Journal of Information Hiding and Multimedia Signal Processing ©2023 ISSN 2073-4212
Ubiquitous International Volume **14**, Number **4**, December **2023**

Head Features-Based Deep Learning Approach for Recognizing Emotion, Gender and Age

Nippon Datta, Juel Sikder, Rishita Chakma and Rahul Kanti Das

Department of Computer Science and Engineering
Rangamati Science and Technology University
Rangamati, Bangladesh
nipponrmstu.cse@gmail.com, juelsikder48@gmail.com, rishita.eee@gmail.com, rahulcse0077@gmail.com

Received July 2023; revised November 2023; accepted November 2023

Abstract. *At the same time, recognizing the emotions, gender and age of a person is a vital issue in real life. This research focuses on identifying the emotions along with age and gender of different ages male, female, common gender and children. In this study, a modified convolutional neural network (M-CNN) has been proposed for feature extraction where classification is followed by a Multi Support Vector Machine (M-SVM). Instead of detecting the face, a head-detecting technique is also used in the prediction part. To detect the head a system is proposed using the You Only Look Once Version 3 (YOLOV3) network. It helps to extract the feature from the whole head instead of the only face. Three datasets named AR-2022, FER-2022 and GR-2022 are introduced in this study. The common gender is also included in the datasets. In the case of age classification, the ages are categorized into 11 categories depending on the similarity of their features, gender into 3 categories and emotions into 7 categories. The proposed model is compared with other existing models and current research, where the proposed model gives better accuracy and the accuracy of classification of emotions, age and gender are 97.47%, 98.43% and 96.65% respectively.*
Keywords: Deep Learning, SVM, YOLOV3, Emotion, Age.

1. **Introduction.** In communication, the face is one of the most significant parts which plays an important role in people's daily life. In recent research, most of the researchers preferred to extract the features from the frontal face and then identify the emotions, age and gender [1]. But in this study, head detection is focused in lieu of face and extracting the features from the detected head. As a result, the features are gathered from the ear, hair along with the face. In this research, a modified convolutional neural network is introduced where classification is followed by a multi-support vector machine in order to create the desired classifier. A callback function has been used to save the best-trained model. The head detection technique was used in the prediction part. A head detector is created using YOLOV3. The head detector detects the head from the whole image and the detected head is used to recognize emotions, age and gender along with their accuracy. The significant contributions of this research are listed below.

- Modification of convolutional neural network for the purpose of feature extraction where multi-support vector machine (M-SVM) used for classification purposes.
- Focused on the head instead of only focusing on the face. As a result, it collected features also from hair, and ears along the face.
- Introduced with YOLO version 3 network for detecting the head from the test image.

- Introduced three datasets also for the classification purpose which are Face Emotion Recognition (FER-2022), Age Recognition (AR-2022), and Gender Recognition (GR-2022).
- Introduced also with the third gender in this research and classify the age level into 11 categories which conclude different stages of life.
- Comparison with other existing models where the proposed model gives better accuracy.
- In order to test the accuracy of the proposed model, the model is trained with several datasets which are currently available.
- Comparison with other current research where the proposed system provides better accuracy.

The following is the organization of this study. The contributions of recent relevant researchers are detailed in Section 2. Section 3 briefly describes the proposed system. Section 4 illustrated the result and analysis of the proposed system, and the last Section 5 focused on the future plan and conclusion of this research.

2. **Literature Review.** In order to recognize emotions along with gender and age, several works have been proposed. some of them are described below. Levi et al. [2] introduced a model that recognizes gender along with age using CNN. In their research, they classified the ages into nine categories and gender into two. Lapuschkin et al. [3] proposed a classifier using a layer-wise propagation technique where face traits are employed for age and gender prediction. On the difficult Adience dataset, they examined the different pictures and discussed how these affect the performances of their proposed system. Duan et al. [4] suggested a structure that is the combination of CNN and ELM. It is able to recognize gender and age. The hybrid design took advantage of their strengths: CNN was utilized to extract features, while ELM was used to classify the intermediate outputs. Hosseini et al. [5] suggested a CNN-based architecture for combined age-gender categorization, in which Gabor filter responses were used as input. The architecture has been taught to categorize the input photographs into eight age groups and two gender categories. Mansanet et al. [6] suggested a Local-DNN. Local-DNN basically refers to the Local Deep Neural Network (Local-DNN) model which is the combination of two major concepts where the concepts are deep architectures and local features respectively. They tested their method on two major benchmarks, demonstrating that their proposed method beats existing neural network-based approaches in both benchmarks, achieving state-of-the-art scores in both. Jain et al. [7] introduced a neural network-based approach for recognizing facial Emotions. The suggested network design consists of Convolution layers followed by an RNN. The combined model captures relations within face images and uses the recurrent network to evaluate temporal relationships in the images during classification. The model is based on two public datasets.

3. **Methodology.** In this section, the proposed system is briefly described. The proposed system is divided into two phases -
1. Head detection stage 2. Classification stage

The working flow of the main diagram is as follows:
Step-01: Take a test image as input.
Step-02: Reshaped the test image into *(416,416,3)* dimension as it is the head detector model dimension.
Step-03: Load YOLOV3 head detector in order to detect the head.
Step-04: Get the detected head.
Step-05: Reshaped the detected head into*(128,128,3)* dimension as it is the dimension of

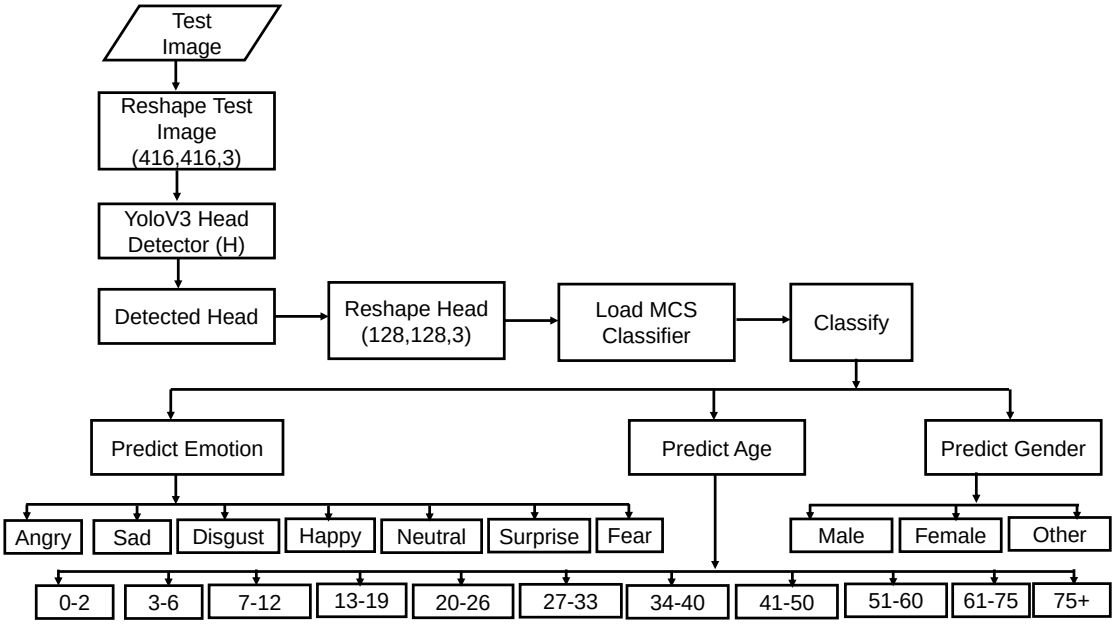

FIGURE 1. Main block diagram of the proposed system.

MCS classifier.
Step-06: Load MCS classifier which is pre-trained.
Step-07: Classify the detected head.
Step-08: Predict the label of emotion, age and gender.
The details of the head detection and MCS classifier are mentioned in the following section.

3.1. **Head Detection Stage.** As mentioned earlier, the proposed system will identify the emotion, gender and age from the head instead of the face, so it the first work to detect the head.

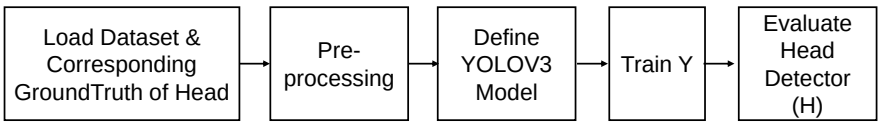

FIGURE 2. Block diagram of head detection.

3.1.1. *Load Dataset and Corresponding GroundTruth.* The system loads the image datasets and their corresponding ground truth in order to evaluate a head detector.

3.1.2. *Pre-processing.* As the dataset and corresponding ground truth are loaded then the dataset images are reshaped into *(416,416,3)* dimensions. Then adjust the images with their corresponding ground truth.

3.1.3. *Define YOLOV3 model.* In this stage, the YOLOV3 model is defined in order to create the head detector. YOLO stands for "You Only Look Once". It can detect the image with its locations and also predict the object [8]. In this study, the YOLOV3 model is denoted by Y.

3.1.4. *Train Y.* After defining the model the next step is to train model Y in order to get the head detector successfully.

3.1.5. *Evaluate Head Detector.* In this section the evaluating process of YOLOV3 will be illustrated and at the end, the head detector will be defined by H.

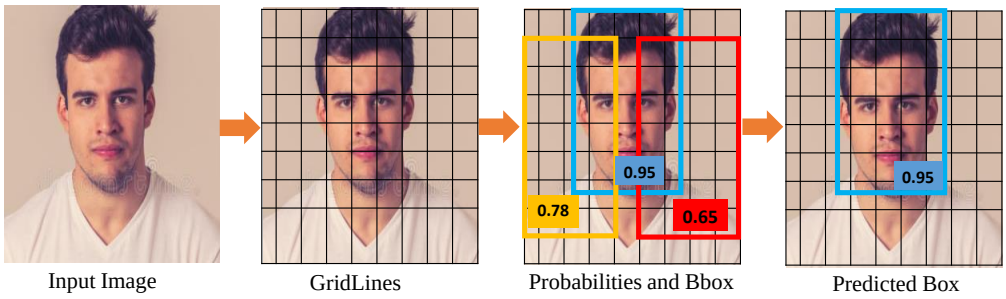

FIGURE 3. Head detection from an image.

- Input Network
  The input consists of a collection of photos with the shape *(n, 416, 416, 3)*. Here n is the picture number, the width and height are the next two numbers, and the number of channels is the last number.
- Detection at 3 Scales
  YOLO version 3 detects 3 different sizes at 3 separate locations throughout the network. Layers 82, 94, and 106 are the three different detection locations. At those different regions of the Network, the network downsamples the input picture by 32, 16, and 8 factors. Assuming that, the stride size is 32 and the input network size is *416 by 416*, then a size *13 by 13* output will be generated [9].
- Detection Kernels
  To create a result at these three points in the network, YOLO version 3 employs one-by-one detection kernels where one-by-one convolution is used to downsample the images into 13×13, 26×26, *and* 52×52. The depth of the detection kernel may be determined using the following equation:

$$(e * ((5 + f)) \tag{1}$$

  Where, e = *bounding box number,* $5 + f$ *bounding box attributes, And* f = *class numbers.* The proposed YOLO version 3 network is trained on the man face dataset then $f = 1$ and the attributes number is $5+1 = 6$. The resulting total attributes are as follows:

$$(3 * ((5 + 1)) \tag{2}$$

  As a result, the generated feature map in this network downsampled dimension depth in 3 separate points that include 18 bounding box characteristics for the man-face dataset.
- Illustrate Detection Grid Cells
  Yolo version 3 predicts three bounding boxes for each feature map cell. Yolo version 3's goal during training is to forecast an item through one of its bounding boxes if the object's centroid fits within the cell's visual field. To begin, it must determine first where the bounding box is present. In order to do that, firstly consider the 1st detection scale which consists of 32 network strides. As computed, the *416 by 416* input picture is downsampled into a *13 x 13* grid of cells. The developed output feature map is now represented by this grid. When the bounding box of ground truth is discovered, YOLO version 3 assigns responsibility for forecasting this object to the center cell. And the objectivity score 1 for that cell.

TABLE 1. The width and height of 9 anchors for the man face dataset.

Scale-1	Scale -2	Scale-3
116 * 90	30 * 61	10 * 13
156 * 198	62 * 45	16 * 30
372 * 326	59 * 119	33 * 60

- Anchor Boxes
  To calculate the bounding boxes Anchors or priors are pre-defined default bounding boxes used by YOLO version 3. There are a total of 9 anchor boxes in use. Each scale has three anchor boxes. It means that each feature map's grid cell may predict three bounding boxes using three anchors at each scale. In YOLO version 3, k-means clustering is used to determine these anchors. They are grouped according to the scale at three separate places in the network.
- Predicted Bounding Boxes
  Anchors are bounding boxes priors and they were calculated by using K-means clustering. To anticipate height as well as width YOLOV3 calculates offsets to the predefined anchors.
- Objectness Score
  YOLO version 3 generates bounding boxes with their properties for each cell. For each class, this bounding box may belong to given attributes which are *Ax, Ay, Aw, Ah, P0, and value of confidences*. These outputs are then utilized to determine the anticipated real width and height of bounding boxes using specified anchors, as well as to choose anchor boxes by computing scores. The so-called objectness score is *P0* in this case. It is the center cell's value and this cell has an objectness score of 1. The objectness score determines the center cell which assists in predicting the head from the input image.
- Detect and Crop Image Classification
  The detected head will be used for the further classification process.

FIGURE 4. Cropped and detected image.

3.2. **Classification Stage.** In this section, the details of evaluating the classifier and the functions related to it are described.

3.2.1. *Load Datasets and Image Categories.* In order to create a well-desired classifier, it is essential to load the datasets and their corresponding image categories. In this study, the three datasets FER-2022, AR-2022 and GR-2022 are loaded with their corresponding classes in order to construct a classifier.

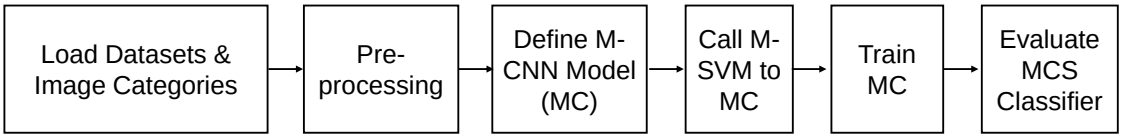

FIGURE 5. Block diagram of evaluating MCS classifier.

3.2.2. *Image Pre-processing.* As the proposed model of this study preferred an input shape of *128 by 128* dimensions it is necessary to convert the dataset's images into these shapes. The images are pre-processed by shearing and horizontal flipping. the range of shearing is kept at 0.25 which assists computers in identifying how people look at different substances from different rotations. [10].

3.2.3. *Define modified CNN model.* CNN refers to Convolutional neural networks that are used for image classification, object detection, segmentation and a variety of other applications [11, 12]. CNN architecture is divided into two components. Feature extraction is a convolution tool that focuses on finding and isolating numerous characteristics from a picture. A fully connected layer that uses the convolution process outputs to forecast the image's class based on the retrieved information [13]. In this study, a modified convolutional neural network (M-CNN) is used for feature extraction. It is essential to extract the features from the head in order to recognize the categories of emotion, age and gender. The M-CNN that extracts the features consists of five convolutional layers along with relu, batch normalization, max-pooling and dropout function. After Convolution layers flattened layer is applied. Then two fully connected layers are used. In this study, the modified CNN is denoted by MC.

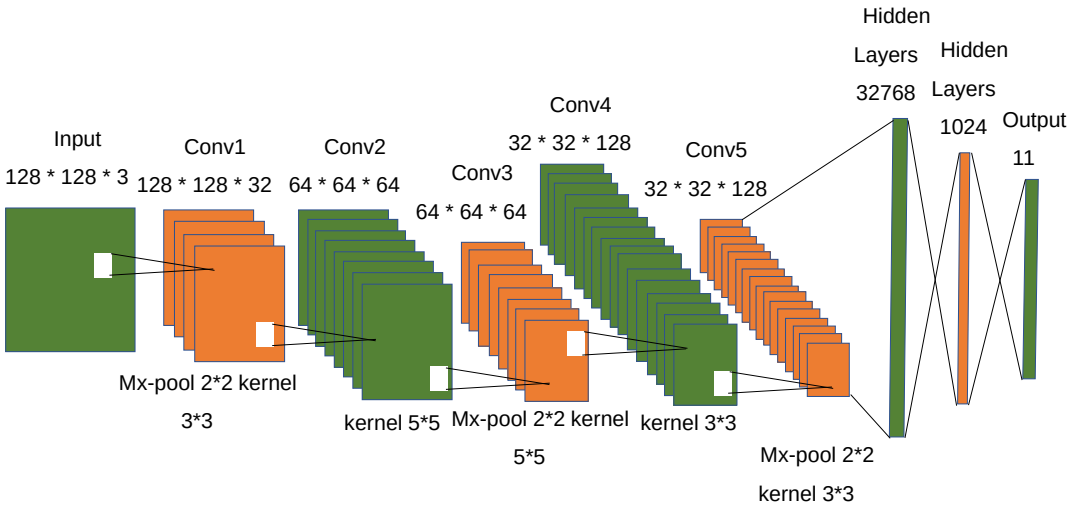

FIGURE 6. Block diagram of proposed modified CNN.

3.2.4. *Call Multi Support Vector Machine.* After extracting the features a multi-support vector machine (M-SVM) is used instead of the default CNN classifier to create a classifier. SVM is one of the most widely used algorithms used for regression and classification [14, 15].

3.2.5. *Train MC.* In this section, the model is trained. A callback function is also used at the time of model training. The callback function assists the classifier in saving the weights when val_loss is decreased [16].

Algorithm 1 Proposed Algorithm for Emotion, Age and Gender Recognition.

$Input \leftarrow TrainSetX1, TestSetX2$
$p \leftarrow themodellearningrate$
$q \leftarrow iterationstep$
$q \leftarrow 0$
$m \leftarrow Maxnumberofiterations$
$n \leftarrow numberofimageperiteration$
Output:
W, weights of CNN
Begin:
1. Define the initial layer of architecture using Modified CNN
2. Set head layers, CNNdropout, CNNflatten, CNNmaxpooling, CNNdense, CNNbatchnormalization, SVMlayer
3. Initialize parameters: p, m, n
4. Image Resizing in the training set by 128 * 128
5. Training and calculating weights of M-CNN
for all $q \leftarrow 1$ to m **do**
  1. Randomly take Batches from X1
  2. Use Forward Propagation, calculate loss using p
  3. Use Back Propagation and Update W using
end for

3.2.6. *Evaluate modified CNN-SVM Classifier.* The best model of training will save in this step and now it is ready to classify emotion, age and gender. In this study, the classifier is denoted as MCS classifier.

4. **Results and Analysis.** The testing data, training data, results and the performance of the proposed method are narrated in this section.

4.0.1. *Datasets.* Three datasets have been introduced in order to train the proposed model. All images of those datasets are collected from authors' friend zones and different social sites like Instagram, Google, and Facebook. After that, the data augmentation technique is used to increase the number of dataset images. Data augmentation is a technique in image processing that assist in increasing the number of image number in a dataset. [17]. The three datasets are AR-2022, FER-2022, and GR-2022. The FER-2022(Face Emotion Recognition) dataset contains 14,216 images in 7 categories. The seven categories are 'Disgust', 'Angry', 'Happy', 'Fear', 'Neutral', 'Surprise', and 'Sad'. The GR-2022(Gender Recognition) dataset contains 9,893 images in 3 categories. The three categories are 'Male','Female', and 'Other'. The AR-2022(Age Recognition) dataset contains 22,597 images in 11 categories. The 11 categories are' 0-2', '3-6',' 7-12',' 13-19',' 20-26',' 27-33',' 34-40',' 41-50',' 51-60',' 61-75',' 75+', respectively.

Depending on the following three factors we have decided to take such age ranges,

- Practically - We think practically about the age ranges. Too many classes make the problems more challenging and require a larger dataset. So, we have decided to take a few classes.
- Common Age Range – Common Age ranges like Baby, Preschooler, Teenager, Young, Young Adult, Adult, Senior Adult, Old, etc. In such kind of categorization, 0-2:Baby; 3-6: Preschooler; 7-12: School-Age Child; 13-19: Adolescent; 20-26: Young; 27-33: Young Adult; 34-40: Pre-Middle Aged Adult; 41-50: Middle-Aged Adult; 51-60: Adult; 61-75: Senior Citizen; 75+: Elderly.

TABLE 2. The comparison of the proposed method with other existing models.

SL.	Model	Accuracy		
		Emotion	Age ,	Gender
[1]	VGG16	49.51%	39.27%	80.49%
[2]	InceptionResnetV2	65.18%	59.23%	92.81%
[3]	DenseNet121	65%	59.48%	91.17%
[4]	CNN	85.28%	9.87%	95.39%
[5]	AlexNet	14.80%	9.44%	35.05%
[6]	VGG19	48.96%	32.77%	73.65%
[7]	MobileNetV2	83.19%	57.63%	95.88%
[8]	ResNet50	16.59%	10.53%	50.80%
[9]	InceptionV3	70.91%	58.09%	92.95%
[10]	DenseNet169	77.51%	64.42%	94.23%
[11]	Proposed Method	97.47%	98.43%	96.65%

- Augmentation – It assists us in increasing the dataset while we have a smaller size of a data sample in a certain class.

4.0.2. *Comparison with existing models.* The proposed model is compared with other existing models, where the proposed model gives better accuracy. Table 3 shows the list of accuracy using the existing model.

4.0.3. *Result of proposed System.* The proposed techniques improve validation accuracy while cutting down on computation time and loss. This neural network-based architecture is tested with different images from outside of the proposed datasets and the results are as follows.

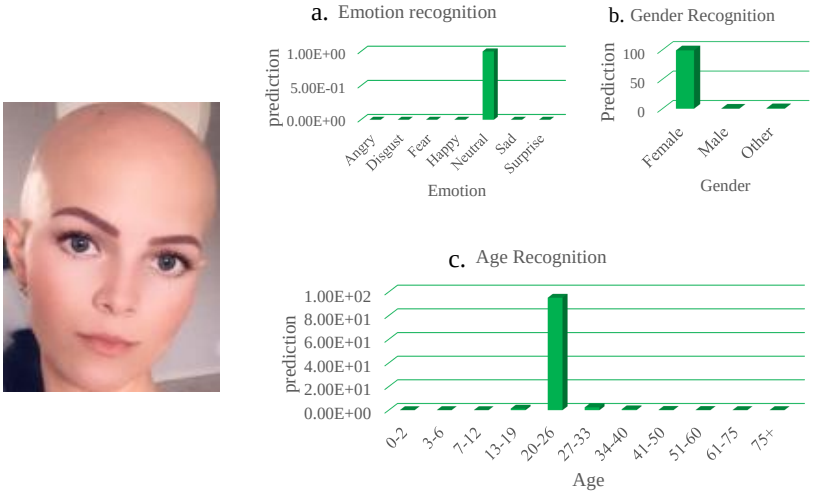

FIGURE 7. Prediction result on the input image 1.

4.0.4. *Result using existing datasets.* The proposed system is trained and tested using currently available datasets to test the accuracy of the proposed system which is shown in Table 4.

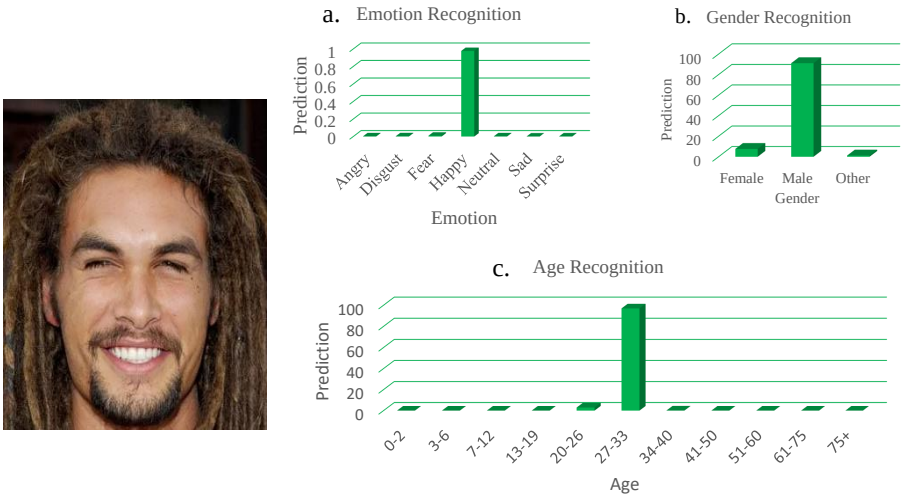

FIGURE 8. Prediction result on the input image 2.

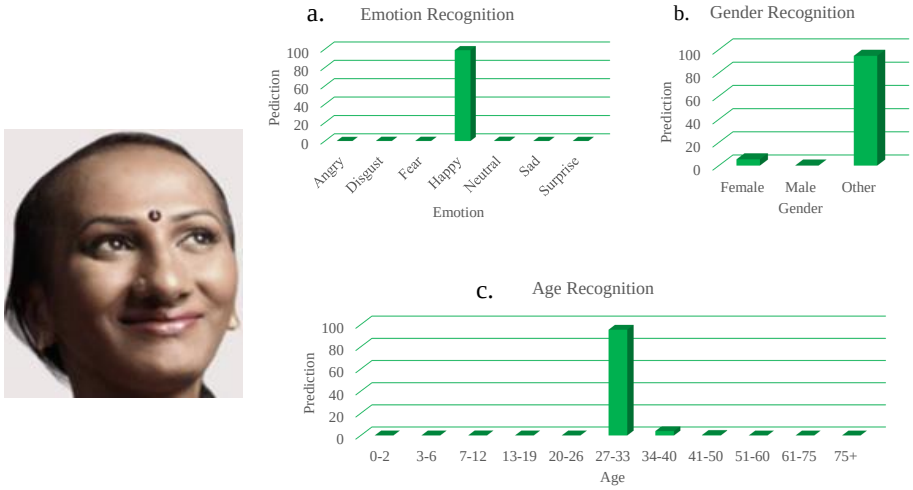

FIGURE 9. Prediction result on the input image 3.

TABLE 3. Result using existing datasets.

SI NO	Dataset Name	Type	Accuracy
1.	FER - 2013	Emotions	88.56%
2.	GAFace Dataset	Gender, Age	91.67%, 87.95%
3.	CK+	Emotions	94.45%
4.	IoG Dataset	Age , Gender	70.10%, 88.90%
5.	RafD	Emotions	93.65%
6.	The LFW-Gender Dataset	Gender	92.25%

TABLE 4. Comparison of current research papers.

Article	Detection	Dataset	Methodology	Accuracy
[18]	Gender	GENDER-FERET	VGGFace CNN	97.45%
[19]	Emotions	MMI, Multipie, FERA,CK+, DISFA, FER-2013	CNN	94.7%, 77.9%, 55%,76.7%, 47.7%, 93.2%
[20]	Emotions	FER-2013	CNN	65%
[21]	Emotions	CK+, JAFFE	SBN-CNN	96.8%, 95.24%
[22]	Emotions	CASME II, MMI	CNN-LSTM	60.9%, 78.61%
[23]	Gender, Age	GAFace Dataset	Deep-CNN	90.33%, 80.1%
[24]	Emotions	AffectNet, RAF-DB	ACNN	54.8%, 80.54%
Proposed Mehodology	Emotions, Gender, Age	FER-2022, AR-2022 GR-2022	Head Detector (H), MCS Classifier	97.47%, 96.65%, 98.43%

4.0.5. *Comparison with current research.* Table 4 shows a comparison of current research with the proposed method where the proposed system shows better accuracy than others.

5. **Conclusions.** In order to recognize facial emotion along with age and gender is a critical issue nowadays. Rather than focusing on a single point, this study focused on a modified method to recognize emotion along with gender and age using a multi-support vector machine (M-SVM), modified convolutional neural network (M-CNN) and YOLO version 3 network. This study also introduced the FER-2022 dataset, GR-2022 dataset, and AR-2022 dataset for the training purpose of emotion, gender and age classifiers respectively. The result of the proposed model is compared with existing models where the proposed model gives the recognition rate of emotion, gender and age are 97.47%, 96.65% and 98.43% respectively.
The future version of that research will focus on facial benchmarks, especially for more accurate results and predictions. To ensure more flexibility of classifiers, some other advanced algorithm-based approaches will use.

Conflict of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and materials. In this study, three datasets have been introduced named FER-2022, AR-2022, and GR-2022 and the datasets are available in "https://github.com/Nippon Age-and-Gender- Datasets" which will be publicly available after the publication of the paper.

REFERENCES

[1] K. Lawrence, R. Campbell, and D. Skuse, "Age, gender, and puberty influence the development of facial emotion recognition," *Frontiers in psychology*, vol. 6, p. 761, 2015.
[2] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–42, 2015.
[3] S. Lapuschkin, A. Binder, K.-R. Muller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 1629–1638, 2017.
[4] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning cnn–elm for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, 2018.

[5] S. Hosseini, S. H. Lee, H. J. Kwon, H. I. Koo, and N. I. Cho, "Age and gender classification using wide convolutional neural network and gabor filter," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, pp. 1–3, IEEE, 2018.

[6] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters*, vol. 70, pp. 80–86, 2016.

[7] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[8] H. Chen and Z.-M. Lu, "Contraband detection based on deep learning," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 13, no. 3, pp. 165–177, 2022.

[9] Q.-C. Mao, H.-M. Sun, Y.-B. Liu, and R.-S. Jia, "Mini-yolov3: real-time object detector for embedded applications," *Ieee Access*, vol. 7, pp. 133529–133538, 2019.

[10] T.-M. Liang, M.-S. Chiu, Y.-C. Wu, M.-T. Yeh, C.-H. Hsu, and Y.-N. Chung, "Applying image processing technology to face recognition," *J. Inf. Hiding Multimed. Signal Process*, vol. 13, no. 2, pp. 106–112, 2022.

[11] J. Sikder, U. K. Das, and R. J. Chakma, "Supervised learning-based cancer detection," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021.

[12] J. Sikder, N. Datta, S. Tripura, and U. K. Das, "Emotion, age and gender recognition using surf, brisk, m-svm and modified cnn," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6, IEEE, 2022.

[13] M. Xu, Y.-P. Feng, and Z.-M. Lu, "Fast feature extraction based on multi-feature classification for color image.," *J. Inf. Hiding Multim. Signal Process.*, vol. 10, no. 2, pp. 338–345, 2019.

[14] M.-S. Chiu, C.-C. Lin, C.-S. Cheng, M.-T. Yeh, Y.-N. Chung, and C.-H. Hsu, "Applying image processing algorithm to dynamic face detection.," *J. Inf. Hiding Multim. Signal Process.*, vol. 12, no. 4, pp. 207–216, 2021.

[15] J. Sikder, R. Chakma, R. J. Chakma, and U. K. Das, "Intelligent face detection and recognition system," in *2021 International Conference on Intelligent Technologies (CONIT)*, pp. 1–5, IEEE, 2021.

[16] J. Sikder, N. Datta, and D. Tripura, "A deep learning approach for recognizing covid-19 from chest x-ray using modified cnn-bilstm with m-svm.," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6, IEEE, 2022.

[17] A. W. Lohmann, "Image rotation, wigner rotation, and the fractional fourier transform," *JOSA A*, vol. 10, no. 10, pp. 2181–2186, 1993.

[18] F. Simanjuntak and G. Azzopardi, "Fusion of cnn-and cosfire-based features with application to gender recognition from face images," in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pp. 444–458, Springer, 2020.

[19] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10, IEEE, 2016.

[20] A. Agrawal and N. Mittal, "Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *The Visual Computer*, vol. 36, no. 2, pp. 405–412, 2020.

[21] J. Cai, O. Chang, X.-L. Tang, C. Xue, and C. Wei, "Facial expression recognition method based on sparse batch normalization cnn," in *2018 37th Chinese control conference (CCC)*, pp. 9608–9613, IEEE, 2018.

[22] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2017.

[23] X. Wang, A. M. Ali, and P. Angelov, "Gender and age classification of human faces for automatic detection of anomalous human behaviour," in *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, pp. 1–6, IEEE, 2017.

[24] G. Yolcu, I. Oztel, S. Kazan, C. Oz, K. Palaniappan, T. E. Lever, and F. Bunyak, "Facial expression recognition for monitoring neurological disorders based on convolutional neural network," *Multimedia Tools and Applications*, vol. 78, pp. 31581–31603, 2019.