9] is the convergence of science and technology to improve well-being of citizens and quality of life. Digital twins are virtual representations of living or non-living physical entities by bridging the physical and the virtual world. An application of digital twin could be used to improve health and well-being, which would assist a person, such as autism disorder patient, to better integrate into communities. As a part of digital twin, one main application of the proposed emotion care system in this paper is to help people with Autism Spectrum Disorder (ASD). People with ASD usually have problems with emotional, social, and communication skills. For children, it is estimated that the prevalence of ASD is up to 1 in 88 [10]. Because it is very difficult for people with ASD to express their emotion and convey how they feel when communicate with others, the DT-FER systems could significantly help them to understand their emotion and take control of their facial expression when an outburst is eminent. Meanwhile, the carers or educators of autistic children could better understand the children's real feelings and make better decisions with the help of emotion care system.

With the breakthroughs of algorithms and computing power in recent years, machine learning, especially deep learning, has become an important topic in pattern recognition, communication sentiment analysis [11] and feature learning [12]. Many CNN models with high accuracy are proposed and trained by researchers in different fields such as hand-written digit recognition, object detection, and FER. However, computers are not portable and hard to use compared to the smartphone. Because of the rapid growth of smartphones, there are so many applications and services developed to provide more convenience for human's daily life. Moreover, most autistic children are visual learners, which means mobile technology could arouse their interest, thus they would enjoy using smart devices to learn and play. Therefore, development of emotion care system based on DT-FER system on mobile phone could be used widely in several applications, specially, for ASD children.

In this paper, we propose a DT-FER based emotion care system, where emotion is detected in terms of facial expression. As shown in Figure 1, the system contains two parts: model and Android application. In the model part, after image preprocessing for the images from dataset, the deep learning model is trained for emotion classification and deployed on a smartphone. The DT-FER Android application (app) detects the multiple faces in the images captured from the smartphone camera by using face detector and predicts facial expressions from detected face images by using CNN model in real time and offline. The application calculates a score for each facial expression and shows the highest score and emotion on the smartphone screen. And the highest score indicates the performed emotion meets the standard better. And the performed facial expression and predicted results can be saved into device to help carers to observe facial expressions of patients at any time and provide better suggestions to patients. In addition, a new dataset for emotion recognition is collected to train the CNN model. In summary, the main contributions of this paper are

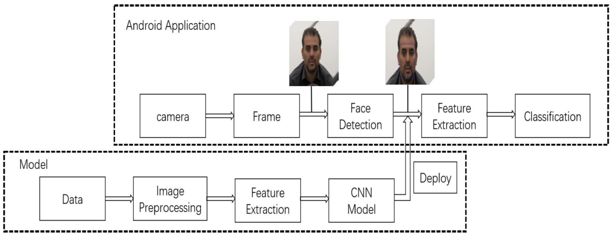


**FIGURE 1.** System architecture.

1) A new low-cost and multi-user framework for emotion detection is proposed. The system is based on an Android application that uses CNN model to classify facial expressions and works in real time and offline.
2) A new challenging and less-constrained dataset called MCFER is introduced for facial expression classification. The dataset is collected in real scenario with more complex conditions, such as movement and light interference.

The rest of this paper is organized as follows. In Section II, we discuss the related work about emotion recognition, FER, and mobile applications for FER. In Section III, we describe related databases for facial expression. In Section IV, we show the CNN model structure and workflow of DT-FER application in detail. In Section V, we show the results of CNN model and the performance of our DT-FER application. Finally, we conclude our work in Section VI.

## II. RELATED WORK

The increasing number of devices connected to the Internet of Things (IoT) generates a massive amount of emotional data, posing a data analytics challenge for IoT applications [1], [13]. With regards to the above concern, one major problem for emotion recognition is the definition of emotion. Darwin [14] assumed that there is a finite amount of emotions that exist in all cultures. Researchers proposed many categorical models, which label the emotions in discrete categories. The basic universal emotions (i.e., angry, disgust, fear, happiness, sadness, surprise, and neutral), widely used by researchers, were classified by Ekman and Friesen [15] in 1975. In this paper, we adopt six kinds of these emotions except neutral, since neutral is a basic emotion that does not need more attention in ASD.

There are many methods for emotion recognition [16], but facial expression is an indispensable part in most of proposed methods. Ma *et al.* [5] proposed an emotion recognition with audio and video data. 2D CNN is used to extract features of audio, while 3D CNN is applied on facial expression image sequence. Then, deep belief network is used to the fusion of emotion features. Finally, emotion can be obtained by the Support Vector Machines (SVM).

Hossain and Muhammad [17] developed an application where ridgelet transform was used for face image features. Two distinct extreme learning machines (ELM) classifiers are used: one for speech features and the other for face feature. Then Bayesian sum law used to get the combined scores of these two classifiers for possible final decision. Although their model achieves an accuracy of 85.06%, there is no description of mobile application nor the integration of the concept of digital twin in their paper.

Hossain and Muhammad [18] proposed a 5G-enabled emotion-aware framework. The emotion is detected using speech and video that captured by wireless cognitive sensors. And the processing of video and speech are performed in cloud. The maximum accuracy obtained in the experiment is 99.8%. To utilize a large amount of unlabeled data, Zhang *et al.* [19] proposed an emotional care using the deep learning.

As an important way to emotion recognition, there are many diverse FER methods that achieve a good performance (e.g., SVM [20], [21], Linear Discriminant Analysis (LDA) [22], Bayesian Network (BN) [23], Neural Network (NN) [24], Gaussian mixture model(GMM) [25], to name a few). These proposed classifiers recognize facial expression according to different facial features, which can be classified into three such as geometric features, appearance features and a hybrid of the aforementioned features. Examples of these features are distance of face landmarks [26], local binary patterns (LBP) [27], Gabor wavelets coefficients [22], histograms of oriented gradients (HOG) [28], to name a few. These features are extracted by hand and then fed into classifiers, which require substantial programming effort and computational power. As a novel method, CNN-based FER classifier combines facial feature extraction and facial expression classification into one stage and makes tremendous progress for FER. Therefore, many CNN architectures for FER have been proposed and obtain much higher accuracy compared with traditional methods [20], [29], [30].

For FER system on smartphone, Suk and Prabhakaran [31] proposed a system that distinguishes between neutral and non-neutral expression frames in video sequences by using facial landmarks. If non-neutral expression is found, the new dynamic features are generated by displacing saved neutral features with current features. Then the new features are fed into SVM models for FER task. The SVM model obtained an accuracy of 86%. The mobile application was tested on Samsung Galaxy S3 with 2.4 frames per second (fps).

Song *et al.* [32] developed a deep learning FER application with DNN model with an accuracy of up to 99.2%. The DNN has 5 layers and recognize 5 facial expressions (i.e., anger, happy, sad, surprise, and neutral). The smartphone application captures first the user's face, then it sends a request to a server, thus the server predicts facial expressions by a trained model and then sends the prediction to mobile phone. Due to the app uses the client-server architecture, it cannot work offline.

Jo *et al.* [33] proposed a new robust FER system against illumination variation, which utilizes Active Appearance Model (AAM) and NN with a Difference of Gaussian (DOG) to fix illumination variation problems and recognize facial expressions. Alshamsi *et al.* [34] developed an automated FER application with an accuracy of 96.3%, but it uses cloud computing for FER based on facial landmarks and SVM.

The existing smartphone applications are compared in Table 1. While most applications work in real time and offline, none of them supports multiple users for emotion detection. Therefore, we develop a multi-user Android application running in real time and offline for emotion care and autism disorder patient training.

**TABLE 1.** Comparison of existing FER smartphone applications.

| Method | Method | Multi-user | Real-time | Offline |
|---|---|---|---|---|
| Suk et al. [31] | SVM | No | Yes | Yes |
| Song et al. [32] | CNN | No | Yes | No |
| Jo et al. [33] | AAM | No | Yes | - |
| Alshamsi et al. [34] | SVM | No | Yes | No |
| Ours | CNN | Yes | Yes | Yes |

## III. DATASET

In this section, we show two popular emotional big data and a new dataset that is named MCFER and its acquisition process. The architecture of CNN model and the framework of application are also described in detail.

### A. FER-2013

Facial Expression Recognition 2013 (FER-2013) dataset [35] was prepared in Challenges in Representation Learning: Facial Expression Recognition Challenge, which is hosted by Kaggle. FER-2013 database has seven facial expression categories (e.g., angry, disgust, fear, happy, sad, surprise, and neutral) and three different sets such as training set (28.709 images), validation set (3.589 images), and test set (3.589 images). All images in this dataset are grayscale with $48 \times 48$ pixels, thus corresponding to faces with various poses and illumination, where several faces are covered by hand, hair, and scarves. Because of FER-2013 is collected from the Internet and has various real-world conditions, it becomes one of the largest and most challenging database for facial expression recognition.

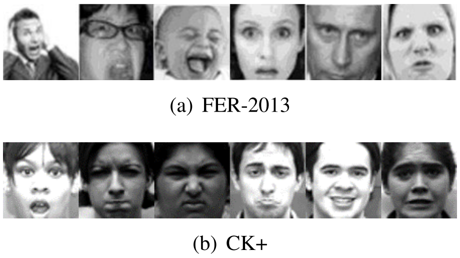
### B. CK+

CK+: the Extended CohnKanade (CK+) database [21] consists of 593 deliberate image sequences from 123 subjects, which is the most used lab-controlled database for evaluation of FER systems. The database was labeled with seven basic facial expressions (anger, contempt, disgust, fear, happiness, sadness and surprise) by adopting the FACS (Facial Action Coding System). The examples of FER-2013 and CK+ are shown in Figure 2.

### C. MCFER

#### 1) DATA ACQUISITION

Unlike most popular databases (e.g., CK+ [21], JAFFE [36]) that were collected in special lab environments with same and specific environment, our proposed new MCFER (Multimedia Communications Research Laboratory Facial



(a) FER-2013

(b) CK+

**FIGURE 2. The examples of two datasets: FER-2013 and CK+.**

Expression Recognition) database is collected in various places of University of Ottawa without particular environment, which makes the data more realistic. Totally, 15 participants (33% female and 67% male) between 23 and 60 years old take part in the experiment. The participants are randomly selected in real life and most of them are students or staff at the University of Ottawa. The experiment is designed as follows: first, the participant is asked to read and sign a consent form to participate in the experiment. Then, the participant reads an instruction about the details of the experiment and is instructed by an experimenter to understand the purpose of the study and the detailed experimental procedure. Once the participant is ready in front of the screen, the researcher turns the camera on and asks the participant to perform a series of facial expressions starting with angry and ended in surprise. The participants start the experiment in a natural place where they are found, instead of performing in a special man-made environment. Moreover, participants can look at the camera at any angle they want.

### 2) DATA DEDUPLICATION

For every participant, one video with six kinds of facial expressions is collected and processed. A haar cascade classifier proposed by Viola and Jones [37] is used to detect the face from video frame by frame. When a face is detected, the face image is saved into the database and labeled according to the facial expression the participant shows. Because the database has a lot of similar images due to the successive frames, we use difference hash (dhash) algorithm to select representative images from the dataset. The difference hash is one of image fingerprint algorithms, and it creates a unique hash value by calculating the difference between adjacent pixel values. To select images form the dataset, we use difference hash to compute our image fingerprints because of its speed and accuracy.

As shown in Figure 3, the image was first shrank to a new size $\{S + 1\} \times S$, which would match any similar images regardless of how it is stretched by ignoring the original size and aspect ratio. Second, the colorful image was converted to grayscale image which reduce hash from $S^2 + S$ pixels to a total of $S^2 + S$ colors. Then, the dhash algorithm calculated the difference between adjacent pixels, which identifies the relative gradient direction. We got $S$ differences from $S + 1$ pixels per row. Therefore, the total differences of the image are $S^2$ bits. Finally, we compared the brightness between adjacent pixels to get the hash value. If the left pixel is brighter than the right pixel, the bit is set to 1, otherwise 0. To compare two hashes, we just counted the number of bits that are different which is the Hamming distance. If the Hamming distance between two images is less than the threshold $D$, one image would be discarded because we regarded these two images were the "same" images containing repetitive information.

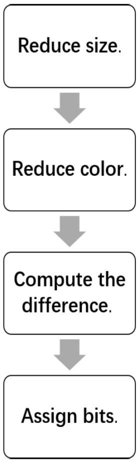
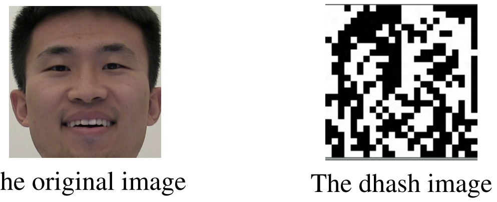


**FIGURE 3. The dhash algorithm.**



The original image

The dhash image

**FIGURE 4. The dhash image is obtained from the original image by using dhash algorithm.**

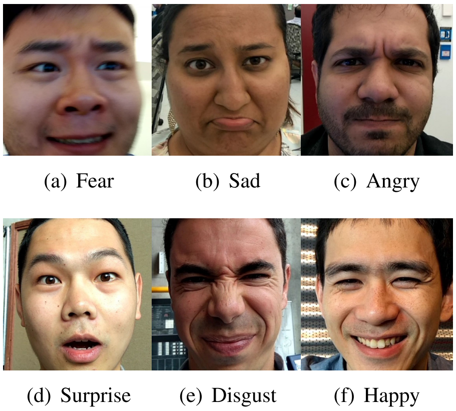
As shown in Figure 4, the original facial expression was encoded as a new hash value. We select 24 as the image size and 170 as the threshold, which means all images with the distance less than 170 were discarded. The dhash image shows the encoded image using dhash.

Then, four volunteers who retain vote power are found to judge whether each image belongs to its corresponding category. Once one of the volunteers cast opposing vote for the image, the image is discarded. We finally obtained 287 images for the proposed MCFER dataset. Figure 5 depicts some collected images of the MCFER dataset.

## IV. METHOD

### *A. PREPROCESSING*

For the system, image preprocessing is necessary before an image is fed into the CNN model. The image preprocessing mainly consists of two stages: face detection, data augmentation. A face detector is adopted for face detection in our system. If faces are detected, the four coordinates of region of interest (ROI) of the faces would be returned


(a) Fear (b) Sad (c) Angry

(d) Surprise (e) Disgust (f) Happy

**FIGURE 5.** **Some examples of the proposed MCFER database.**

to the system, the system would crop the faces and discard irrelevant background. Data augmentation is used to process the detected face images and increase the quantity of data, because training process of deep learning model usually needs huge amounts of data. The images are cropped by the random bounding boxes that have different cropped ranges from 0.85 to 1. Then the data are randomly flipped and rotated. The amount of data has increased by 200 times.

### *B. CNN MODEL*

The MobileNet V1 [38], which is used as our model architecture, was proposed in 2017. It is a lightweight deep neural network, which already became in an underlying network structure. Because its key point is to construct a small neural network, it can be widely used on mobile and embedded devices with remarkable speed and good accuracy compared with other architectures. The main idea of the MobileNet V1 is to decouple standard convolution into a 1 × 1 pointwise and depthwise convolutions to extract features from big data.

Depthwise convolution is used to extract features, where those features are combined into new features by pointwise convolution. Thus, Mobilenet V1 architecture has smaller model size and complexity because of fewer number of parameters and fewer additions and multiplications. The reduction of computation between traditional convolution and combination of depthwise and pointwise convolutions are compared in Equation 1. The computation amount of traditional convolution is represented as denominator and the numerator shows the computation amount of MobilNet V1.

$$\frac{D_F * D_K * D_F + M * N * D_F * D_F}{D_K * D_K * M * N * D_F * D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{1}$$

where $D_K$ is the convolutional kernel size and $D_F$ denotes an input feature map, while M and N denote the number of input and output channels, respectively. In addition, the MobileNet V1 also has two hyperparameters such as the resolution multiplier $\rho$ to control the size of the feature map and the width multiplier $\alpha$ to reduce the calculation amount. Because of the MobileNet V1 can reach a satisfactory trade-off between accuracy and speed compared to other popular CNN structures (e.g. AlexNet [39], VGG16 [40]), we select this lightweight CNN model as our framework to train a FER classifier. The model is shown in Figure 6.

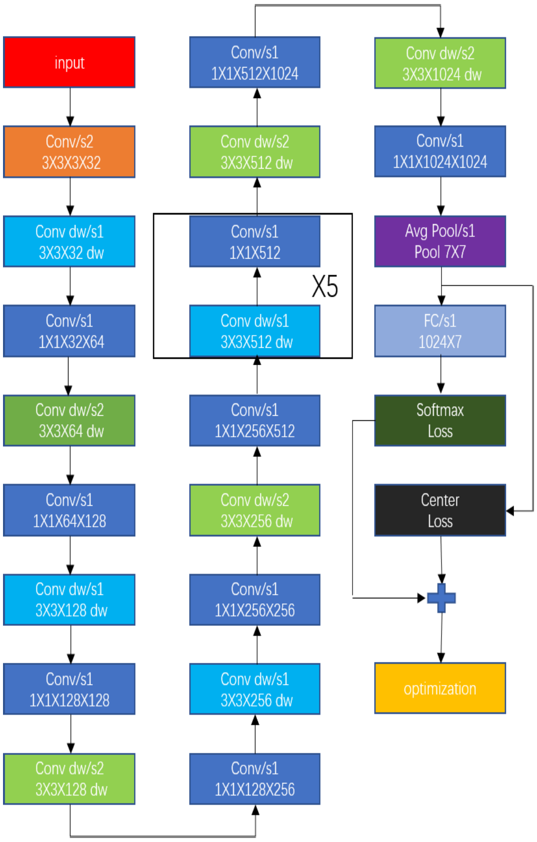


**FIGURE 6.** **The proposed architecture with joint loss.**

### *C. JOINT LOSS*

Center loss [41] shown in Equation 2 can be used in training step to increase discriminatory power by reducing the distance constraint between the feature and its corresponding class center. Therefore, it can be used to learn discriminative feature and improve model performance.

$$L_{CL} = \frac{1}{2}\sum_{i=1}^{N} \|x_i - c_{y_i}\|_2 \tag{2}$$

where $x_i$ denotes the ith features extracted from $y_i$th class and $c_{yi}$ denotes the learned center for the $y_i$th class. The softmax loss increases the distance between different classes, the center loss reduces the distance within a class. Therefore, the joint loss containing center loss and softmax loss not only enlarges inter-class feature difference, but also reduces intra-class feature variation, as shown in Equation 3.

$$\begin{aligned} L &= L_S + \lambda \cdot L_{CL} \\ &= -\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{m} e^{W_j^T x_i + b_j}} + \frac{1}{2}\sum_{i=1}^{N} \|x_i - c_{y_i}\|_2 \end{aligned} \tag{3}$$

where $L_S$ and $L_{CL}$ are the softmax and center loss, respectively. $\lambda$ denotes a hyperparameter to balance two loss functions and $b \in \Re^n$ is the bias term. $W_j \in \Re$ is the jth column of the weights $W_j \in \Re^{d \times n}$ in the last fully connected layer. The network is optimized using SGD (Stochastic Gradient Descent) [42].

### D. TWO-STAGE STRATEGY

Training a new model from scratch takes a tremendous amount of time and effort and needs a lot of data to achieve high performance. We use a two-stage approach [7] to train the CNN model in order to deal with the problem of insufficient size of small datasets. We implement the first stage fine-tuning with FER-2013 dataset by using the model with pre-trained weights from the ILSVRC-2012 (ImageNet) [43]. After the best trained model is obtained from the FER-2013 dataset, the second stage fine-tuning is implemented with our new dataset to obtain the final model.

For the two-stage strategy, the last fully connected layer is replaced by new fully connected layer with six facial expression classes output. For the first fine-tuning stage, due to the FER-2013 dataset has 48 × 48 pixels images and the input size of original MobileNet V1 is 224 × 224 pixels, the Gaussian distribution is adopted to initialize the parameters of the first convolutional layer. For the second fine-tuning stage, the input image of MCFER dataset has the same size of 48 × 48 pixels as the size of FER-2013 dataset after preprocessing. The weights of the first convolutional layer are initialized with the weights from the FER-2013.

### E. DT-FER APPLICATION

The CNN model is developed using TensorFlow platform, which is an end-to-end open source platform for machine learning. In order to use CNN model on the mobile phone, the model first needs to be converted to a new data (i.e.,.lite) file by TensorFlow Lite. TensorFlow lite provides a set of tools to run TensorFlow models on mobile and embedded devices. After the model is converted to a lite file, it is deployed on mobile phone with the TensorFlow Lite interpreter.

When the app starts to work, it continuously captures the images from the front camera or rear camera. The haar-like feature is employed to detect faces in the application and then the detected faces are cropped and resized to 48 × 48 size. After normalization and other processing methods, the face images are fed into the model as input. Finally, the prediction and other results are shown on the screen in real time. The workflow of facial expression application is shown in Figure 7.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

### A. PERFORMANCE OF CNN MODEL

To train a model for FER, all images of the datasets have to be detected by using the haar cascade classifier in order to determine if there is a face in the image. If the classifier detects the face, this will be cropped. After data augmentation and normalization for the cropped images, the processed images are used to train the model. As shown in Figure 8, with 8000 training steps, the model trained on FER-2013 database in the first fine-tuning stage obtained an accuracy of 67.03%. Based on the first stage pre-trained model on FER-2013 dataset, the accuracy of model on MCFER dataset in the second stage is about 95.89%. As shown in Figure 9, the model also has a precision of 97.58% and a recall of 100%. In addition, as shown in Figure 10 F1-score of the model is 98.79%.

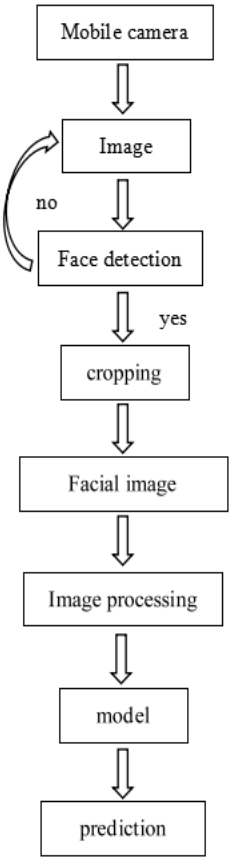


**FIGURE 7.** Workflow of the proposed mobile application.

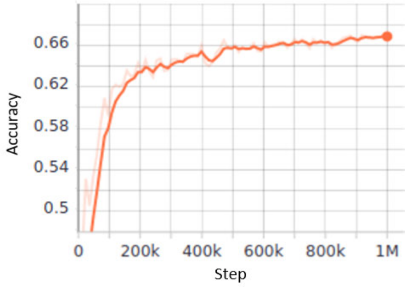
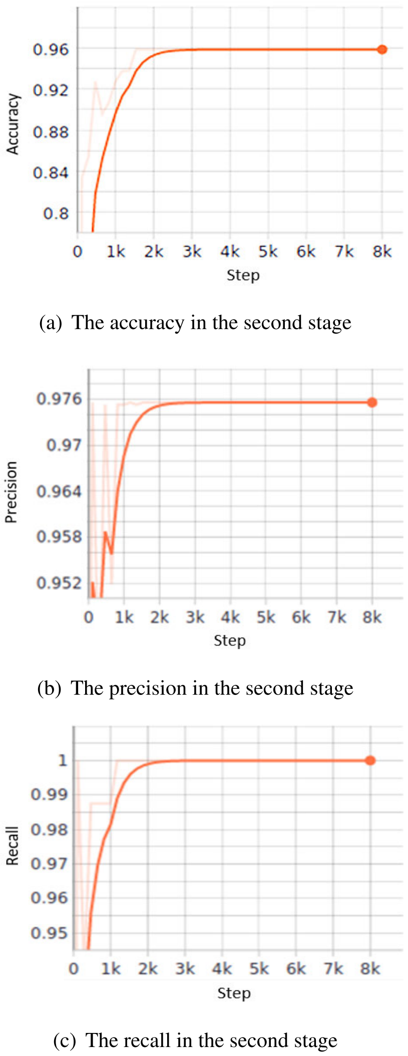


**FIGURE 8.** The accuracy of the first stage model with 1 million steps on FER-2013 dataset.

We also test three methods with different training stages and loss functions on two datasets, CK+ and MCFER, as shown in Table 2. One-stage with softmax loss method



(a) The accuracy in the second stage

(b) The precision in the second stage

(c) The recall in the second stage

**FIGURE 9. The performance of the second stage model with 8000 training steps on MCFER dataset.**

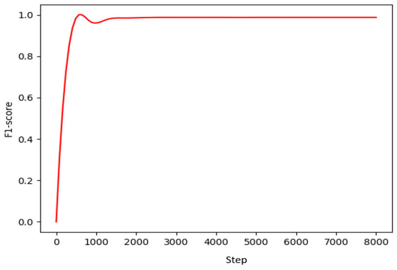


**FIGURE 10. The F1-score of the second stage model with 8000 training steps on MCFER dataset.**

**TABLE 2. Comparison of different strategies.**

| Strategy | CK+ | MCFER |
|---|---|---|
| One-stage with softmax loss | 92.16% | 82.36% |
| two-stage with softmax loss | 95.12% | 94.03% |
| two-stage with joint loss | 96.85% | 95.89% |

indicates that the CNN model is fine-tuned on the pre-trained ImageNet model with softmax loss function, the FER-2013 dataset is not used in this method. Two-stage with softmax loss means the model is first fine-tuned on pre-trained ImageNet model with FER-2013, then FER-2013 model is fine-tuned on CK+ or MCFER. The last one, two-stage with joint loss, is the method we used in the paper, the center loss and softmax loss are combined together to get the joint loss function. As shown in Table 2, two-stage strategy improves the performance of the model significantly, because large dataset first coarsely adjusts the feature extractor to FER task, then small dataset fine tunes the model again to get better performance. Meanwhile, the joint loss from softmax loss and center loss also improves the accuracy of the model, as the center loss make the model to update the weights to better learn deep discriminant features from data.

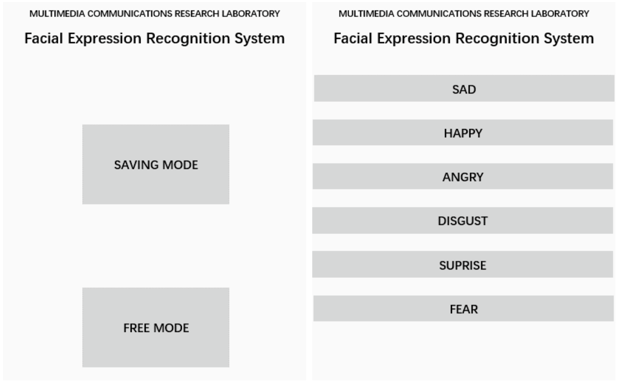
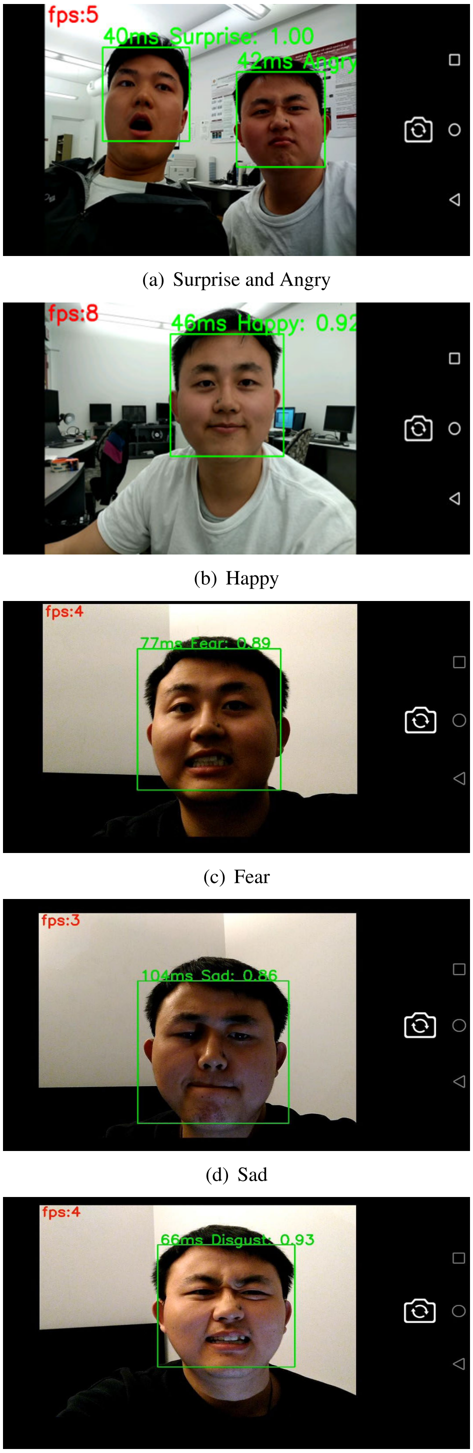

MULTIMEDIA COMMUNICATIONS RESEARCH LABORATORY
Facial Expression Recognition System
SAVING MODE
FREE MODE
MULTIMEDIA COMMUNICATIONS RESEARCH LABORATORY
Facial Expression Recognition System
SAD
HAPPY
ANGRY
DISGUST
SUPRISE
FEAR

**FIGURE 11. Left: Saving mode and free mode in the proposed mobile application. Right: When the saving mode is selected, the predicted facial expressions will be saved in their corresponding directory.**

### B. PERFORMANCE OF DT-FER APPLICATION

To monitor patients in real time and offline, the model is integrated into the Android application instead of remote server. The application has two kinds of working modes: saving mode and free mode as shown in Figure 11. The saving mode means the images and predicted results are saved on the phone, which would help carers to observe behaviors of people with autism at any time and provide better suggestions to patients, but the app would run slower because of saving image. The free mode, in turn, means nothing needs to be saved, so the app runs at higher speed compared with saving mode. In addition, the application support multiple facial expressions recognition, which can detect multiple faces and recognize facial expression for each face at the same time. As shown in Figure 12, when users use the developed system, four results, time spent in prediction, frame per second, predicted facial expression, and prediction confidence, are shown on the screen in real time. Time spent in prediction indicates the running time that CNN model spent in predicting facial expressions. Frame per second shows that the speed of the system. The predicted facial expression shows the users' facial expression detected by the system, and prediction confidence means confidence score for the predicted result. With the help of the developed system, doctors or parents could monitor the training process of patient.



(a) Surprise and Angry

(b) Happy

(c) Fear

(d) Sad

(e) Disgust

**FIGURE 12. Screenshot of the DT-FER app for basic expressions.**

The system is tested on a Huawei G9 VNS-AL00 smartphone, which has Qualcomm MSM8952 CPU Processor, 3GB of RAM, and 16GB of ROM. As shown in Figure 12, the prediction is shown on the screen, as well as the time taken by the model to make the prediction (e.g., 40 ms, 46 ms, etc.). The App can recognize facial expressions with 12 fps in free mode. However, when the system works in saving mode, it is slower with a speed of 9 fps. Meanwhile, the app can detect multi-face and recognize facial expression. But if there are more faces, the speed is lower. For instance, when two faces are detected by the app, the speed would be about 5 fps, which is half of the speed recognizing one face. Because of the limitation of computation power on the mobile device, we tested the application with saving mode under the condition of full load. The obtained speed was about 4 fps, which is the lowest speed for the application in the smartphone.

## VI. CONCLUSION

In this paper, we adopt a deep learning technique to process emotional big data and develop an emotion care system using facial expression recognition system working on smartphone. The CNN model is trained with two emotional datasets: FER-2013 and a new proposed dataset, MCFER, which was collected at the University of Ottawa. We employ a two-stage strategy and joint supervision to train our model and test the method on two datasets: CK+ and MCFER, which shows the performance is improved. The model on MCFER obtains a good performance with 95.89% accuracy. The Android application can detect multiple faces and recognize facial expression for every face in real time and offline. We test the application in a smartphone with two modes, saving mode that saves the predicted results to help doctors or parents monitor autism disorder patient and free mode that make the system work at high speed. The application has a good performance with a speed of up to 12 fps. Because of the weak computation power and limited memory for the embedded system, the model could be quantified to save memory space and improve speed in the future.

## REFERENCES

[1] E. Ahmed, I. Yaqoob, I. A. T. Hashem, I. Khan, A. I. A. Ahmed, M. Imran, and A. V. Vasilakos, "The role of big data analytics in Internet of Things," *Comput. Netw.*, vol. 129, pp. 459–471, Dec. 2017.

[2] K. Lin, F. Xia, W. Wang, D. Tian, and J. Song, "System design for big data application in emotion-aware healthcare," *IEEE Access*, vol. 4, pp. 6901–6909, 2016.

[3] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Inf. Fusion*, vol. 53, pp. 209–221, Jan. 2020.

[4] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Comput. Hum. Behav.*, vol. 93, pp. 309–317, Apr. 2019.

[5] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, Mar. 2019.

[6] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio-visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.

[7] Y. Miao, H. Dong, J. M. A. Jaam, and A. E. Saddik, "A deep learning system for recognizing facial expression in real-time," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2, pp. 1–20, Jun. 2019, doi: 10.1145/3311747.

[8] G. Muhammad, M. S. Hossain, and N. Kumar, "EEG-based pathology detection for home health monitoring," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 603–610, Feb. 2021.

[9] A. E. Saddik, "Digital twins: The convergence of multimedia technologies," *IEEE Multimedia*, vol. 25, no. 2, pp. 87–92, Apr./Jun. 2018.

[10] M. Chuah and M. Diblasio, "Smartphone based autism social alert system," in *Proc. 8th Int. Conf. Mobile Ad-hoc Sensor Netw. (MSN)*, Dec. 2012, pp. 6–13.

[11] Z. Long, R. Alharthi, and A. E. Saddik, "NeedFull—A tweet analysis platform to study human needs during the COVID-19 pandemic in New York state," *IEEE Access*, vol. 8, pp. 136046–136055, 2020.

[12] Y. Zhou, H. Dong, and A. E. Saddik, "Deep learning in next-frame prediction: A benchmark review," *IEEE Access*, vol. 8, pp. 69273–69283, 2020.

[13] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Gener. Comput. Syst.*, vol. 91, pp. 563–573, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X18311099

[14] C. Darwin, *The Expression of the Emotions in Man and Animals* (Cambridge Library Collection—Darwin, Evolution and Genetics). Cambridge, U.K.: Cambridge Univ. Press, 2013.

[15] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*. Upper Saddle River, NJ, USA: Prentice-Hall, 1975.

[16] Y. Hao, J. Yang, M. Chen, M. S. Hossain, and M. F. Alhamid, "Emotion-aware video QoE assessment via transfer learning," *IEEE Multimedia Mag.*, vol. 26, no. 1, pp. 31–40, Jan. 2019.

[17] M. S. Hossain and G. Muhammad, "Audio-visual emotion recognition using multi-directional regression and ridgelet transform," *J. Multimodal User Interfaces*, vol. 10, no. 4, pp. 325–333, 2016.

[18] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2399–2406, Aug. 2018.

[19] Y. Zhang, Y. Qian, D. Wu, M. S. Hossain, A. Ghoneim, and M. Chen, "Emotion-aware multimedia systems security," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 617–624, Mar. 2019.

[20] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun. 2016.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010 pp. 94–101.

[22] H. B. Deng, L.-W. Jin, L.-X. Zhen, and J.-C. Huang, "A new facial expression recognition method based on local Gabor filter bank and PCA plus LDA," *Int. J. Inf. Techn.*, vol. 11, no. 11, pp. 86–96, 2005.

[23] L. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understand.*, vol. 91, no. 1, pp. 160–187, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107731420300081X

[24] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Sep. 2015, pp. 1–6.

[25] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, "Audio–visual emotion-aware cloud gaming framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.

[26] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314285710041

[27] S. Liao, W. Fan, A. C. S. Chung, and D.-Y. Yeung, "Facial expression recognition using advanced local binary patterns, Tsallis entropies and global appearance features," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 665–668.

[28] P. Kumar, S. L. Happy, and A. Routray, "A real-time robust facial expression recognition system using HOG features," in *Proc. Int. Conf. Comput., Anal. Secur. Trends (CAST)*, Dec. 2016, pp. 289–293.

[29] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *CoRR*, vol. abs/1509.05371, pp. 1–8, Sep. 2015. [Online]. Available: http://arxiv.org/abs/1509.05371

[30] S. Qian, T. Zhang, C. Xu, and M. S. Hossain, "Social event classification via boosted multimodal supervised latent Dirichlet allocation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 2, pp. 1–22, Jan. 2015.

[31] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system—A case study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 132–137.

[32] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2014, pp. 564–567.

[33] G.-S. Jo, I.-H. Choi, and Y.-G. Kim, "Robust facial expression recognition against illumination variation appeared in mobile environment," in *Proc. 1st ACIS/JNU Int. Conf. Comput., Netw., Syst. Ind. Eng.*, May 2011, pp. 10–13.

[34] H. Alshamsi, V. Kepuska, and H. Meng, "Automated facial expression recognition app development on smart phones using cloud computing," in *Proc. IEEE 8th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2017, pp. 577–583.

[35] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014002159

[36] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[37] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, pp. 1–9, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

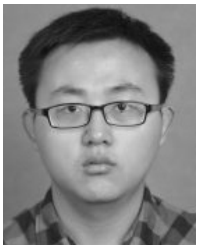
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[41] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, 2016, pp. 499–515.

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.



**HAOPENG WANG** received the M.Eng. degree in electronic and communication engineering and the B.Eng. degree in information and electronics from Beijing Institute of Technology, Beijing, China, in 2017 and 2015, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Ottawa. His research interests include computer networks, mixed reality, and multimedia.

**DIANA P. TOBÓN V.** received the B.Sc. degree in electronic engineering and the M.Sc. degree in engineering, with a subject on the quality of service in wireless body area networks (WBANs), from the University of Antioquia, Medellín, Colombia, in 2008 and 2012, respectively, and the Ph.D. degree from INRS-EMT affiliated with MuSAE, and Optical Zeitgeist Laboratories, under the supervision of Prof. Tiago Henrique Falk and co-supervision of Prof. Martin Maier. She continued as a Postdoctoral Fellow with the MuSAE Laboratory for three months working on machine learning applied in neuromarketing. She was a Postdoctoral Fellow with the Multimedia Communications Research Laboratory, University of Ottawa, working on emotional state detection under the supervision of Prof. Abdulmotaleb El Saddik for seven months. She was a full-time Professor with the Universidad del Valle, Cali, Colombia, for one semester. She is currently a full-time Professor with the Universidad de MedellÃn, Medellín. Her research interests include WBANs, signal processing of noisy signals, machine learning, and deep-learning techniques in the biomedical field.

**M. SHAMIM HOSSAIN** (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa, ON, Canada, in 2009. He is currently a Professor with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is also an Adjunct Professor with the School of Electrical Engineering and Computer Science, University of Ottawa. He has authored and coauthored more than 315 publications including refereed journals, conference papers, books, and book chapters. His research interests include cloud networking, smart environment (smart city, smart health), AI, deep learning, edge computing, the Internet of Things (IoT), multimedia for health care, and multimedia big data. Recently, he co-edited a book on *Connected Health in Smart Cities*, published by Springer. He has served as the Co-Chair, the General Chair, the Workshop Chair, the Publication Chair, and a TPC in several IEEE and ACM conferences. He is the Chair of IEEE Special Interest Group on Artificial Intelligence (AI) for Health with IEEE ComSoc eHealth Technical Committee. Currently, he is the Organizing Co-Chair of the Special Sessions with IEEE I2MTC 2022. He is also the Co-Chair of the 1st IEEE GLOBECOM 2021 Workshop on Edge-AI and IoT for Connected Health. He was the Co-Chair of the Special Session "AI-enabled technologies for smart health monitoring," held with IEEE I2MTC 2021. He was the Co-Chair of the 3rd IEEE ICME Workshop on Multimedia Services and Tools for Smart-Health (MUST-SH 2020). He is a recipient of a number of awards, including the Best Conference Paper Award and the 2016 ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) Nicolas D. Georganas Best Paper Award, and the 2019 King Saud University Scientific Excellence Award (Research Quality). He is on the Editorial Board of the IEEE Transactions on Instrumentation and Measurement (TIM), IEEE Transactions on Multimedia (TMM), IEEE Multimedia, *IEEE Network*, IEEE Wireless Communications, IEEE Access, *Journal of Network and Computer Applications* (Elsevier), *International Journal of Multimedia Tools and Applications* (Springer), *Games for Health Journal*, and *International Journal of Information Technology, Communications and Convergence* (Inderscience). He also serves as a Lead Guest Editor for *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), and *Multimedia Systems* Journal. Previously, he served as a Guest Editor for more than two dozens of Sis, including *ACM Transactions on Internet Technology*, *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), *IEEE Communications Magazine*, *IEEE Network*, IEEE Transactions on Information Technology in Biomedicine (currently JBHI), IEEE Transactions on Cloud Computing, Future Generation Computer Systems (Elsevier). He is a Senior Member of the ACM. He is an IEEE ComSoc Distinguished Lecturer (DL).

**ABDULMOTALEB EL SADDIK** (Fellow, IEEE) is a Distinguished University Professor and the University Research Chair with the School of Electrical Engineering and Computer Science, University of Ottawa. His research interests include the establishment of digital twins to facilitate the well-being of citizens using AI, AR/VR, and tactile internet, hence allowing people to interact in real-time with one another as well as with their digital representation. He has authored and coauthored ten books and more than 550 publications and chaired more than 50 conferences and workshop. He has received research grants and contracts totaling more than U.S. $20 M. He has supervised more than 120 researchers and received several international awards, among others, are an ACM Distinguished Scientist, a fellow of the Royal Society of Canada, a fellow of the Engineering Institute of Canada, and a fellow of the Canadian Academy of Engineers. He has received the IEEE I&M Technical Achievement Award, IEEE Canada C.C. Gotlieb (Computer) Medal, and A.G.L. McNaughton Gold Medal for important contributions to the field of computer engineering and science.