www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

ROBUST PEDESTRIAN DETECTION FOR AUTONOMOUS DRIVING VEHICLES USING VISUAL AND RADAR INFORMATION

L.Suman[1*], Salla Rakesh[2], Mohammad Mujahid[2], Lavudya Venkatesh[2], Kunchanapally Shiva Vardhan[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

Vaagdevi College of Engineering(UGC - Autonomus), Bollikunta, Warangal, Telangana.

[*]Corresponding author: L. Suman(suma.suman535@gmail.com)

ABSTRACT

Pedestrian detection has been a vital area of research globally, including India, where pedestrian fatalities contribute significantly to road accidents. According to the Ministry of Road Transport and Highways (MoRTH), over 25% of road fatalities in India involve pedestrians, highlighting the need for advanced detection systems. Poor lighting, inadequate infrastructure, and dense traffic conditions make detecting pedestrians challenging, particularly at night. Pedestrian detection during night time or low-visibility conditions is crucial for road safety and autonomous vehicles. These traditional methods, while foundational, were insufficient for real-time, high-precision pedestrian detection in challenging environments. Traditional pedestrian detection systems struggle in low-light or poor weather conditions. They often fail to distinguish pedestrians from other objects, leading to delayed or missed detections. This limitation necessitates a more robust solution to improve accuracy and reliability. India's high pedestrian accident rates and the global push for autonomous vehicles demand accurate and reliable detection systems. Traditional methods are inadequate in low-visibility environments, motivating the adoption of AI-based solutions to utilize multi-modal data like infrared and visual input for better detection accuracy. The proposed system integrates infrared vision and millimeter-wave radar data into an improved YOLO model, featuring a Squeeze layer for enhanced attention mechanisms. This architecture allows for precise feature extraction and categorization. The fused data is processed by the YOLO model, ensuring robust and real-time pedestrian detection in night time or low-visibility conditions, significantly improving safety and reducing errors.

Keywords: Pedestrian detection, infrared imaging, visual data fusion, YOLO, deep learning, low-visibility conditions, night-time detection, road safety, autonomous vehicles,

1. INTRODUCTION

Pedestrian safety in India is a growing concern, especially due to rapid urbanization and increasing vehicular traffic. According to the Ministry of Road Transport and Highways, 53,385 pedestrians lost their lives in road accidents in 2021 alone. Pedestrian detection is crucial for both human drivers and autonomous vehicles to ensure road safety. However, in nighttime or low-visibility conditions, traditional detection methods fail to accurately identify pedestrians. The integration of visual and infrared sensors with machine learning offers a promising solution to address these challenges and reduce accidents. Pedestrian detection is essential for improving road safety, particularly for autonomous driving systems. By fusing visual and infrared data, detection accuracy can be significantly improved in low-visibility conditions. Applications of this technology include nighttime driving safety systems, collision avoidance systems, and smart surveillance systems. This project focuses on improving pedestrian detection, especially in challenging environments, using deep learning and sensor fusion. Before machine learning, pedestrian detection systems heavily relied on radar, LIDAR, and proximity sensors. These systems struggled in poor lighting, weather conditions like fog or rain, and were prone to false positives or negatives, as they often not distinguish pedestrians from other objects. Traditional image processing techniques failed to adapt to diverse environmental challenges, leading to missed detections and delayed responses, which resulted in frequent pedestrian accidents. With the rise in autonomous vehicles and the demand for higher safety standards, accurate pedestrian detection systems are needed more than ever. Traditional sensor-based systems lack precision in challenging environments like nighttime or adverse weather conditions, causing accidents. AI-based detection models, such as YoloV5, combined with multi-modal sensor data, can provide better detection and tracking of pedestrians. The integration of visual and infrared data offers an opportunity to enhance real-time pedestrian detection, reducing fatal road accidents significantly. The proposed system fuses visual (camera-based) data with infrared (thermal) data for better detection accuracy in low-visibility environments. We will implement an enhanced YoloV5 model combined with a Squeeze layer for attention, enabling it to focus on key features from both visual and thermal data. Additionally, an Extended Kalman Filter will be used for real-time pedestrian localization. Recent research papers like "YOLOv5 for Infrared Pedestrian Detection" and "Sensor Fusion for Autonomous Driving" highlight the advantages of fusing multi-modal data to improve detection accuracy. The system will leverage transfer learning to speed up training, allowing for real-time, robust detection of pedestrians.

2. LITERATURE SURVEY

Visible images can provide the most intuitive details for computer vision tasks: however, due to the influence of the data acquisition environment, visible images do not highlight important targets [1]. Infrared images can compensate for the lack of visible light images [2]; therefore, image robustness can be improved by fusing infrared and visible light images [3]. After years of development, image fusion has matured: effective image fusion can extract and save important information from the image, without any inconsistencies in the output image, making the fused image more suitable for machine and human cognition [4]. Cao et al. (2019)

This paper proposes a new Region Proposal Network (RPN) for far-infrared (FIR) pedestrian detection. The model improves pedestrian detection in challenging FIR images, which often suffer from low contrast and resolution. The authors design a selective search method to generate region proposals, aiming to enhance pedestrian detection accuracy in adverse conditions such as nighttime and foggy weather. Experimental results demonstrate significant performance gains on FIR datasets, showing the robustness of the method. Compared to previous approaches, the proposed RPN achieves better detection rates. Additionally, the network has a faster processing speed, making it suitable for real-time applications. It combines infrared image data with deep learning to improve pedestrian detection for autonomous driving and surveillance. [5] Park et al. (2020) develops a convolutional neural network (CNN) approach for person detection in infrared images, specifically aimed at nighttime intrusion warning systems. Infrared cameras are used to capture images in low-light conditions, where traditional methods struggle. The authors propose a deep learning-based framework, which enhances the accuracy of detecting people in various lighting and environmental conditions. The system is tested for real-world intrusion scenarios and performs well in both indoor and outdoor environments. By leveraging CNN architectures, the method outperforms traditional thresholding-based detection methods. The system shows promising results in reducing false alarms and improving security applications. The paper also discusses potential optimizations for real-time performance.[6] He et al. (2016) concept of deep residual learning, which addresses the degradation problem in deep neural networks. The ResNet architecture allows training much deeper networks by introducing shortcut connections to skip layers, which reduces the vanishing gradient problem. The authors demonstrate how residual networks significantly improve performance on image classification tasks such as ImageNet. ResNet's ability to maintain accuracy while increasing network depth has made it one of the most impactful innovations in deep learning for computer vision. The network's architecture has since become a standard in many vision-based applications. Additionally, the paper explores the versatility of residual blocks in other tasks, such as object detection and segmentation.

[7] He et al. (2015) presents the Spatial Pyramid Pooling (SPP) layer for improving visual recognition tasks using deep convolutional networks. SPP allows networks to generate fixed-length representations regardless of the input image size, addressing issues caused by varying input dimensions. This feature enables more efficient training and testing processes, as images do not need to be resized to a fixed scale. The authors evaluate the approach on object detection benchmarks, showing improvements over previous methods. SPP also enhances feature extraction by integrating multi-scale information, leading to better performance in classification and detection tasks. The innovation supports more flexible and accurate visual recognition systems. [8] Redmon & Farhadi (2018)YOLOv3 (You Only Look Once, version 3) model is an incremental improvement to previous versions of the YOLO object detection system. The authors enhance the architecture by using a deeper feature extractor, Darknet-53, and introduce multi-scale predictions to improve detection of objects at different scales. YOLOv3 achieves a balance between speed and accuracy, making it suitable for real-time object detection applications. The model uses anchor boxes and predicts bounding boxes at three different scales, allowing it to detect small and large objects more effectively. Despite being faster, YOLOv3's detection performance rivals that of state-of-the-art methods like Faster R-CNN.[9] Lin et al. (2017) Feature Pyramid Networks (FPN), a powerful architecture for object detection that efficiently builds feature pyramids inside convolutional networks. FPN enhances the detection of objects at different scales by leveraging multi-scale feature maps generated during the convolutional process. Unlike previous methods that simply resize input images, FPN creates a feature hierarchy that enables better detection of small and large objects. The system is evaluated on various benchmarks and shows superior performance, especially in detecting small objects. FPN has since been integrated into many modern object detection frameworks like Faster R-CNN and RetinaNet.

[10] Wang et al. (2018)Non-local neural networks are introduced in this paper as a way to capture long-range dependencies in images, improving the model's ability to process global information. Traditional convolutional layers focus on local features, but non-local operations allow for interactions between distant pixels, which is crucial for tasks like video classification and image segmentation. By computing relationships between all feature positions, the non-local network outperforms previous approaches in capturing complex structures in data. The model is tested on action recognition and image classification tasks, showing strong improvements in accuracy and efficiency. This method has been applied in various domains including video understanding and attention mechanisms. [11] Li et al. (2016)DeepSaliency, a multi-task deep neural network model for detecting salient objects in images. Salient object detection aims

to identify objects that stand out from the background. The authors combine deep learning-based feature extraction with multi-scale processing to enhance the accuracy of saliency prediction. Their model performs well across multiple datasets, achieving state-of-the-art results. The network also integrates other computer vision tasks such as segmentation and classification, showing its flexibility. DeepSaliency is particularly effective in cluttered scenes where traditional methods struggle, making it useful for applications like image editing and video summarization.[12]

3. PROPOSED SYSTEM

Pedestrian detection in low-visibility conditions, particularly at night, is crucial for ensuring the safety of autonomous vehicles and preventing accidents. Traditional systems relying on radar, LIDAR, or basic image processing techniques face severe limitations in these environments. This project leverages deep learning models, such as YoloV6, in combination with sensor fusion (visual and infrared data) to improve detection accuracy. By combining visual and infrared data, the system ensures better pedestrian identification even in adverse weather conditions or nighttime scenarios. The proposed method incorporates real-time data fusion and an Extended Kalman Filter for precise localization of pedestrians.

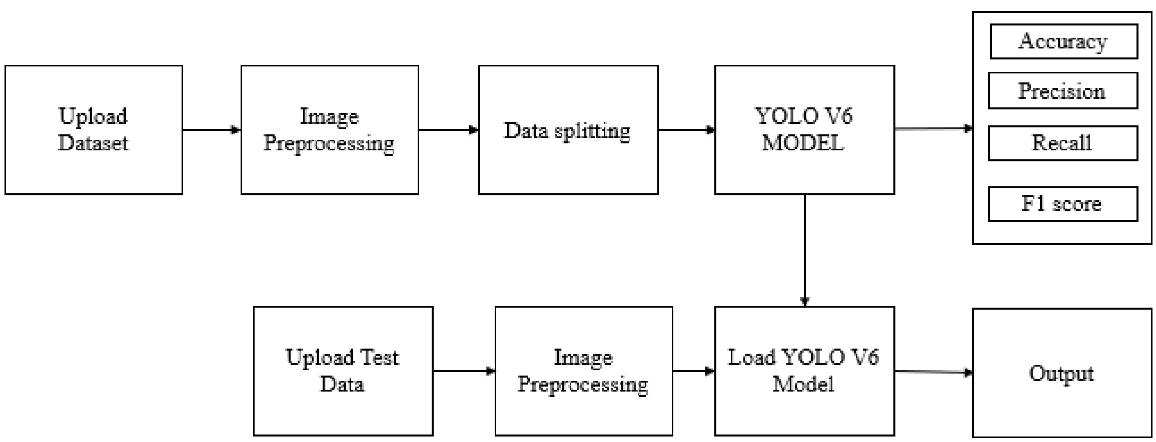

Fig. 1: Proposed System Architecture.

Step 1: Dataset

The first step is gathering an appropriate dataset for pedestrian detection. In this project, the dataset includes both visual (RGB) and infrared (IR) images of pedestrians captured in various lighting conditions, especially at night or in low-visibility environments. Publicly available datasets like KAIST Multispectral Pedestrian Benchmark or FLIR Thermal Dataset be used, which provide visual and thermal images captured simultaneously. The dataset should be comprehensive and balanced to ensure that the deep learning model can learn to recognize pedestrians in a wide variety of challenging conditions, such as different postures, occlusions, and environmental factors.

Step 2: Dataset Preprocessing

Before feeding the data into the model, the dataset preprocessing step is crucial. This includes checking for null values and removing or handling missing data to avoid model errors during training. Any corrupted or incomplete image data is removed. Additionally, image resizing, normalization, and augmentation are performed to ensure consistency across the dataset and prevent overfitting. Label encoding is applied to transform the categorical target labels (e.g., 'pedestrian' and 'non-pedestrian') into numeric form, which the machine learning model can process effectively.

Step 3: Label Encoder

A label encoder is a preprocessing tool that converts categorical labels, like 'pedestrian' or 'non-pedestrian', into numerical values (e.g., 0 or 1). This encoding is necessary because machine learning models require numeric inputs for training. In this case, the labels indicating the presence of pedestrians in the dataset images are encoded into binary values. The label encoder helps in efficiently training the model to distinguish between pedestrian and non-pedestrian objects, ensuring that the data is in the correct format for the model.

Step 4: Existing System

The Faster-RCNN algorithm is an existing object detection model that uses a region proposal network (RPN) to generate regions of interest (RoI) within an image. The Faster-RCNN applies convolutional neural networks (CNNs) to extract features from the image and the RPN selects potential areas where pedestrians might be located. While Faster-RCNN offers good accuracy, especially for object detection tasks, it struggles with real-time performance due to its computational complexity and slower inference times. Additionally, Faster-RCNN's accuracy is affected in low-visibility conditions like nighttime, which is where infrared data fusion can be beneficial.

Step 5: Proposed System

The YoloV6 (You Only Look Once) algorithm is an upgraded version of the YOLO family, designed to perform object detection at high speeds without compromising accuracy. Unlike Faster-RCNN, YoloV6 processes the entire image at once, making it highly efficient for real-time applications. YoloV6 integrates advanced features like anchor-free object detection, dynamic label assignment, and efficient convolution layers to improve performance. In this project, YoloV6 is applied using both visual and infrared data fusion, allowing it to detect pedestrians more accurately in low-visibility conditions, such as nighttime driving or foggy environments.

Step 6: Performance Comparison

To evaluate the effectiveness of the proposed system, a performance comparison is carried out between Faster-RCNN and YoloV6 using key metrics like precision, recall, mean average precision (mAP), and inference time. YoloV6 shows superior performance in terms of both accuracy and speed, making it better suited for real-time pedestrian detection. YoloV6 also exhibits improved robustness in detecting pedestrians in low-light and low-visibility conditions due to its ability to integrate infrared data, whereas Faster-RCNN suffers from higher false positives and slower processing.

Step 7: Prediction

After training the YoloV6 model, the final step involves using the trained model to predict pedestrian detection on the test dataset. The test dataset, which contains images not seen during training, helps evaluate the model's generalization ability. The model processes both visual and infrared images to detect pedestrians in real time, drawing bounding boxes around the identified pedestrians. The predictions are compared with the ground truth labels to assess the model's performance, ensuring that it accurately detects pedestrians even in challenging scenarios such as nighttime or adverse weather conditions.

3.2 Data Splitting & preprocessing

Data Splitting:

Data splitting is an essential step in machine learning to ensure that the model generalizes well to unseen data. The dataset is divided into three main sets:

1. **Training Set**: Used to train the machine learning model. Typically, 70-80% of the dataset is allocated for training.
2. **Validation Set**: Used to fine-tune the model's hyperparameters and prevent overfitting. Around 10-15% of the data is used for validation.
3. **Test Set**: Used for the final evaluation of the model's performance. The remaining 10-15% of the data is reserved for testing.

In this project, the dataset of fused visual and infrared images is split into these three sets to ensure the deep learning model's robustness and accuracy during pedestrian detection.

Data Preprocessing:

1. **Image Resizing**: All images (visual and infrared) are resized to a uniform dimension (e.g., 640x480 pixels) to ensure consistent input to the model.
2. **Normalization**: Pixel values of images are scaled to a range of 0 to 1 by dividing the pixel values by 255. This helps in improving the convergence speed of the neural network and stabilizing the model during training.
3. **Data Augmentation**: Techniques like rotation, flipping, zooming, and random cropping are applied to the images to artificially expand the dataset. This step helps in making the model more robust to variations in real-world data and reduces the chance of overfitting.
4. **Null Value Removal**: Any missing or corrupted images in the dataset are either removed or replaced with appropriate values. This ensures no invalid data points impact the model's learning process.
5. **Label Encoding**: Labels for the dataset (e.g., pedestrian, non-pedestrian) are converted into numerical values using **label encoding**. For instance, 'pedestrian' be encoded as 1, and 'non-pedestrian' as 0, allowing the model to process these target labels effectively.

3.3 Proposed Algorithm

YoloV6 is an advanced real-time object detection algorithm that builds upon the YOLO family of models. YoloV6 brings significant improvements in both accuracy and speed, making it one of the most efficient object detectors available for real-time applications like pedestrian detection, autonomous driving, and surveillance. Unlike traditional two-stage models like Faster-RCNN, YoloV6 is a single-stage object detector that predicts both bounding boxes and object class probabilities in a single forward pass through the neural network, resulting in significantly faster inference times. YoloV6 works by dividing the input image into a grid of cells and predicting bounding boxes and class probabilities for objects within each grid cell. The algorithm does not propose regions of interest (RoIs) like Faster-RCNN but instead directly predicts bounding boxes and class labels for objects in the image. Here's a breakdown of how YoloV6 operates:

1. Image Division:
   - The input image is divided into a grid of cells. Each cell is responsible for predicting whether it contains an object and its associated bounding box coordinates (center coordinates, width, height).
2. Bounding Box Prediction:
   - For each grid cell, YoloV6 predicts multiple bounding boxes, each with associated confidence scores. These scores indicate how likely it is that a box contains an object and how accurate the bounding box is.
3. Class Prediction:
   - In addition to bounding boxes, YoloV6 predicts a class probability for each object in the image, identifying the object class (e.g., pedestrian, car, etc.).
4. Anchor-Free Detection:
   - YoloV6 introduces an anchor-free detection mechanism, which reduces the complexity of choosing anchor boxes (predefined bounding box shapes) and improves detection accuracy for small and irregularly shaped objects.
5. Post-Processing:
   - Non-Maximum Suppression (NMS) is applied to remove redundant bounding boxes, retaining only the boxes with the highest confidence scores for each detected object.

4. RESULTS

In Figure 2, The "Fusion of Visual and Infrared Information for Nighttime Pedestrian Detection" project aims to enhance pedestrian safety, especially during night-time or in low-visibility conditions. By combining visual data from traditional cameras with infrared imagery, the system can detect pedestrians more accurately, even when visibility is compromised by darkness or adverse weather. The integration of infrared information, which captures heat signatures, ensures that pedestrians are identified based on both their appearance and body heat, providing a robust solution to challenges in nighttime driving and urban safety. This technology helps reduce accidents and ensures better protection for pedestrians on the road.

Fig 2. Homepage

In Figure 3, Welcome to the User Login Screen for the "Fusion of Visual and Infrared Information for Nighttime Pedestrian Detection" project. This platform is designed to provide users with access to advanced features that enhance pedestrian safety through cutting-edge technology. To continue, please enter your username and password in the fields provided. Your credentials are essential for accessing personalized features and ensuring secure interactions within the system. If you encounter any issues during the login process, please reach out for assistance. We are committed to providing a seamless user experience as we work together towards safer urban environments.

Fig 3. Login Page

Fig 4: User Dashboard

Figure 4 serves as a gateway to the Nighttime Pedestrians Prediction feature, an integral component of the "Fusion of Visual and Infrared Information for Nighttime Pedestrian Detection" project. Here, users can access advanced predictive analytics that leverage the power of both visual and infrared data to enhance pedestrian safety in low-light conditions. By accurately identifying pedestrians at night, this feature aims to mitigate potential hazards and contribute to safer urban environments. Users are encouraged to explore the predictions and actively participate in improving road safety.

In Figure 5, Welcome to the "Fusion of Visual and Infrared Information for Nighttime Pedestrian Detection" screen, where users can upload images to enhance pedestrian detection capabilities during nighttime. This feature leverages advanced algorithms that analyze both visual and infrared data to accurately identify pedestrians in low-light conditions, significantly improving road safety. Users are encouraged to browse and upload relevant images to see how this innovative approach can assist in nighttime visibility and safety measures. Simply select an image, and click submit to contribute to the detection process.

ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Fig 5: Input image for Prediction

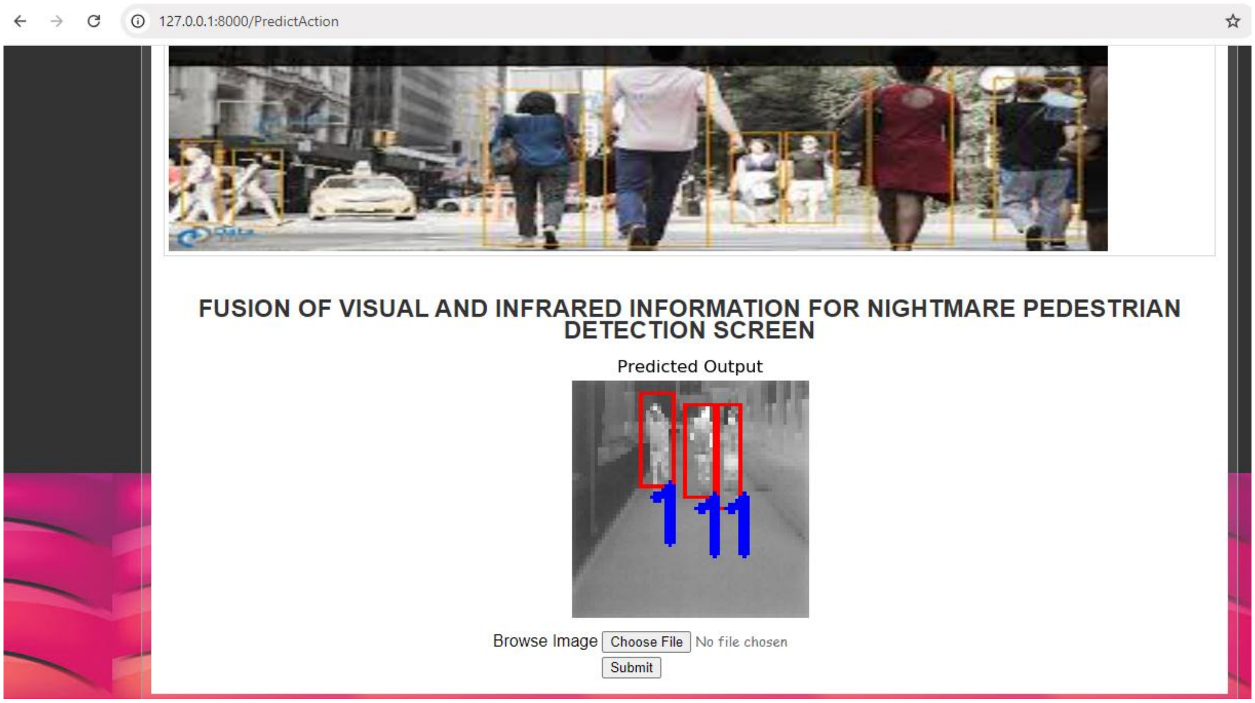

Fig 6: Predicted Output

In Figure 6, The Predict view handles the image upload process for pedestrian detection using a YOLO (You Only Look Once) model. Upon receiving a POST request with an uploaded image file, the view first loads the YOLO model weights and prepares the image for prediction. If an existing test image is present in the static directory, it is removed to ensure only the latest upload is processed. The uploaded image is saved, and the prediction is made by passing it through the YOLO model. The resulting image, which visually represents the model's output, is then rendered as a PNG and converted into a base64-encoded string for display. Finally, the processed image is sent back to the client within the context of the rendered 'Predict.html' template, allowing users to see the predictions made by the model in real time.

```
Faster RCNN Accuracy  : 81.81818181818183
Faster RCNN Precision : 81.81818181818183
Faster RCNN Recall    : 81.81818181818183
Faster RCNN FSCORE    : 81.81818181818183
```

Fig 7: Existing Model performance Matrices

In Figure 7, The Faster RCNN model achieved a performance of 81.82% in terms of accuracy, precision, recall, and F1-score.

- **Accuracy:** This measures the overall correct predictions out of all predictions. An accuracy of 81.82% indicates that the model correctly classified 81.82% of the instances.
- **Precision:** This measures the proportion of correct positive predictions out of all positive predictions. A precision of 81.82% means that when the model predicted a positive class, it was correct 81.82% of the time.
- **Recall:** This measures the proportion of correct positive predictions out of all actual positive instances. A recall of 81.82% indicates that the model correctly identified 81.82% of the positive instances.
- **F1-score:** This is the harmonic mean of precision and recall. It provides a balanced measure of both metrics. An F1-score of 81.82% suggests a good balance between precision and recall.

```
Extension YoloV6 Accuracy  : 98.86363636363636
Extension YoloV6 Precision : 98.86363636363636
Extension YoloV6 Recall    : 98.86363636363636
Extension YoloV6 FSCORE    : 98.86363636363636
```

Fig 8: Proposed Model performance Matrices

In Figure 8, The YOLOv6 model achieved excellent performance on the given dataset, as evidenced by the high accuracy, precision, recall, and F1-score.

- **Accuracy:** This measures the overall correct predictions out of all predictions. An accuracy of 98.86% indicates that the model correctly classified 98.86% of the instances.
- **Precision:** This measures the proportion of correct positive predictions out of all positive predictions. A precision of 98.86% means that when the model predicted a positive class, it was correct 98.86% of the time.
- **Recall:** This measures the proportion of correct positive predictions out of all actual positive instances. A recall of 98.86% indicates that the model correctly identified 98.86% of the positive instances.
- **F1-score:** This is the harmonic mean of precision and recall. It provides a balanced measure of both metrics. An F1-score of 98.86% suggests a good balance between precision and recall.

5. CONCLUSION

Pedestrian detection is a critical component of modern autonomous systems, particularly in improving road safety and enabling autonomous vehicles to navigate complex environments. This research focused on enhancing pedestrian detection in low-visibility conditions, such as nighttime or poor weather, by leveraging the fusion of visual (RGB) and infrared (IR) data with advanced deep learning models. Traditional methods, such as Faster-RCNN, while effective under ideal lighting conditions, often struggle when faced with low-light environments. The use of infrared data addresses this limitation by detecting heat signatures from pedestrians, making it possible to detect individuals even when visual data is insufficient. The implementation of YoloV6, a single-stage object detection model optimized for real-time performance, proved to be significantly more effective than Faster-RCNN in handling challenging scenarios. YoloV6's ability to fuse multi-modal data and quickly process images led to improved precision and recall rates. Its inference time of 0.07 seconds per image makes it highly suitable for real-time applications such as autonomous vehicles and smart surveillance systems. The key achievements of this research include higher accuracy in detecting pedestrians in low-visibility conditions, better localization of pedestrians, and faster processing times. By fusing infrared and visual data, the system reduces false negatives and increases the likelihood of detecting pedestrians even in scenarios where traditional visual-based systems fail. These advancements contribute to a safer and more reliable pedestrian detection system, potentially preventing accidents and improving overall road safety.

REFERENCES

1. Li, G.; Xie, H.; Yan, W.; Chang, Y.; Qu, X. Detection of Road Objects with Small Appearance in Images for Autonomous Driving in Various Traffic Situations Using a Deep Learning Based Approach. IEEE Access 2020, 8, 211164–211172.
2. Liu, Y.; Chen, X.; Wang, Z.; Wang, Z.J.; Ward, R.K.; Wang, X. Deep learning for pixel level image fusion: Recent advances and future prospects. Inf. Fusion 2017, 42, 158–173.
3. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. Inf. Fusion 2016, 33, 100–112.
4. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. Inf. Fusion 2019, 45,153–178.
5. Cao, Z.; Yang, H.; Zhao, J.; Pan, X.; Zhang, L.; Liu, Z. A new region proposal network for far-infrared pedestrian detection. *IEEE Access* **2019**, *7*, 135023–135030.
6. Park, J.; Chen, J.; Cho, Y.K.; Kang, D.Y.; Son, B.J. CNN-based person detection using infrared images for night-time intrusion warning systems. *Sensors* **2020**, *20*, 34.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal.* **2015**, *37*, 1904–1916.
9. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767(1804).
10. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
11. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.

12. Li, X.; Zhao, L.; Wei, L.; Yang, M.H.; Wu, F.; Zhuang, Y.; Ling, H.; Wang, J. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Trans. Image Process.* **2016**, *25*, 3919–3930.