Gender-Based Crowd Categorization and Counting Employing YOLOv8

Jalil Akbarzai[1,2], Muhammad Qasim[1,2], Zainab[1,2], Sayed Shahid Hussain[2] and Shahzad Anwar*[2,3]

[1] Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan
[2] Artificial Intelligence in Healthcare Lab, National Centre of Artificial Intelligence, Pakistan.
[3] Department of Mechatronics Engineering, University of Engineering and Technology Peshawar, Pakistan

***Abstract*-**. Gender-based crowd-counting is a complex and important area of research thus appealing to a wider research community, which has notable applications in fields such as security surveillance, worship places, hotels etc. which is vital for the comprehension of demographics, public safety and city planning efficiently. This research incorporates an advanced deep learning algorithm YOLOv8 famous for its high accuracy and efficiency for object detection. The dataset was annotated considering various demographic factors, such as ethnic diversity and attire variations to have robustness and reliability in gender classification. The main purpose of the developed method is to improve how the crowd is analyzed commissioning advanced methods of computer vision aiming to improve the decision-making process in urban management, enhancing recourses allocation in event planning. The proposed method paves the way for more advanced crowd analysis techniques for practical scenarios.

Index Terms- Crowd Counting, Gender Classification,

YOLOv8, Computer Vision

I. INTRODUCTION

In many fields, computer vision has advanced considerably in the past few years. Commissioning computer vision accurately for human counting and gender classification have importance in many day-to-day applications. To be more specific, crowd counting is important to consider in building public spaces to accommodate a specific number of individuals, for instance, places for worship, hotels, playgrounds, airports and different types of stations for better management and security monitoring etc. [1]. Currently various machine learning (ML) algorithms have been deployed for such tasks. AdaBoost based on Harr features [3], SVM based on Hog [4], and LBP [5] and others are examples of traditional ML detection techniques. The primary idea of these detection systems is based on manually derived characteristics. Typically, the method requires extracting features from videos/images, building a classifier for detection/classification, and obtaining the targets. However, there are associated limitations with respect to accuracy and efficiency. These traditional methods employ complicated preprocessing steps and complex algorithms [2]. Comparing this, 'You Only Look Once' (YOLO) object detection framework is proposed in the study, which is well known for real-time performance and its high accuracy in object detection. YOLO framework has many benefits compared with traditional methods, for instance, the ability to handle large datasets and preprocessing pipelines. In addition, the system performance is improved by training models on custom datasets which are tailored for certain applications [2]. YOLOv8 utilizes a similar backbone to YOLOv5. A complete architecture of YOLOv8 is shown in Figure 1, with some modifications in the CSPLayer, now denoted by the C2f module. The C2f module, which is composed of a two-convolution cross-stage partial bottleneck, combines high-level features with contextual information to improve detection accuracy. YOLOv8 uses an anchor-free model with decoupled heads to independently handle object detection, classification, and regression tasks. This specific design lets each branch focus on its particular task, leading to improved overall model accuracy. The activation function for object scores in the YOLOv8 output layer is the sigmoid function, which represents the likelihood that an object is inside the bounding box. Class probabilities are expressed as the probability that an object belongs to each potential class, and the softmax function is employed to represent these probabilities [6]. In this study data has been collected from various sources and have carefully been labeled for accurately classifying gender and individual counting, which covers different scenarios and demographic distribution.

The paper has been arranged as per following: an overview of the literature around gender classification and crowd counting is presented in section II. Methodology, data collection, preprocessing is discussed in section III followed by result, discussion and potential application in section IV. Finally, section V concludes this manuscript.

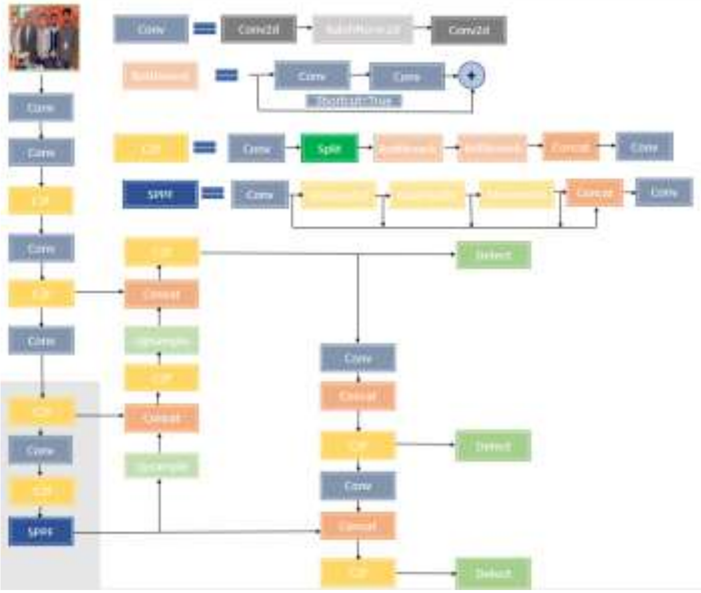

Figure 1: Architecture Diagram of YOLOv8 [6].

II. LITERATURE REVIEW

The literature on crowd counting is extensive, with numerous researchers conducting in-depth experiments. This section highlights some of these studies, detailing the techniques used and the results obtained. For example, Dasari et al. [7] proposed a Facial Analysis System for Gender Identification and Counting. The authors introduced an advanced dimension to face detection and gender classification by incorporating gender counting. Their study introduces Multi-Task Cascaded Convolutional Networks (MT-CNN) and Convolutional Neural Networks (CNN) to achieve the project objectives accurately and efficiently. In another study, Singh [8] presented an Embedded VGG 22 model for gender classification in crowd videos. To classify crowd footage based on gender, the study developed and optimized an embedded VGG model. According to their study even though used VGG models give reliable feature extraction for image classification, their high processing requirements make them unfeasible for real-time analysis in embedded systems with limited resources. Chen et.al [9] suggested a multi-scale dilated convolution of a convolutional neural network for crowd counting. This research presents a population density estimate approach using a multi-scale dilated convolution feature maps fusion convolutional neural network (Multi-scale-CNN). Based on their research the algorithm considers the advantages and disadvantages of other crowd-counting algorithms currently, which are being used. The density map regression method used by the algorithm counts the crowd in any resolution image without the need for pre-processing. The convolutional neural network used has been divided into two sub-networks: a multi-scale dilated CNN for convolution feature extraction and a convolutional neural network as the front-end for feature extraction. This approach addresses the challenges of human size variation and small target counts in crowd photos. The research considered the following datasets: ShanghaiTech, UCF_CC_50, and WorldExpo'10. For crowd counting, Deepak Babu and his mates used an unsupervised learning method. The study utilized the layered auto-encoding convolution model. Due to the large amount of unsupervised learning, the grid winner-take-all (GWTA) strategy was used to develop this model. ShanghaiTech-A dataset was used to test this system, resulting in an MSE of 229.4 and an MAE of 154.7. Using UCF-CC-50 dataset, the same system displayed MAE and MSE values of 433.7 and 583.3, respectively. On the other hand, this system performs over 25% worse than state-of-the-art methods [crowd counting]. The study also explored the concept of expanding networks, focusing on the hierarchical recursive tree model of the CNN density regressor. Their network is resourceful because it splits into recursive sub-networks to handle scene inconsistency [10].

Yuhong Li et.al [11] recommended a network for Congested Scene Recognition called CSRNet to give a data-driven and deep learning method that recognized highly congested scenes and performed accurate count estimation as well as presented high-quality density maps. CSRNet has two main components, a CNN for 2D feature extraction task and a backend CNN. CSRNet was used to compare four distinct datasets. It obtained an MSE of 115.0 and an MAE of 68.2 on the ShanghaiTech-A dataset. The MAE and MSE for ShanghaiTech-Part-B were 10.6 and 16.0, respectively. It achieved an MAE of 266.1 and an MSE of 397.5 on the UCF-CC-50 dataset. Similarly, CSRNet achieved an MSE of 1.35 and an MAE of 1.07 on the UCSD dataset. The obtained MAE for the WorldExpo'10 dataset was 8.6.

M. Hossain et.al [12] in an article have considered the issue of crowd counting in images. Given an image of a crowded scene, their aim was to estimate the density map of this image, where each pixel value in the density map corresponds to the crowd density at the corresponding location in the image. A deep learning-based crowd counting model was developed to localize crowd, having scale difference, their model has four sub modules: (1) A multi-scale feature extractor (MFE), (2) Global scale attention (GSA), (3) Local scale attention (LSA) and (4) fusion network (FN). Multi-scale convolutional neural network (MSCNN) was proposed by Zeng et al. [13] to count the number of people in a picture frame. The MSCNN is capable of taking out scale-relevant features from the multi-scale blobs. The approach was cost effective. The technique was tested using the Shanghai-Tech dataset which got comparable results [13]. Haroon Idrees et al. [14] proposed a competitive dataset and crowd-counting model. This deep CNN is designed for human counting, density map estimation, and human localization. The model handles highly dense crowd situations effectively. The model was evaluated on their own dataset UCFQNRF which have 1535 images with total 1251642 number of annotated people count and have following MAE 132.0 and MSE 191.0. A crowd counting methodology based on semantic analysis of human bodies and head count was proposed by Siyu Huang et.al [15]. Semantic analysis was first used to identify the individuals in the frame, and then the counting challenge was transformed into several parallel learning problems. A convolutional neural network was then used to estimate the number of pedestrians present.

Using the following datasets—ShanghaiTech-B, UCF-CC-50, UCSD, and WorldExpo'10—the results achieved were MAE values of 20.2, 1.0, 409.5, and 10.5, respectively, and MSE values of 35.6, 1.4, and 463.7.

D. Kang and Chan [16] proposed a novel crowd-counting model based on CNN that employs an image pyramid for training. The use of an image pyramid is highly beneficial for addressing distortion and scale variations of individuals. For testing purposes, their system was evaluated on three datasets, yielding the following results: MAE of 90.1 and MSE of 137.5 on ShanghaiTech-A, MAE of 12.4 and MSE of 22.0 on ShanghaiTech-B, and MAE of 1.16 and MSE of 2.29 on UCSD. This system solved human scalability problems using state-of-the-art approaches but lacks speed [16]. Xinya Chen Yanrui Bin Nong Sang Changxin Gao [17] proposed the Scale Pyramid Network for Crowd Counting. In this study, the authors introduced crowd-counting design known as the Scale Pyramid Network (SPN). Their approach includes using a single column structure as the foundation and extracting high-level features from many scales via parallel dilated convolutions at varying rates. On the ShanghaiTech Part A dataset, their model achieved results with a 9.5% lower MAE and a 13.5% improvement in MSE. On the TRANCOS vehicle counting dataset, the results showed a 5.9% reduction in GAME (0), a 10% reduction in GAME (1), a 24.5% reduction in GAME (2), and a 38.7% reduction in GAME (3).

Amirali Abdolrashidi, Mehdi Minaei, Elham Azimi, Shervin Minaee propose a deep learning framework, based on the ensemble of attentional and residual convolutional networks, to predict gender and age group of facial images with high accuracy rate [18]. Junaid Hussain Muzamal et al. proposed the use of Faster R-CNN-based detection for age and gender-based crowd counts. In order to overcome issues with crowd counting, gender recolonization, age estimation and localization of people in visual frames, their study used Faster R-CNN [1].

III. METHODOLOGY

This section outlines the proposed methodology for people detection and classification into male and female categories, as illustrated in Figure 1. The study aims to utilize existing monitoring cameras, commonly installed in malls, transport stations, and other areas where crowd management is challenging. Overhead monitoring video footage serves as the system's input. The algorithm processes the video footage to detect individuals within the frame and subsequently classifies them into male and female categories. The output provides information on crowd management, including counts of males, females, and the overall total.

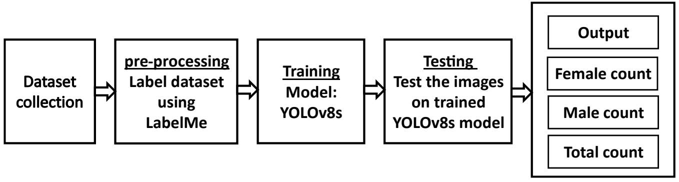

Figure 2: Block diagram of proposed work

A. Dataset

The dataset has been generated by collecting the images from different sources and has been manually labeled for ensuring the accuracy having total of 2000 hundred pictures on which data augmentation has been employed the data set has been split in the following order 80% for training and 15% for validation and 5% for testing.

B. Annotation:

Image annotation is the process of labeling data for computer vision and also known as image tagging it is done by marking the features in order to make an algorithm understand. The application used for labeling the data was LabelMe, which was developed at MIT and is accessible as an open-source project. Datasets for computer vision tasks such as classification, recognition, and segmentation are created using LabelMe, which supports annotations with rectangles, circles, polygons, lines, and points. After annotation, the dataset is uploaded to computer vision algorithm in its required format.

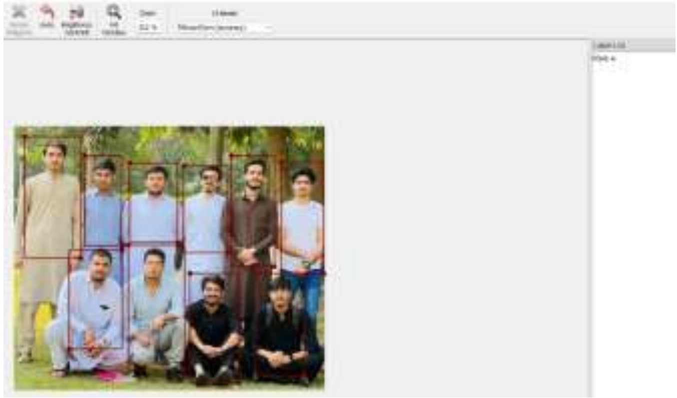

Figure 3: LabelMe interface

C. Algorithm selection

Deep learning techniques are extensively used for human detection and counting due to their speed and accuracy in processing large amounts of data and detecting targets effectively. These target detection techniques is broadly categorized into two types: the first category comprises the two-stage object detection R-CNN (Region-CNN) target detection algorithm, which includes Fast R-CNN, Faster R-CNN. This type algorithms are not suitable for real time detection because of its computational complexity. The second type of target detection technique is one stage object detection this type of technique includes SSD (Single MultiBox Detection), YOLO (You only look once) comparing this technique with two stage object detection algorithm it is faster and more accurate [20]. This study employed a comparative analysis of several algorithms, including YOLOv5, YOLOv8, YOLOv9, and YOLOv10. Among these, YOLOv8 demonstrated superior performance compared to the other algorithms tested.

IV. RESULTS AND DISCUSSION

This research aimed to address gender-based people counting using various detection algorithms. The four different types of algorithms were employed such as YOLOv5s, YOLOv8s, YOLOv9s and YOLOv10s, while evaluating the performance of the trained model MAP50&MAP50-95 (mean average precision) metrics were considered. After evaluation, YOLOv8s achieved a MAP50 of 98.5% and a MAP50-95 of 82%.

According to this comparison, YOLOv8s performed better than the rest of the applied algorithms. the obtained results indicate the better capacity to perform the task of gander-based people counting the results are present in table 1.

Table 1: Comparative Analysis of Algorithms used in this Study

Algorithms	mAP50	mAP50-95
YOLOv5s	98.5%	78.9%
YOLOv9s	98.3%	80.6%
YOLOv10s	98.3%	79.3%
YOLOv8s	98.5%	82%

To provide a comprehensive overview of the results, each employed algorithm was thoroughly tested. Figure 4 and Figure 5 represent the result YOLOv8s model. Each figure demonstrates the detected persons, their gender and individual and total count. Figure 4 highlights the model potential to correctly detect and classify classes based on gender in a less populated scene.

A challenging scenario is highlighted in Figure 5, with dense crowd, regardless of this complex scenario, the YOLOv8s model has precisely detected all the classes, keeping high accuracy in detection and classification. This indicates the model's effectiveness even in crowded environments. Overall, the YOLOv8s model is outstanding in accurately detecting and classifying individuals based on gender. performance of YOLOv8s model is evident in its capacity to accurately detect and perform classification tasks, significantly outperforming the other used algorithms.

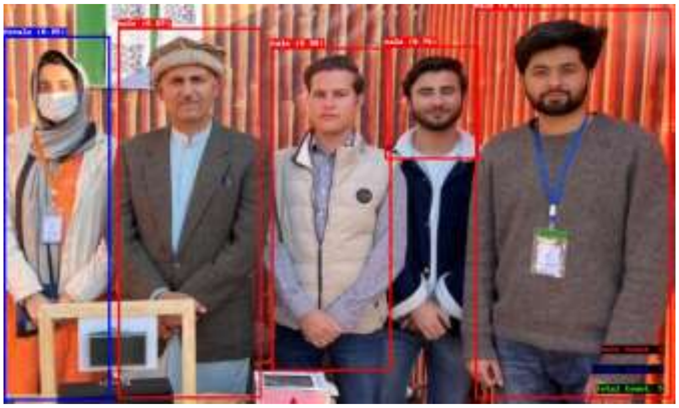

Figure4: YOLOv8s Result

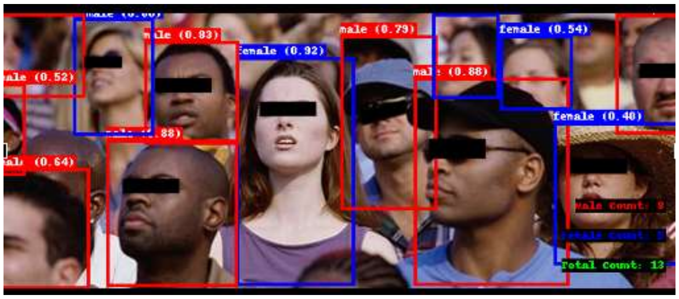

Figure5: YOLOv8s Result of dense crowd

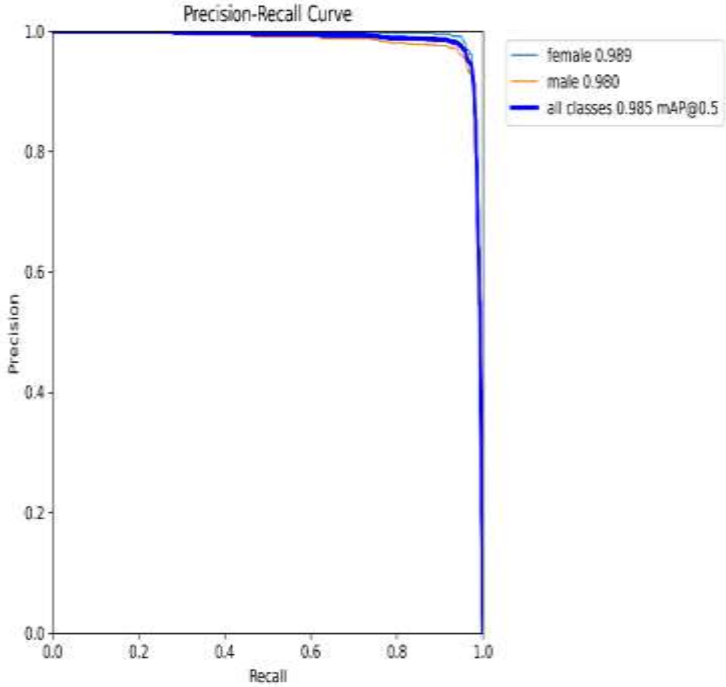

Figure 8: P-R curve of train model.

The Precision-Recall (PR) curves for each class in the crowd dataset illustrate the average precision for each class. Figure 8 indicates that the categories 'female' and 'male' perform exceptionally well, with average precisions of 0.989 and 0.980, respectively. Precision is calculated using Equation (1) while recall is calculated by Equation (2). The high average precision values for both 'female' and 'male' classes demonstrate the model's effectiveness in distinguishing these categories in the dataset. Overall, the mean average precision (mAP@50) for all classes is 0.985.

$$Precision = \frac{True\ Positives}{True\ positives + False\ Positives} - - - - - (1)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} - - - - - (2)$$

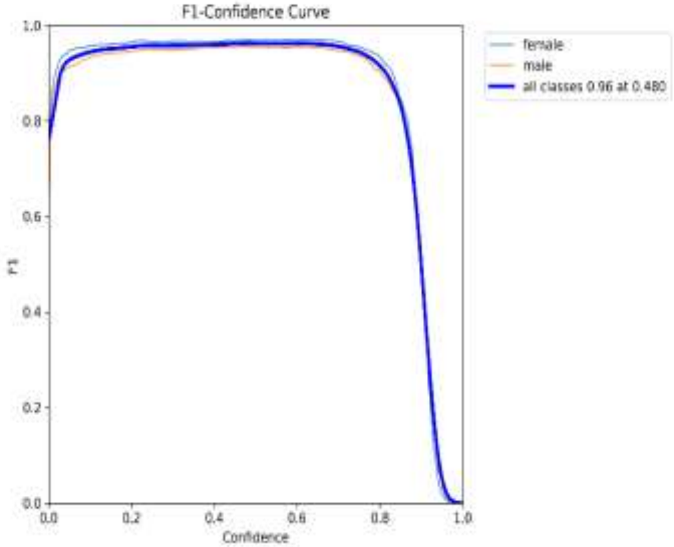

Figure 9: F1-curve

Figure 9 shows the F1 score curve for each class in the dataset. The graph shows that all included classes have achieved an F1 confidence score of 96%, indicating the model's effective performance in this task. Equation (3) outlines the calculation of the F1 score.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad - - - - - (3)$$

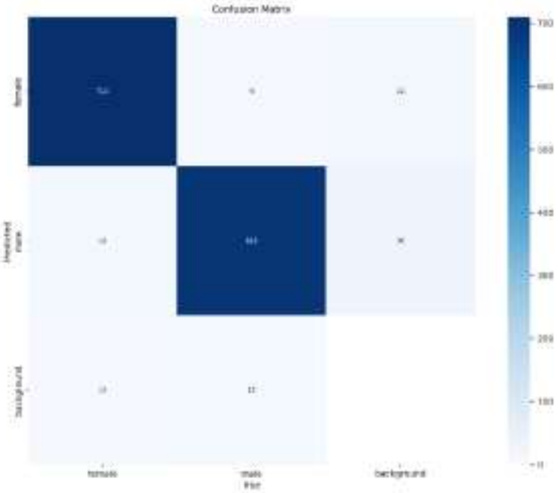

Figure 10: Confusion matrix of trained model YOLOv8s.

To evaluate the performance of the YOLOv8s model on the crowd dataset, a confusion matrix was generated for model evaluation, as depicted in the figure. The rows and columns of the confusion matrix represent the actual and predicted categories, respectively, with diagonal values indicating the number of correct predictions for each category.

Table 2: Comparison with previous findings

Reference	Technique	accuracy
[20]	SVM	81%
[21]	Adaboost	87.37%
[22]	CNN	87%
Proposed Technique	YOLOv8s	92.9%

The comparative analysis of our results indicates that the proposed methodology based on YOLOv8s algorithms, surpasses the accuracy of previously employed algorithms for crowd counting based on gender task. By utilizing the advanced detection and classification method, our approach has proven to be more reliable and effective. This achievement highlights the ability of our proposed work in field of computer vision technology by offering significant advancements over traditional methods. Gender based people counting has significant applications in security, public safety and retail and marketing. Security staff and various event organizers may have crowd control and safety by using this technology, ensuring a peaceful environment by keeping track of gander distribution, which is important in situation where imbalance of gender could cause safety risk. In the marketing and retail sector, gender-based crowd counting improves the product placement and advertising. By understanding the predominant gender of customers, shopkeepers are able to develop targeted promotions and adjust store layouts, enhancing the shopping experience and potentially increasing sales. Additionally, city planners are able to utilize gender-based crowd data to design more effective public transportation systems and facilities.

V. CONCLUSION

This research focused on real-time application while resolving issues related to people counting and gender-based classification by employing the YOLOv8 framework. By carefully selecting, classifying, and training a unique custom dataset with YOLOv8 models, a notable improvement was observed. The obtained results indicate how well YOLOv8 carry out the crowd analysis and give important information for the crowd management in different real-time situations. The proposed highlights how important it is to select the right algorithms and high-quality datasets for the sake of achieving precise and accurate results in computer vision applications. The recommended YOLOv8s algorithm demonstrated notable accuracy compared to traditional methods such as SVM, AdaBoost, and CNN. This significant improvement underscores the importance of selecting advanced algorithms and high-quality datasets to achieve accurate and reliable results in computer vision applications. This research highlights the important role of advanced algorithms in deep learning to address the challenges of crowd analysis and sets the foundation for future research.

ACKNOWLEDGMENT

We would like to express our gratitude to the AIH lab for their support and encouragement throughout this research work.

VI. REFERENCES

[1] Muzamal, Junaid Hussain, Zeeshan Tariq, and Usman Ghani Khan. "Crowd Counting with respect to Age and Gender by using Faster R-CNN based Detection." *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*. IEEE, 2019.

[2] A. Ranjan, N. Pathare, S. Dhavale and S. Kumar, "Performance Analysis of YOLO Algorithms for Real-Time Crowd Counting," *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet, India, 2022, pp. 1-8, doi: 10.1109/ASIANCON55314.2022.9909018.

[3] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 39, no. 06, pp. 1137-1149, 2017.

[4] Liu W. et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525, 2017.

[6] Wu, T.; Dong, Y. YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition. Appl. Sci. 2023, 13, 12977. https://doi.org/10.3390/app132412977.

[7] Sai, Dasari Rohith Satya, and G. Ravi Kumar. "A Facial Analysis System for Gender Identification and Counting." *Journal of Engineering Sciences* 15.02 (2024).

[8] Singh, Priyanka, and Rajeev Vishwakarma. "An Embedded VGG 22 Model for Gender Classification in Crowd Videos." *International Journal of Intelligent Systems and Applications in Engineering* 12.12s (2024): 11-33.

[9] Wang, Y., Hu, S., Wang, G. et al. Multi-scale dilated convolution of convolutional neural network for crowd counting. Multimed Tools Appl 79, 1057–1073 (2020). https://doi.org/10.1007/s11042-019-08208-6.

[10] Sam, Deepak Babu, et al. "Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[11] Li, Yuhong, Xiaofan Zhang, and Deming Chen. "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[12] Hossain, Mohammad, et al. "Crowd counting using scale-aware attention networks." *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019.

[13] Lingke Zeng, Xiangmin Xu, Bolun Cai, Suo Qiu, and Tong Zhang. Multi-scale convolutional neural networks for crowd counting. In 2017 IEEE International Conference on Image Processing (ICIP), pages 465– 469. IEEE, 2017.

[14] Idrees, Haroon, et al. "Composition loss for counting, density map estimation and localization in dense crowds." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[15] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, Rongrong Ji, and Junwei Han. Body structure aware deep crowd counting. IEEE Transactions on Image Processing, 27(3):1049–1059, 2018.

[16] Di Kang and Antoni Chan. Crowd counting by adaptively fusing predictions from an image pyramid. arXiv preprint arXiv:1805.06115, 2018.

[17] Chen, Xinya, et al. "Scale pyramid network for crowd counting." 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, 2019.

[18] Abdolrashidi, Amirali, et al. "Age and gender prediction from face images using attentional convolutional network." arXiv preprint arXiv:2010.03791 (2020).

[19] Wu, Shuai, et al. "Enhanced YOLOv5 Object Detection Algorithm for Accurate Detection of Adult Rhynchophorus ferrugineus." Insects 14.8 (2023): 698.

[20] Degadwala, Sheshang, et al. "Intelligent Crowd Counting System with Gender Classification." (2020).

[21] Yoo, Jang-Hee, So-Hee Park, and Y. J. Lee. "Real-Time Age and Gender Estimation from Face Images." Proceedings of the 1st International Conference on Machine Learning and Data Engineering (iCMLDE2017), Sydney, Australia. 2017.

[22] Adene, Gift, et al. "Detection and Classification of Human Gender into Binary (Male and Female) Using Convolutional Neural Network (CNN) Model." Asian Journal of Research in Computer Science 17.6 (2024): 135-144.

AUTHORS

Jalil Akbarzai – completed BSc Engineering in Electrical from the University of Engineering and Technology Peshawar, Pakistan. Currently working as a Research intern at Artificial Intelligence in Healthcare Lab, NCAI. His research focuses on computer vision, Artificial intelligence in Electrical systems, Power Electronics.

Muhammad Qasim – completed BSc Engineering in Electrical from the University of Engineering and Technology Peshawar, Pakistan. Currently working as a Research intern at Artificial Intelligence in Healthcare Lab, NCAI. His research areas encompass computer vision, the integration of artificial intelligence with electrical systems, and power electronics.

Zainab – completed BSc Engineering in Electrical from the University of Engineering and Technology Peshawar, Pakistan. Currently working as a Research intern at Artificial Intelligence in Healthcare Lab, NCAI. Her research interests include artificial intelligence, optimizing communication systems and computer vision technology.

Sayed Shahid Hussain – holds a Master's degree in Computer Systems Engineering from the University of Engineering and Technology Peshawar, Pakistan. He is currently a Research Associate at the AI in Healthcare Lab, National Center of Artificial Intelligence, UET Peshawar. His research interests include AI, machine learning, and deep learning, with significant contributions to data analysis, signal processing, image processing, NLP, and AI model development for practical applications.

Shahzad Anwar –received M.Sc., MS, in Electonics Engg, and PhD. degree from the University of the West of England, Bristol (Frenchay Campus), U.K. He is currently serving as an Associate Professor of Mechatronics Engineering, University of Engineering & Technology at Peshawar, Pakistan. He is also leading AI in Healthcare lab which is part of the National Centre of AI. His work focuses on computer vision and artificial intelligence with particular attention into innovative intelligent system techniques.