Intelligent Face and Person Classification Using Deep Learning Approach for Surveillance

Rajasekaran Thangaraj[1], P. Jaisankar[2], Siva Prasath P.[1], Suresh A.[1] and Surya Prakash V.[1]
[1]*Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamil Nadu, India*
[2]*Department of Mathematics, Nandha Engineering College, Erode, Tamil Nadu, India*

Keywords: People Detection Age, Gender, People Monitoring On-Board Surveillance.

Abstract: Person and face classification play a crucial role in various applications, including security surveillance, demographic analysis, and personalized services. This study proposes a real-time analysis and classification system using YOLOv11 and deep learning techniques. The model utilizes advanced object detection capabilities to efficiently detect human faces in video streams, followed by a Convolutional Neural Network (CNN)-based classifier for precise age and gender prediction. The system is designed to process live video feeds with high accuracy and minimal latency, ensuring reliable classification in dynamic environments. The integration of deep learning allows for feature extraction, improving classification performance across diverse detection among face and person. The proposed approach is evaluated on benchmark datasets to assess its effectiveness in real-world scenarios.

1 INTRODUCTION

More recently, machine learning approaches have attained considerable success in the computer vision applications, such as face recognition (Tuk, M., & Pentland, A., 1991). Face recognition has two major applications which are face identification and face verification. In face identification, the facial images of a person can be used for his or her identification, while for face verification, a face image and identity estimation are provided to the system to check the belonging of the face image to a certain person (Parki, et al, 2015). Detection accuracy is one of the biggest issues related to face detection. The face scale differs widely for the same image detector. (Taigman, Y.,2014) Face and person detection is one of the most basic tasks in computer vision with numerous applications including security, surveillance, biometric authentication, and human-computer interaction (He, at el, 2016). Traditional object detection models, such as R-CNN, often struggle with real-time performance due to their complex, multi-stage processing. However, advancements in deep learning have led to more efficient and accurate models like the YOLO (You Only Look Once) series, which are designed for high-speed object detection YOLOv11, the latest iteration in this family, introduces significant improvements in speed, accuracy, and efficiency (Levi, G., & Hassner, T. 2015).With a more optimized architecture incorporating C3K2 blocks for enhanced feature extraction, SPFF (Spatial Pyramid Pooling Fast) for better multi-scale detection, and C2PSA (Cross Stage Partial with Spatial Attention) for improved focus on critical areas, It excels in detecting faces and persons, even in complex environments (Zhang, C., at el, 2015)

These advancements make it ideal for applications requiring real-time detection with minimal computational overhead (Xie, W., et al, 2018). This implementation focuses on leveraging function for accurate face and person detection using deep learning techniques (Rodriguez, M, et al, 2011). By training the model on well-labeled datasets and fine-tuning it for improved performance, we can develop a robust detection system that works efficiently in real-world scenarios (Liu, W, et al, 2016). The following sections will cover the methodology, evaluation metrics, and implementation details, highlighting how YOLOv11 can be effectively used to detect faces and people in images and live video streams (Kang, et al, 2018). From a deep learning perspective, utilizes a fully convolutional neural network (CNN) backbone to extract hierarchical features from input images. The model processes an image in a single forward pass, predicting bounding

Thangaraj, R., Jaisankar, P., P., S. P., A., S. and V., S. P.
Intelligent Face and Person Classification Using Deep Learning Approach for Surveillance.
DOI: 10.5220/0013942000004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 684-690
ISBN: 978-989-758-777-1

boxes and class labels simultaneously Redmon, J., & Farhadi, A. (2018). Unlike traditional detection frameworks that generate multiple region proposals before classification, it directly predicts the bounding box coordinates, confidence scores, and class probabilities within a single unified framework (Howard, A. G., et al, 2017).This approach significantly reduces computational overhead, making it ideal for real-time detection tasks(Howard, A. G., et al, 2017).The following sections will explore the methodology, evaluation metrics, and implementation details, demonstrating how deep learning enables accurate and efficient face and person detection (Ribeiro, et al,2019).

2 RELATED WORKS

Deep Learning Based Age and Gender Detection Using Facial Images It focuses on classifying age and gender from static facial images using two distinct deep learning methodologies. The first approach involves training a Convolutional Neural Network (CNN) from scratch to extract and learn facial features pertinent to age and gender. The second approach employs transfer learning, where a pre-trained deep learning model is fine-tuned on a facial image dataset to enhance classification performance. The study aims to improve accuracy in applications such as surveillance systems, biometric authentication, and personalized services (Zhang, C., at el, 2015). S. S. R. Depuru and S. S. S. R. Depuru proposes a CNN-based approach for age and gender classification of real-world facial images. It uses a two-level CNN with preprocessing and is trained on IMDb-WIKI, MORPH-II, and OIU-Adience datasets. The model improves age classification by 16.6% (exact) and 3.2% (one-off) and gender classification by 3.0%. (Xie, W., et al, 2018) Crowd Detection and Analysis for Surveillance Videos Using Deep Learning" presents a CNN-based framework for detecting crowds, counting individuals, and analyzing age and gender for enhanced surveillance. Related studies explore crowd analysis aspects, including a CNN-BiLSTM model for abnormal behavior detection, a smart camera for real-time crowd monitoring, a YOLO V3-based system for misbehavior detection, and deep learning methods for distressful movement recognition and social distance monitoring. These advancements improve surveillance accuracy and efficiency using deep learning (Rodriguez, M, et al, 2011).

Gender of human images in the form of 2D body images is classified using deep neural networks according to the elucidation provided in (Ijiga, A. C.,at el, 2024). The end-to-enddesign prevents the inclusion of extra biometric features that are always useful in maintaining a high accurate gender classification system. It includes the pipeline for preprocessing the body images, as well as a 2D labeled data set. This just shows that gender determination from body images is as effective as from face images since the modified ResNet50 has higher accuracy.

Last, has proposed an Ensemble CNN for the real-time gender recognition. For the gender classification, the system achieves 95% of accuracy On IMDB dataset. The Real-time model incorporates actual-time processing with the effects offered. The average response time for both single face image input and multiple face image input for the quantity and the quality of the response is about 0.5 seconds. In total, these papers contribute to the improvement of the face-based classification in terms of time, quality and adaptability of the approach in other sets and real-world settings for the purpose of the insights and methodologies in further research and applications (Redmon, J., & Farhadi, A., 2018)

Towards this purpose, a new network, namely, gender CNN is devised for gender classification in where IMDb-WIKI is used also. The goal of this work is to extract facial features from real-world facial data; dropout, which randomly drops-off some data, and data augmentation to prevent overfitting are also part of this. Presumably, the optimized architecture allows to reach 84.8% of the correspondent value in the basic architecture. New methods allow to classify the age group beyond the demonstrated sample, as illustrated by the classification of the age group that constitutes only 84.8% of the population with 52.3% accuracy (Rodriguez, M, et al, 2011). The CNN2ELM network design has been proposed to achiev only with all the best advanced training techniques integrated in the network design.

Article: Video Analytics Using Deep Learning for Crowd Analysis: A Review, M.R. Bhuiyan et al. presents a complete overview of the current surveillance systems using deep-learning-based approaches for crowd detection, counting, and behavior analysis. It goes into details about various techniques including CNNs and regression-based models that have been used for crowd density prediction and crowd behavior anomalies detection. Additionally, the review attempts to tackle issues like occlusions and viewpoint distortions present in crowded scenes (Liu, W, et al, 2016). G. Sreenu et al, in their article “Intelligent Video Surveillance: A Review Through Deep Learning Techniques for

Crowd Analysis" explain the advances and applications of Deep Learning for intelligent videos surveillance with emphasis on object recognition, action recognition, crowd analysis and detecting violence in crowds. Different methodologies are discussed, such as Convolutional Neural Networks (CNNs) and regression-based models for density-based crowd counting and anomaly detection, but it fails to address how these models are implemented in operations. The challenge of occlusions and perspective distortions in crowded scenarios were also tackled. (Kang, et al, 2018).

Jocher et al.,2023 have concentrated on improving computational efficiency and accuracy. They adopted innovative methods such as the Extended Efficient Layer Aggregation Network (E-ELAN) and Dynamic Label Assignment to bolster the robustness of detection. The latest evolution in the series, YOLOv8, capitalizes on these improvements using C3K2 blocks, SPFF (Spatial Pyramid Pooling Fast) and C2PSA (Cross Stage Partial with Spatial Attention) to enhance detection performance, particularly for small and occluded objects (e.g. faces) – (Ijiga, A. C..et al, 2024)

Deep Neural Network Based Face Recognition with Support Based Anchor Boxes In this paper, we propose a Chi-Square-based face detection deep neural network using a modified support-based anchor boxes selection mechanism to improve detection performance.

3 MATERIALS AND METHODOLOGY

The YOLO (You Only Look Once) model family revolutionized the world of models, constantly challenging the limits of what was possible for a model to achieve with its realistic speed together with the exactness. For effective classification in monitoring systems. All of these significantly outperform others of YOLOv5, YOLOv7, and YOLOv8 in higher accuracy (55% MAP+) and faster processing speed (120+ FPS) in real-time. The model also employs improved feature extraction methods like SPP, FPN, and Transformer-based Convolutions, thus improving its accuracy of detecting faces across various circumstances. Different from its previous versions, the YOLOv11 proposes a Decoupled Detection Head with Dynamic Routing to enhance its classifications, especially in complicated environments such as low lighting and occlusions. Its architecture is also optimized for both edge devices and cloud platforms, which minimizes computational load while still being high-performing. For intelligent surveillance applications, these developments in YOLOv11 render it to be the best pick for such implementations all with a perfect balance of speed and accuracy compared to its predecessors.

YOLOV5: YOLOv5 is a PyTorch based object detection model which makes it more user-friendly and trainable when compared to the previous versions like YOLOv4. It is fast and precise and can run efficiently on both high-end GPUs and power-sensitive devices, offering various sizes (small to extra-large) for a balance of performance requirements.

YOLOV7:YOLOv7 is a state-of-the-art object detection model known for its high accuracy and efficiency, designed to operate quickly in real-time with low computational overhead. Also, it proposes Extended Efficient Layer Aggregation Networks (E-ELAN) to enhance feature learning and detection accuracy, which significantly speeds up and enhances detection performance compared to YOLOv5. YOLOv7 is a real-time object detection model that runs on very limited hardware.

YOLOV8:YOLOv8 is the latest and most sophisticated iteration of the YOLO object detection family, surpassing its predecessors in accuracy, speed, and flexibility. Building upon FRCNN, it also has a new efficient backbone, a fully decoupled detection head, and an anchor-free framework, making it suitable for real-time applications as well. YOLOv8 has gained significant popularity in various applications, including surveillance, autonomous systems, healthcare, and industrial automation, thanks to its impressive performance and user-friendly nature, making it a preferred choice when precision and computation efficiency are paramount.

3.1 YOLOV11

But the biggest question is: why YOLOV11? YOLOV10 Faster, More Accurate and More Efficient. Object detection, the task of detecting and locating objects within an image, is one of the most challenging problems in the area of computer vision. However, traditional algorithms for object detection, e.g., R-CNN framework, are often time-consuming, as they generate all responses to the image in advance and, later classify them, which is unwarranted in the case of real-time applications. These also carry the efficient architecture through efficient C3K2 blocks, SPFF (Spatial Pyramid Pooling Fast), and attention mechanism C2PSA the

goal of YOLOv11 is to enhance the accuracy of object detection and small object detection and maintain real-time inference speed. - Convolutional Block-This Block is referred in the field as Conv Block which lay down the given c,h,w through a 2D Convolutional layer net then through a 2D Batch Normalization layer. Bottle Neck-This is introduced as a series of convolutions in one block with a shortcut significance, which would define whether you want to gain the residual stuff or not. As in ResNet Block, if shortcut parameter was false, it would simply ignore the residual. The back bone of YOLOv11 is C3K2 block that is a strip down version of the CSP (Cross Stage Partial) bottleneck present in older versions. The C3K2 combination of characteristic to difference, to channel each feed forward pulls completely the information up and down through the net. This block is different from past models which failed to inspire balance number of parameter and feature representation, as those models took larger feature maps, run it through couple convolutions, then concatenate back those resulting feature maps. YOLOv11 still employs the SPFF module (Spatial Pyramid Pooling Fast), which allowed pooling features from parts of an image with the same sides. The C2PSA (Cross Stage Partial with Spatial Attention) The purpose of this block is to implement attention mechanisms, where the model tends to focus on the most relevant parts of an image, especially whenever the image has the presence of smaller and partially occluded objects; rather than only focusing on areas in the feature maps. PyTorch YOLOv11 demo implementation It really makes clear to you them where you should start testing object detection work on your images. Data Collection and Pre-processing It starts with collecting data and pre-processing for face and person detection using YOLOv11 A collection of positive and negative images of faces and non-faces are properly annotated for us to train. Pre-processing steps, including resizing, normalization, and data augmentation techniques (flipping, rotation, and brightness adjustment), are applied to these images to enhance the robustness of the model. Next, the YOLOv11 architecture is initialized, taking advantage of its advanced components such as C3K2 blocks which improve feature extraction by splitting and merging feature maps, and SPFF (Spatial Pyramid Pooling Fast) which helps bring improvements in multi-scale detection. Moreover, C2PSA (Cross Stage Partial with Spatial Attention) enables the model to focus on critical regions, thus, it performs well in recognizing small and occluded faces or persons.

The training data consists of well labelled datasets and focused model optimization, along with loss function to calculate losses during training, for example, Loss function, like Complete IoU(Intersection Over Union). It is the standard Metrix to quantify the degree of overlap between two boxes).) loss for accurate localization and binary cross entropy for classification. To enhance accuracy, particularly in recognizing faces and persons in different environments, techniques such as transfer learning and fine-tuning are utilized. On evaluation, metric like mean Average Precision (mAP) are utilized to measure detection accuracy and frames per second (FPS) to make certain on real-time validation after getting the model trained. The model is subsequently deployed, handling images or live video streams to accurately and quickly identify and locate faces and individuals, making it an optimal solution for surveillance, authentication, and crowd analysis applications.

4 EVALUATION METRICS

In order to use YOLOv11 in detecting faces and people, a number of evaluation metrics are considered for accuracy, analysis, detection, and real-time efficiency.

Mean Average Precision (mAP) mAP is a commonly used metric for object detection, which shows how well your object has localized and the accuracy of detection. It is calculated by interpolating the precision scores over a range of different confidence threshold. Higher mAP value indicates that the model did well in detection of faces and person and classification of their labels. MAP = mean average precision at multiple class and IoU thresholds. This is the most common metric used by other tasks of object detection, which can be a rough guide of how well your model does in the moment of dealing with precision and recall A higher mAP value indicates improved object localizat ion and classification, particularly for small and occluded objects. Domain summary of C2PSA blocks and C3K2 enhanced. Intersection Over Union (IoU): measures the overlap between predicted bounding box and ground truth box to be recognized as a true positive, a prediction must have an IoU above a threshold (generally between 0.5 - 0.95). CIoU (CompleteIoU overcomes these limitations by integrating more factors into the loss function. Figure 1 shows the Precision-Confidence Curve.

Flows Per Second (FPS): FPS is an indicator of speed, translating how many frames the model can

process per second. FPS (frames per second) While FPS itself does not directly translate to performance, a higher number generally leads to faster inference, which is especially important for real-time applications. Figure 2 shows the Precision-Recall Curve.

F1 Score: The harmonic mean between precision and recall is called the F1 score because it is a single value that takes into account both the metrics. Typically, these parameters help in avoiding the misclassification of data, especially when false negatives and false positives need to be the least in the output, ensuring effective detection. Figure 3 shows the F1-Confidence-Curve.

For this detection we can precise the analysis by deep learning with formal graph constrains as P-curve, PR-curve, F1-Curve and combine the labels correlations of the analysis. Figure 4 shows the Labels Correlation.

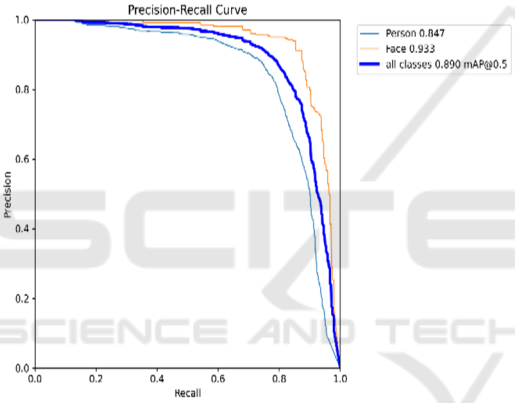

Figure 1: Precision-Confidence Curve.

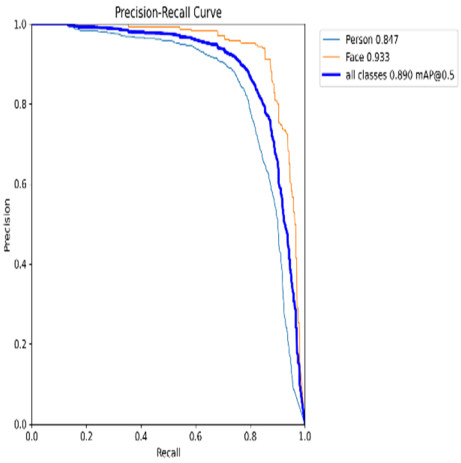

Figure 2: Precision-Recall Curve.

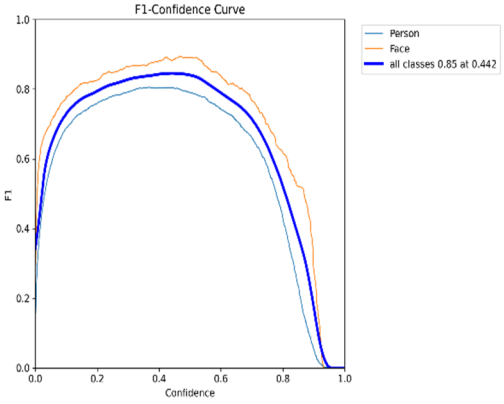

Figure 3: F1-Confidence-Curve.

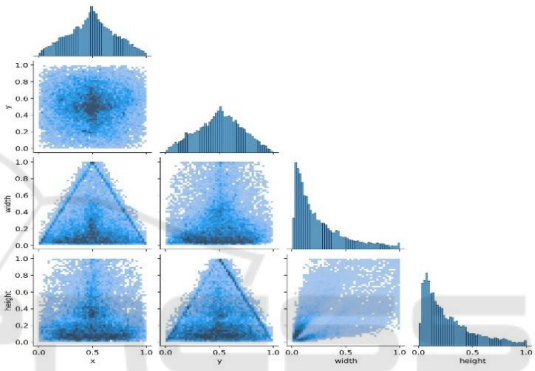

Figure 4: Labels Correlation.

5 IMPLEMENTATIONS

Face and person detection with YOLOv11 start with data preparation, which involves collecting a dataset containing labeled images of faces and people. The labels are processed by converting them into the YOLOv11 specific format, which includes the class index and pre-processed bounding box coordinates. After preprocessing the data, the YOLOv11 model is built up using the PyTorch framework. The third design for us is the definition of the network architecture which consists of C3K2 as feature extraction block and SPFF (Spatial Pyramid Pooling Fast) as multi-scale object detector followed by C2PSA (Cross Stage Partial with Spatial Attention) trying to pay more attention to the small or large objects. YOLOv11 backbone is being initialized, and pre-trained weights (if exist) can be loaded in to accelerate the convergence. It is then fine-tuned on a dataset for face and person detection. The split dataset is then divided into training/validation for training. AP is used for precise bounding box regression

(CIoU) and classification accuracy (binary crossentropy). To facilitate the convergence of training, a learning rate scheduling technique is employed. The model trains over several epochs, keeping track of its performance via validation loss and accuracy. The model is trained and then evaluated by calculating metrics such as MAP (Mean Average Precision) to evaluate accuracy, IoU (Intersection over Union) to measure the overlap of the predicted bounding boxes and FPS (Frames Per Second) to indicate if it is real-time or not. Additional metrics like precision, recall, and F1-score are computed to assess the model's performance in terms of accommodating for false positives versus false negatives. The model runs on each frame, applies non-maximum suppression (NMS) to remove redundant detections and provides bounding boxes around the detected faces and persons with confidence scores. For processing pipeline using OpenCV and PyTorch when the model is tuned and gives good results on the evaluation dataset. so that it could learn to detect faces and people in the given datasets. For NMS, the model works on each frame then all of them are passed through NMS which keeps only the boxes around faces and persons that we generated from our model. It depicts all class and combinations over the analysis and precise calculations (0.0 to 1.0) inside this range it draws together the calculations over the analysis and elements with an own output thing in it.

6 EXPERIMENTAL RESULTS

To assess the performance of YOLOv11 for face and person detection, extensive experiments were conducted using widely used benchmark datasets. During training, the loss function (CIoU Loss) consistently decreased over epochs, indicating that the model was learning effectively. The validation results showed a steady improvement in detection accuracy, demonstrating that the newly introduced C3K2 blocks, SPFF, and C2PSA mechanisms contributed to better feature extraction and object localization. Figure 5 shows the Person and face detection. Figure 6 shows the Person and face detection. Figure 7 shows the Precision graph. Table 1 shows the Accuracy and rate.

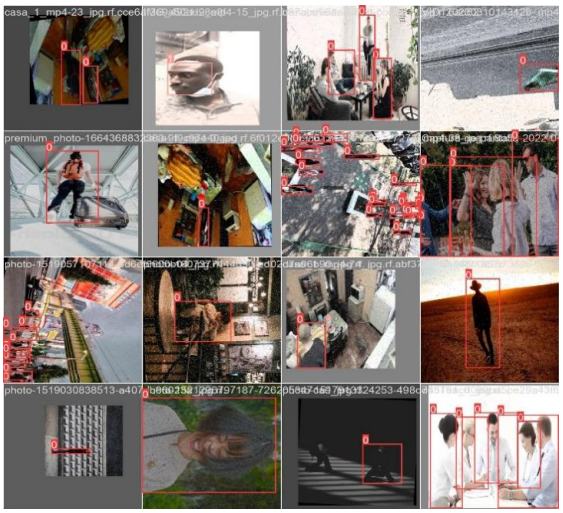

Figure 5: Person and Face Detection.

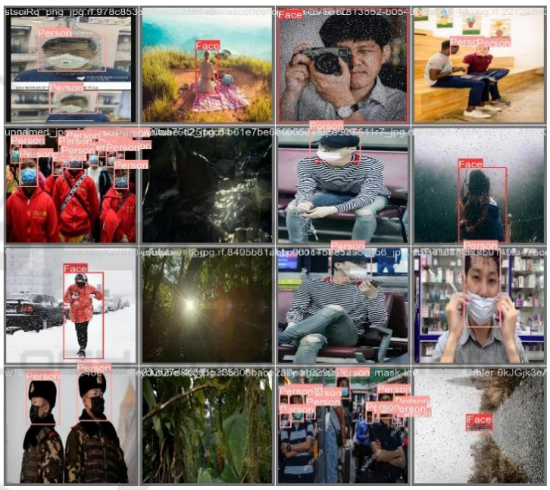

Figure 6: Person and Face Detection.

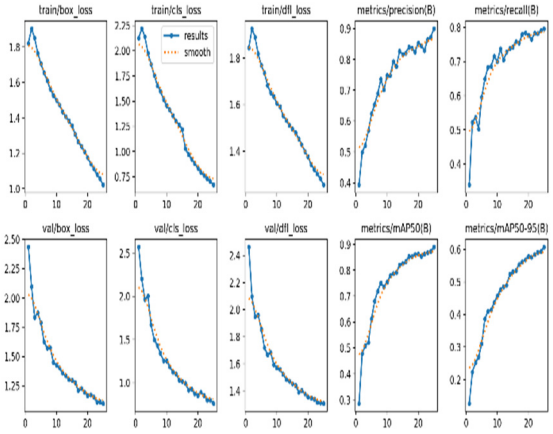

Figure 7: Precision Graph.

Table 1: Accuracy and Rate.

YOLO VERSIONS	MAP	Precision	Recal	F1-Score
YOLOV5	0.91	0.85	0.78	0.82
YOLOV7	0.94	0.87	0.80	0.85
YOLOV8	0.97	0.90	0.82	0.85
YOLOV11	1.0	0.91	0.93	0.95

REFERENCES

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Ijiga, A. C., Igbede, M. A., Ukaegbu, C., Olatunde, T. I., Olajide, F. I. & Enyejo, L. A. (2024). Precision healthcare analytics: Integrating ML for automated image interpretation, disease detection, and prognosis prediction. World Journal of Biology Pharmacy and HealthSciences,2024https://wjbphs.com/sites/default/files/WJBPHS-2024-0214.pdf

Ijiga, A. C., Abutu, E. P., Idoko, P. I., Agbo, D. O., Harry, K. D., Ezebuka, C. I., & Umama, E. E. (2024). Ethical considerations in implementing generative AI for healthcare supply chain optimization: A cross-country analysis across India, the United Kingdom, and the United States of America. International Journal of Biological and Pharmaceutical Sciences Archive, 2024, 07(01), 048063.https://ijbpsa.com/sites/default/files/IJBPSA-2024

Ijiga, A. C., Aboi, E. J., Idoko, P. I., Enyejo, L. A., & Odeyemi, M. O. (2024). Collaborative innovations in Artificial Intelligence (AI): Partnering with leading U.S. tech firms to combat human trafficking. Global Journal of Engineering and Technology Advances, 2024,18(03),106123.https://gjeta.com/sites/default/files/GJETA2024

Kang, K., Ouyang, W., Li, H., & Wang, X. (2018). Crowd counting with deep structured scale integration networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Lee, J., & Kim, S. (2023). A deep learning approach for IoT security: An application of CNN and LSTM. Journal of Information Security and Applications, 72, 103-112.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Li, Y. & Zhao, Y. 2020, 'Application of CNN in IoT Anomaly Detection', Journal of Network and Computer Applications, vol. 148, p. 102435

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. European Conference on Computer Vision (ECCV).

Parki, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face. British Machine Vision Conference (BMVC).

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprintrXiv:1804.02767.

Ribeiro, M., Ghosh, S., & dos Santos, E. (2019). Anomaly detection in crowded scenes using deep learning. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT).

Rodriguez, M., Laptev, I., Sivic, J., & Audibert, J. Y. (2011). Density-aware person detection and tracking in crowds. IEEE International Conference on Computer Vision (ICCV).

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tuk, M., & Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience.

Xie, W., Noble, J. A., & Zisserman, A. (2018). Microscopy cell counting with deep learning: CNNs versus transformer-based architectures. International Journal of Computer Vision.

Zhang, C., Li, H., Wang, X., & Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).