

Convxf-Net: An Explainable Ai Approach For Recognizing Pedestrian Behaviour Using Convolutional Xgboost Pyramid Fractional Network

Dr. Amol Raoji Madane¹ Dr. Ho Chin Kuan²

¹Post-Doc fellow, Asia Pacific University of Technology, Jalan Teknologi 5, Taman Teknologi Malaysia, 57000¹amol.madane@gmail.com

²Vice Chancellor and Post-Doc Research Guide, Asia Pacific University of Technology, Jalan Teknologi 5, Taman Teknologi Malaysia, 57000² ckho@apu.edu.my

Abstract

In Advanced Driver Assistance System (ADAS), identifying pedestrians and predicting their future behaviour is essential for preventing accidents, however during heavy occlusion, the prevailing Pedestrian Behaviour Detection (PBD) techniques fails to identify the appropriate behaviour. Therefore, in this research, a novel Explainable AI based Convolutional XGBoost pyramid fractional Network (ConvXPF-Net) is proposed. The existing PBD techniques struggles to identify the patterns in the frames due to orientation-invariant regression head, and hence, YOLOv8 based Visual Casual Model (Y8-VCM) is employed, which eliminates the Boundary Discontinuity problem in Angle prediction. Besides, for addressing missed detections and issues with distracted walking during classification, a Convolutional XGBoost Pyramid Fractional Network (ConvXPF-Net) is utilized. The complexities in detecting the pedestrians with less radial velocities due to minimal Doppler shift, is reduced by Pyramid Net Technique (PNT), and then for recognizing the subtle behavioural changes that leads to oversimplification of the features, a Fractional Calculus Technique is employed. Further, based on the classified PB, the movement of pedestrians is determined, which further aids in performing speed and brake automation. Finally, the proposed work is validated using Explainable AI, and the outcomes attained shows the effectiveness of proposed work in Pedestrian behaviour prediction.

Key Words: Pedestrian Behaviour, Advanced Driver Assistance System, Deep Learning, Pedestrian Detection and Classification.

1. INTRODUCTION

The most vulnerable and complex road users known as Pedestrian's, are one of the fundamental reasons for traffic-based accidents and is essential to analyse the behaviour of the pedestrians, so as to enhance comfort and safety. By enhancing the road traffic safety systematically and comprehensively, it is possible to reduce the pedestrians-based traffic accidents. An effective and potential technology to address the road-based challenges is the Vehicle Automation and the launch of Advanced Driver Assistance System (ADAS) has given a steep rise in the level of automated vehicles, as specified by the Society of Automotive Engineers (SAE). In consideration with autonomous driving, many different topologies and techniques have been developed to analyse the next actions of the pedestrians. However, challenges exist with respect to the holistic views of the behaviour of pedestrians and the decision making. Moreover, not only exact prediction is necessary, in addition, explainability for the prediction is also necessary, which further helps in the validation and standardization processes carried by the autonomous vehicles [1-5]. For determining the factors which influence the behaviour of pedestrians, a force-based model has been proposed in [6] and it effectively measures the force acting over the pedestrians crossing with higher accuracy. However, the defined technique less explores about the repulsion and attraction generated by the surroundings. The traditional pedestrian-based decision-making systems, falls short in

meeting the demands of the individuals and the ADAS, which prompted the researchers to explore with the Machine Learning (ML) and Deep Learning (DL) models. When, an analysis is carried out to identify the effectiveness, with the combination of the multi-criteria decision analysis and ML using ranking and weighting techniques, it effectively addressed the constraints with the conventional techniques [7]. Besides, for detecting the pedestrians and to trigger the cross-walk signal automatically, a Convolutional Neural Network (CNN) is employed, which captures and analyses the images in real time and the outcomes attained automatically activates the systems with a crosswalk light. However, the defined work concentrates only on detection and lags in object classification [8].

Meanwhile, a Holistic Spatio-Temporal Graph Attention (HSTGA) technique is proposed in [9] for vehicle pedestrian interaction, which outperforms the state-of-art techniques with respect to curvilinear, linear and piece-wise linear trajectories corresponding to pedestrians and vehicles. Nevertheless, the defined technique less concentrates over the proximity amid the pedestrian and the vehicle, with respect to the climatic conditions. Then, for Vulnerable Pedestrian Prediction (VPP), a VPP network (VPPN) has been developed in [10] and is integrated with the Watchoutped dataset, which uses an auto-labelling method for ground region analysis and pedestrian state estimation. However, if only the area surrounding the pedestrian is considered, excluding the pedestrian, an 1% drop occurs in its efficiency. Besides, a Spiking Neural Network (SNN) along with Deep Neural Network has been proposed in [11] and it address the issues with street scene featuring and capturing the various illumination and elevation levels, but the sensitivity towards overlapping features leads to missed detection.

Besides, a Crowd activity perception framework has been introduced in [12] and it employs uses Scene Adaptive Congestion risk assessment based Multi-object tracking model (SAV-M). The SAV-M uses mutual mapping amid the geographical spaces and the image, facilitating effective identification of spatio-temporal detection. Even though, it is flexible and accurate, the reasonable assessment of risk zones, also the weight distribution necessitates an in-depth analysis. Meanwhile, an index called Extended Time to Collision (ETC) has been introduced in [13] for capturing the field image of pedestrians, as soon as they first approach the vehicle, in twilight condition. The ETC measures the time available for preventing the danger and the analyses results depicted that ETC at the time when pedestrians approach the vehicle is shorter in twilight than in the day-time. The defined technique effectively addresses the number of accidents, however, to improve the reduction rate of traffic accidents, pedestrian awareness of the vehicles approaching has to be promoted during twilight condition.

Moreover, an analysis has been performed in [14] with a naturalistic driving dataset using Hidden Markov Model (HMM), which considers the pedestrian movement, vehicle, driver gaze and instance of driving. The HMM identifies that the encounters do not involve reciprocal communication at all times and the patterns shift overtime. Nevertheless, the HMM less concentrates over the communication breakdowns occurring at a specific instance. Whereas, in [15], a novel variable is introduced for measuring the braking distance, which captures the deceleration process and reflects both the driver and pedestrian safety. Moreover, the study concentrates only with the leading vehicles and not over the other vehicles behind, also the defined technique necessitates improvements with the fast response. From the analysis, it is identified that the use of force-based model for pedestrian prediction possess issues with repulsion and attraction of the surrounding, while CNN aims only on detection and lags in object classification. Besides, the HSTGA technique, does not concentrate over the proximity of pedestrians

and vehicles based on climatic conditions, also the VPPN possess 1% drop with accuracy, due to surrounding consideration.

Similarly, the SNN based technique has higher sensitivity towards overlapping features and it leads to missed detection, while, the introduction of SAV-M requires reasonable assessment of risk zones, also the weight distribution necessitates an in-depth analysis. Meanwhile, the use of ETC technique, requires pedestrian awareness of the vehicles approaching, which has to be promoted during twilight condition. Furthermore, the use of HMM technique faces communication breakdown and the introduction of measuring variable based pedestrian detection faces constraints with the fast response. Hence, to mitigate these issues in pedestrian behaviour recognition, a novel Explainable AI based DL technique is proposed in this research work.

1.1 Contributions of the proposed work

To address the complexities with the conventional pedestrian behaviour detection techniques, a significant technique has been proposed in this research and the main contributions concerning are stated below.

- To eliminate the issues with the Boundary Discontinuity Problem in Angle Prediction, due to orientation-invariant novel YOLOv8 based Visual Casual Model (Y8-VCM) is proposed, which assists in making decisions with the orientation of the facing angles of pedestrians.
- To reduce the higher false negatives due to missed detections, a Convolutional XGBoost Pyramid Fractional Network (ConvXPF-Net) is proposed and it aids in detecting the behaviour of the pedestrians.
- To lessen the Error of Emission due to minimal Doppler shift, a Pyramid Net Technique (PNT) is employed with the ConvXPF-Net, which uses multi-scale extraction technique for reducing the occlusion.
- To address the superficial understanding due to subtle behavioural changes, a Fractional Calculus Technique is integrated to ConvXPF-Net and it employs its memory property to reduce the loss of context.
- To validate the significance of the proposed ConvXPF-Net based pedestrian behaviour prediction model, using Explainable AI (XAI), which gives the appropriate explanation about the defined work.

1.2 Organisation of the research work

In order to assist the pedestrian behaviour prediction, a literature review is performed in section 2 and based on the motivation attained from the review, the proposed research work introduced is discussed in section 3. Besides, the results attained from the proposed work and the performance metrics evaluated from the research work, also the comparison of proposed with conventional techniques are also discussed in section 4. Then, the defined work is concluded in section 5, along with its significance and the future suggestions.

2. LITERATURE SURVEY

For effectively analysing the behaviour of the pedestrians, a review is carried out concerning the conventional pedestrian detection techniques along with the effectiveness of the autonomous driving system in this section, along with its pros and cons.

Sahar Hassani et al. [16] introduced a pedestrian behaviour monitoring approach that employs advanced signal processing and approaches for enhancing the essential measurement features. For effective monitoring of the pedestrians, Vibrational

Decomposition algorithm along with data fusion was employed, which reduces the errors and noises in the input data. Then, a 2D CNN is employed independently for sensing the sources and was further extended to fuse raw, deionised and decomposed data. Further, the defined work was validated with the aid of experimental data attained from the pedestrian simulation carried out over a bridge panel. The defined technique improves the structural and pedestrian management by implementing an early warnings so as to predict safety risks and gives out better performance than other techniques. However, challenges occur with respect to prediction amid the pedestrians and the infrastructures, which is necessary for the safety management.

Chaeyeon Han et al. [17] presented a new strategy for scaling the urban pedestrians with the aid of Audio based topology and it uses a large-scale dataset known as ASPED that incorporates high quality audio recordings and audio recordings, in order to label the data regarding the pedestrian count. As audio sensing was combined with the prediction techniques, the accuracy and reliability of the defined technique enhances. The audio-based sensing gives better significance across different lighting conditions and was efficient than the video surveillance. Even though audio sensing was effective, the performance of it lags with the video sensing and this is due to that few audio features were trained only for audio event detection and is complex than the detection of pedestrians.

Mohammad BaniSalman et al. [18] made a study to address the limitations with the multiple sensor data and insufficient critical edge coverage of pedestrians using a Vulnerable Road Deep Safety (VRDeepS) simulation platform. It generates datasets that features synchronized multi-vehicle interaction and then using a Bayesian inference, the VRFormer uniquely fuses V2X with the VR sensor data. The defined technique gives out an Accident Prediction Accuracy of about 85% at 2second prediction horizon, which significantly improves the Autonomous Driving (AD) simulation, also the safety setting. From the approach, the limitations in the single vehicle systems were revealed and necessitates the need for the effective and robust AD system. However, the defined work does not realize the V2X potential completely and moreover further refinement is essential for maximizing the real-world performance. Sundas Iftikhar et al. [19] discussed about the design and implementation of the real pedestrian detection with the aid of Deep Learning Network, in order to identify the rejection in pedestrians. The main goal of this work was to compute the significance of 2D detection system that employs 3D CNN (YOLOv3) in identifying the true false pedestrians from the images attained from the cameras of the on-board, as well as the Light detection and Ranging Sensors. Further, an analysis was carried out single phase and two phase based meta-structures beneath different attribute extractors and image resolutions. For executing the framework effectively, data augmentation method has been used and it was identified that the defined work gives out better detection rate with respect to the false and true pedestrians, bit issues occur with respect to the run-time performance.

Eunmok Yang et al. [20] have introduced a Bio-inspired Garra Rufa Optimized DL (BGRO-DL) for classification of the pedestrians on the walkways. The defined technique initially analyses the existence of pedestrians and other objects in the surveillance video. To assist with better classification, a Ghosh Net that gives out a set of feature vectors has been employed. then, for tuning the hyper-parameters of the classifier, a GRO algorithm was used, from which the classification of pedestrians was performed using Attention based Long Short-term Memory (ALSTM). The analysis results depicts that the defined work performs better than the conventional techniques, however, in case of real time video streams, the defined work necessitates enhancements in the scalability, also the

applicability has to be extended in order to improve the safety of the pedestrians and to enhance the classification accuracy.

Nang Htet Aung et al. [21] presented a MetaPed, which is a synthetic pedestrian dataset generated from Metaverse, in which the avatars expose the behaviour of the pedestrians, excluding the original image of the individuals. Moreover, the Metaverse incorporates programmed pedestrian behaviour, as well as the regulated avatar movements that ensures the realistic behaviour of the pedestrians. To analyse and validate the effectiveness of the proposed work, it was evaluated with both intra and cross dataset. The results from it depicts that the defined work gives out better precision and recall performance in terms of pedestrian detection. The defined dataset not only enhances the detection, but also shows the potential of the model in complementing the real-world data. Nevertheless, challenges exist diverse pedestrian behaviours, which depicts the concerns with the real time data. Mehmet Kilicarslan et al. [22] introduced a Spatial-Temporal Walking Pedestrian Detection technique and it uses One-phase Motion based Pedestrian Detection, in order to prevent collision in driving. A DeepStep Network, which is based on CNN, instantly identifies the Leg trajectories of the walking pedestrians at the motion profile image patches. The non-smooth movements from long and short distance are effectively classified by the learned network. The defined work also has the ability to identify the pedestrians amid the crowd, also among different environments with less False -Positive rates. When certain motion profile images are applied to highly effective DL technique, significant real-time pedestrian prediction is attained, but fails in identifying off-road pedestrians occluded by other objects and other pedestrians.

Siva Senthil D and Sivarani T S [23] have made a discussion about detecting Suspicious Event Detection (SED) with the aid of SEDNet, in which the input images are collected from two datasets and is transformed into frames. It is then pre-processed using Gaussian Adaptive Bi-lateral Filter (GABF), which decreases the noises, while preserving the edges. In order to identify the finer edges in the processed image, a Sobel Edge Detector was employed and it enhances the structural details. Then, for extracting the features necessary, a Di Stream Neural Network was employed, which retrieves both the pose and the motion features. Further, for classifying the suspicious activity, the motion and pose features are combined using Weight Average Fusion Technique, which is then classified using Simplified Spiking Neural Network with better accuracy. Besides, optimizing SEDNet in real-time detection with lightweight architectures with hardware accelartion is complex.

Jie Xu et al. [24] have presented a Social Force based Method (SFM), which is characterised by loosely defined Leader and Follower structure. For measuring the variations in information flow amid the pedestrians, a field-based Phase Transfer Entropy was used. Then for detecting the leader in the collective, initially, the detecting threshold was set with three sigma principle and a Radial Basis Function along with SFM is utilized. The integration of these models aids in simulating the collective behaviour. In order to validate the model, the data from Fairground Dusseldorf was used. The walking trajectories reproduced reflects possess higher accuracy with less swing, which guides the optimization about the design of walking that aids in pedestrian prediction. Meanwhile, the defined model solely depends on the prolonging time interval, which is directly proportional to the stability of the structure. Deepak Kumar Jain et al. [25] introduced a robust Multi-Modal Pedestrian Detection (MMPD) with the aid of DCNN along with Ensemble Learning. For accurately predicting the Pedestrians, the Computer Vision and Ensemble Learning are leveraged. Further, for extracting the required

features, a Sim based Efficient Net Model was presented and it uses Stochastic Gradient Descent optimizer for the extraction. Besides, Ensemble classification includes three DL models namely Nested LSTM, Deep Belief Networks and Extreme Learning Machine. Furthermore, for identifying the significance of MMPD, it is analysed with INRIA dataset, which resulted in robust prediction, even with variations in the pose and appearance. Nevertheless, the defined model lags with the pedestrian detection with different climatic conditions, also necessitates effective optimization algorithms for fine-tuning the model. From the detailed review carried out with respect to pedestrian behaviour detection, it is identified that the use of advanced signal processing approach with 2D CNN in [16] possesses challenges with the prediction amid pedestrians and infrastructures, while the ASPED in [17] exhibits issues with training the audio features. Meanwhile, VRDeepS and 3D CNN (YOLOv3) of [18] and [19] has problems with maximizing the real-world performance and detection rate with respect to the false and true pedestrians. Besides, BGRO-DL and MedPed of [20] and [21] has constraints regarding the safety of pedestrians, classification accuracy and potential of the model. Moreover, the STW-PD and SED in [22] and [23] fails in identifying off-road pedestrians occluded by other objects and detection with lightweight architectures. Besides, the use of SFM in [24] gives out better accuracy, but solely relies on the prolonging time-interval, while the MMPD necessitates effective optimization algorithms for fine-tuning the model. Hence, for addressing these challenges, a novel DL based Pedestrian Behaviour Detection model is proposed in this research work.

2.1 Motivation of the Work

In consideration with the Pedestrian behaviour detection, the visual models employ Bounding boxes for identifying and localizing the objects within a video frame or an image and it mainly uses Regression Head for predicting the co-ordinates of the bounding box. As the conventional techniques solely rely on manual features like Histogram of Oriented Gradients (HOG), Deformable Part Model (DPM), etc. and possess issues with respect to scale invariance orientations. This is mainly due to orientation mis-alignment, in which the regression head is not orientation-invariant and struggles to identify the patterns present in the data, leading to inappropriate predictions. When a pedestrian is facing an angle (left, right, or diagonally), their body's orientation might not align perfectly with the grid's axes and the grid fails to capture the full extent of the bounding box. This leads to Boundary Discontinuity Problem in Angle Prediction, in which pedestrians react to close-time collisions by taking into account the shape of their body, which can be challenging to model and predict. Moreover, during heavy occlusion of pedestrians, the prevailing Pedestrian Detection Models (PDMs) fails to identify them, due to missed detections that leads to higher false negatives. Whenever, a model or system incorrectly specifies that an object is absent (negative), which is actually present, lags to reject the False Null Hypothesis (FNH). This is mainly due to the fact that, even if the echo of the target is robust than the clutter, the saturation masks the target signal, as it is complex to detect the pedestrians with less radial velocities due to minimal doppler shift. This makes them invisible to the radar display, which leads to Error of Omission, which equals to one minus the power of the test, leading to less false positives, which further degrades the performance of the model. As the level of occlusion increases, the Accuracy and Recall of the PDSs decreases and the model fails to recognize the pedestrians effectively.

Besides, the pedestrians frequently use mobile phones, headphones and other devices at the time of walking, which reduces the attention about surrounding leading to Erratic movements with distracted walking. Due to this, constraints occur in recognizing the subtle behavioural changes, which are often left unidentified. This is because, most of the

conventional models prioritizes only the observable behaviour and excludes the nuanced and subtle changes that precedes significant variations in the behaviour of the pedestrians. This leads to oversimplification of the features with loss of context that brings superficial understanding and hinders effective decision making.

3. PROPOSED CONVXPF-NET BASED EXPLAINABLE AI APPROACH FOR RECOGNIZING PEDESTRIAN BEHAVIOUR

For addressing the issues encountered by conventional techniques with respect to detection and classification of pedestrian behaviour, a novel "Explainable AI based Convolutional XGBoost pyramid fractional Network (ConvXPF-Net)" is proposed in this research work.

Initially, for extracting the video frames, the input video considered is subjected to video frame extraction, in which the specific frames necessary for video analysis are identified and selected. In order to address the issues with scale invariance orientations, a novel YOLOv8 based Visual Casual Model (Y8-VCM) that uses YOLOv8 for reducing the issues with capturing full extend of the boundary box is employed. At first, the YOLOv8 identifies the pedestrians in cross walks from the video frame, with respect to the lighting, crowd density, and weather, which gives out bounding boxes along with confident scores, which specifies the location of the pedestrians and then, from the bounding boxes, the features necessary are extracted. Then, the features extracted are fed to the Visual Casual Model (VCM) for reasoning about the relationship amid various features. This mainly aids in contextual understanding, which thereby gives out the likelihood of the pedestrians. Further, the output aids in making the decisions about the orientation (left, right or diagonal) of the pedestrians facing angle, which thereby eliminates the Boundary Discontinuity problem in Angle prediction.

Then, for reducing the higher false negatives due to missed detections and to rectify the issues with the distracted walking, a novel Convolutional XGBoost Pyramid Fractional Network (ConvXGB-Net) is proposed, which is highly effective in identifying the behaviour of the pedestrians. The ConvXGB-Net model includes Convolutional eXtreme Gradient Boosting (ConvXGB), which involves two hidden layers, which include Fractional Calculus and Pyramid Net. For dealing with the minimal Doppler shift, the Pyramid Net Technique (PNT) is employed and is a multi-scale extraction technique that identifies the occluded parts by capturing the features at various spatial resolutions. This reduces the occlusion and assists in decreasing the Error of Omission, which better aids in recognizing the pedestrians. Further, for considering the subtle behavioural changes of the pedestrians, Fractional Calculus Technique is used, which uses a classical calculus for dealing with the functions of integration and differentiation of the non-integer order. The defined technique, takes into account the correlation amid the pixel values and hence the fractional derivative assists in attaining the behaviour of the pedestrians, which are left unidentified. Besides, the efficient memory property of the fractional differentiation, includes the nuanced, as well as the subtle changes and reduces the issues with the loss of context that rectifies the superficial understanding, which makes improvement in decision making. Furthermore, based on the decision, the behaviour of the pedestrians (i.e.) Walking, Standing, Looking and Crossing at the traffic are classified at the output layer of the ConvXGB, using Convolutional Neural Network (CNN) and the prediction accuracy is improved by XGBoost with the aid of decision trees. Furthermore, based on the classified output, if a pedestrian is detected to be Crossing or Walking, then as a response to it, automatic break or slowing down is carried out, which brings better

pedestrian safety. Then, for validating the effectiveness of the proposed ConvXPF-Net based pedestrian behaviour prediction model, an Explainable AI (XAI) is employed, which analyses the features or inputs of the model, which uses the most for predicting the pedestrian behaviour.

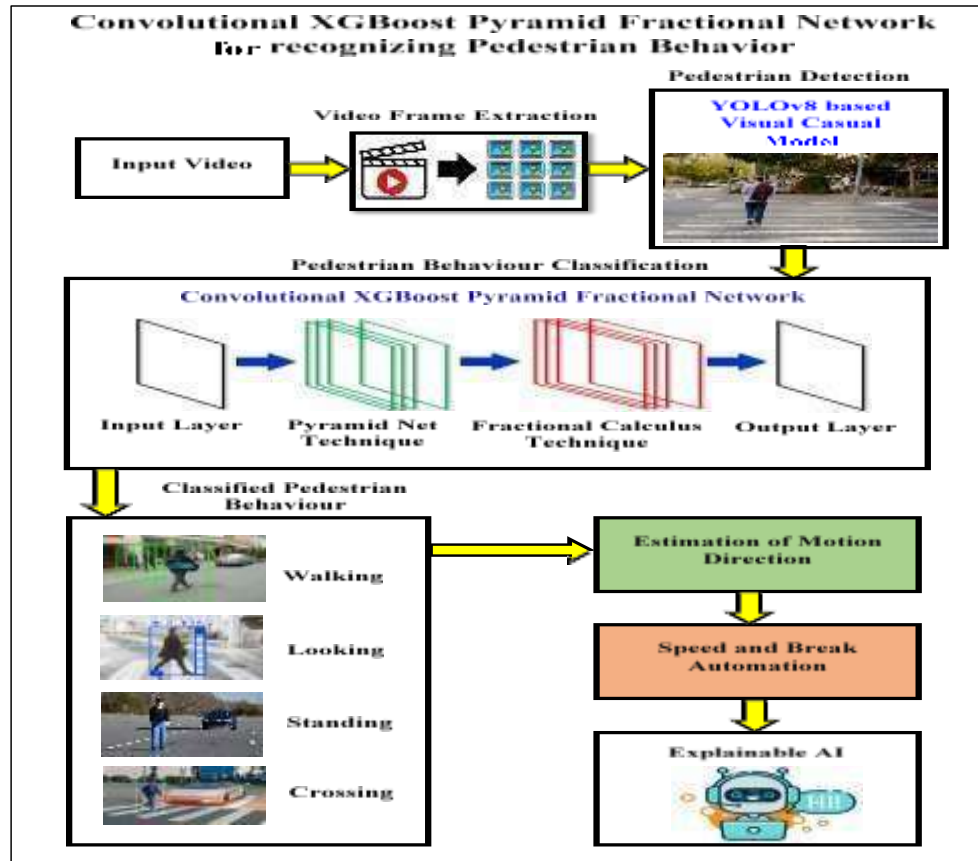


Figure 1: Block Diagram of the proposed ConvXPF-Net based Pedestrian Behaviour prediction

The Block Diagram of the proposed ConvXPF-Net based Pedestrian Behaviour prediction is illustrated in Figure 1. The functioning of the proposed ConvXPF-Net based Pedestrian Behaviour prediction starts by taking the input from the preferred dataset and the necessary video frames from it has been extracted. Then, with the aid of Y8-VCM, the cross walks of the pedestrians with respect to climatic conditions and lighting are identified. Further, the behaviour of the pedestrians is identified using ConvXGB-Net, in which PBT and FCT, along with ConvXGB involved, helps with effective classification and prediction of the pedestrian's behaviour. Further, for validating the significance of the model, the proposed model is analysed by XAI. In order to identify the effectiveness of the defined work, it is implemented in Python Tool.

3.1 Video Frame Extractor (VFE)

Initially, the input video taken from the JAAD (Joint Attention in Autonomous Driving) Dataset, is fed to the Video Frame Extractor [26], in which individual frames necessary for pedestrian behaviour detection has been extracted from the video file. The VFE includes the following process for extracting individual frames from the video.

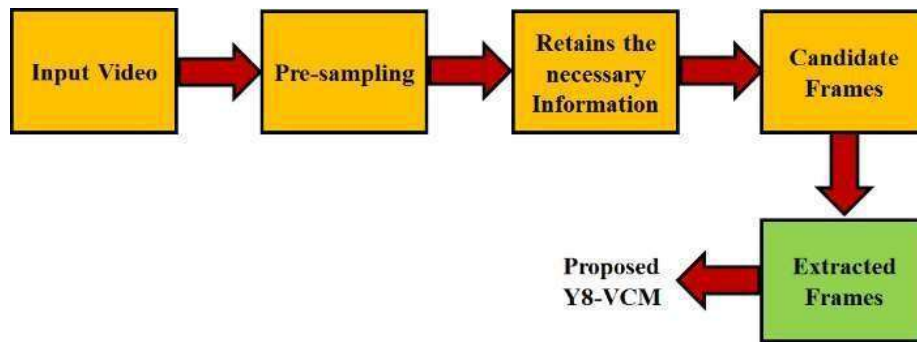


Figure 2: Video Frame Extraction

Uploading the Video: The input video considered is pre-processed and the noises in the video are removed, retaining the necessary informations and the candidate frames are given to the frame extractor. The diagrammatic representation of the proposed block diagram was depicted in Figure 2. The extractor in the VFE reads the video file and identifies its structure, which is further transformed using a decoder into a compressed video stream with an arrangement of frames that are necessary for the processing.

Frame Extraction: The process by which the video file is converted into a sequence of individual frames is known as Frame extraction. Many different tools have been employed for frame extraction, which permits the users for capturing individual frames from video at appropriate intervals. Then, for identifying scale invariance orientations, Y8-VCM model is employed.

3.2 Proposed YOLOv8 based Visual Casual Model (Y8-VCM)

Generally, the visual models employ regression head for predicting the co-ordinates of the bounding-box in the video frame. As the regression head struggles with identifying the patterns of the frames, in-appropriate prediction occurs. Hence, to bring effective prediction of the patterns, a novel YOLOv8 based Visual Casual Model is used. The VCM employs YOLOv8 for detecting the pedestrians and then uses Scene Graph Generation (SGG) for enhancing the YOLOv8. The steps involved in the proposed YOLOv8 based Visual Casual Model is depicted in Figure 3.

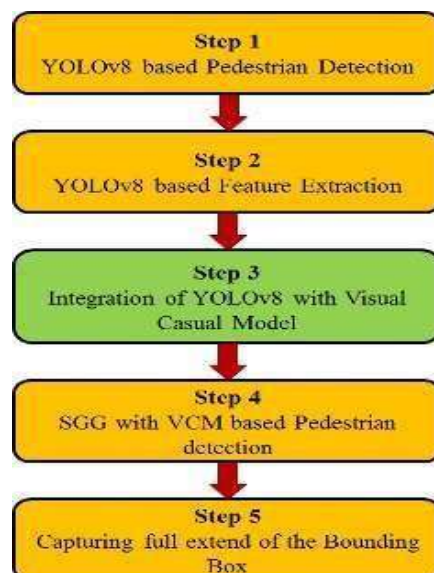


Figure 3: Proposed YOLOv8 based Visual Casual Model

3.2.1 YOLOv8 in Pedestrian Prediction

Initially, for detecting the pedestrians in the video frame, a YOLOv8 network is employed, which significantly specifies the bounding boxes over the pedestrians with class labels and confident scores. The YOLOv8 aids in object localization and classification tasks inside a single, End-to End differential network framework sustaining a balance amid the speed and the accuracy. The structure of YOLOv8 [27] possess components like Backbone, Neck and Head; the object classification stage involved in YOLOv8 is illustrated in Figure 8

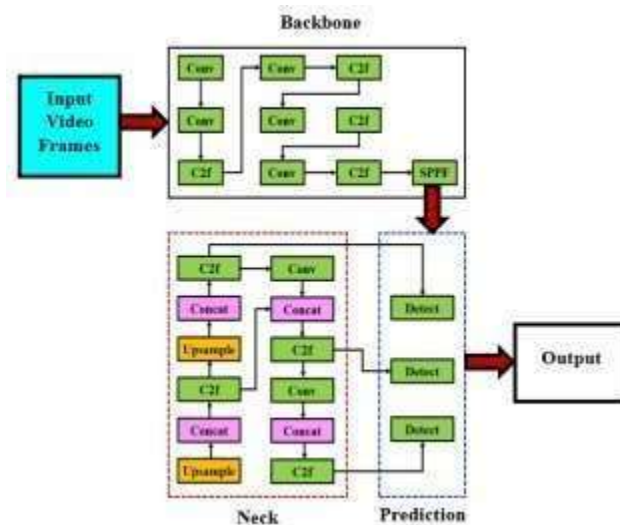


Figure 4: Object classification process using YOLOv8.

Backbone: The YOLOv8 employs CNN as the backbone, for extracting multi-scale features from the input and it captures the hierarchical feature maps significantly, which specifies the lower level and higher-level semantic information that is essential for object detection. The backbone mainly concentrates over the accuracy and speed, while with holding a separable convolution for reducing the computational overhead and for sustaining the power representation.

Neck: The features extracted from the backbone are refined and fused by the neck module and it leverages a path aggregation network, which is improved by enhancing the information flow across various feature levels.

Head: The final predictions from the refined images, are generated by the Head module and it incorporates the co-ordinates of the bounding-box, confidence score of the objects and the class labels. For accurate prediction, the YOLOv8 presents an anchor-free technique, which reduces the prediction process by decreasing the count of the hyperparameters, also enhance the adaptability of the model to the objects having different scales and aspects. Finally, the YOLOv8 predicts the bounding boxes, with effective reduction of the boundary discontinuity problems with angle prediction, also gives out the confidence scores that specifies the location of the pedestrians.

Detection of Pedestrians: A real time-based video frame is given as output to the YOLOv8 framework.

Extraction of Frames: From the detected bounding boxes, the features necessary for prediction are extracted and it includes movement, environmental and contextual features. With the use of movement features, the angle faced by the pedestrian (i.e. Left, Right and diagonal) are identified. Besides, the environmental features give out the conditions of the lighting, weather conditions and the lighting, etc. Moreover, the contextual features explain about the pedestrian's interactions with the background. Using the significant architecture, the YOLOv8 improves the performance of the object

detection, offering better flexibility, speed and accuracy. With this, the YOLOv8 effectively extracts the frames and is then fed to the Visual Casual model for identifying the relationships amid the various variables.

3.2.2 YOLOV8 with Visual Casual Modelling

In pedestrian prediction, after extracting the key frames, the associations amid different variables are identified using Visual Casual Model (VCM). The VCM incorporates contextual, behavioural and event-based prediction, in which the contextual understanding analyses the movement of the pedestrians (pedestrian is walking, pedestrian is crossing), beneath different weather, lighting and crowd density. The Scene Graph Generation (SGG) involved in the VCM aids in analysing the movement of the pedestrians. The SGG specifies the scenes with nodes (objects) and edges (relationships), which further organizes the information into the graphical structure. Generally, the standard SGG model possess an object detector and a relationship estimator, in which, the former identifies the object information and latter specifies the inter-object relation.

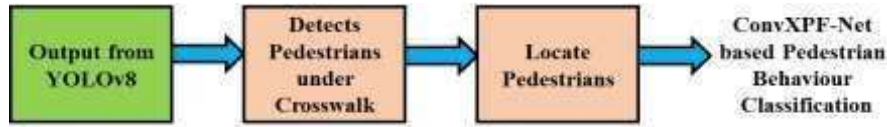


Figure 5: YOLOV8 with Visual Casual Modelling

The flow of YOLOV8 with Visual Casual Modelling is illustrated in Figure 5. The probability of the relationship instance includes, object from categories like A_i and A_j , which belongs to k th category and is given in equation (1) [28].

$$Prob(A_{ijk} = k) = \sigma(f_d(f_e(Z_{ij}))) \quad (1)$$

Here, f_e specifies the object detector, f_d denotes the relationship classifier, A_{ijk} specifies the triplet with i -th, j -th and with k -th relationship category, Z_{ij} denotes the original feature of image A_{ijk} . Moreover, the σ denotes function that converts outputs in probabilities. In recent times, the SGG research focus on model bias, in which the long-tail relationship is considered. Even though, it is counterintuitive, the counterfactual reasoning captures the variations and the total direct effect simulates the counterfactual reasoning, thereby erases specific parts of the image, so as to attain debiasing predictions and is given in equation (2).

$$Prob(A_{ijk} = k) = \sigma(f_d(f_e(Z_{ij})) - f_d(f_e(Z_y)))) \quad (2)$$

In which, the Z_{ij} and Z_y specifies the instance before and after the process. Hence, the VCM detects and estimates the relationship among the variables using SGG, which thereby improves the detection of pedestrians effectively, using more structured illustration of the scene. Further, the behaviour of the pedestrians is detected with the aid of ConvXPF-Net. The feed-forward computation is carried out at the bottom-up path, in which the feature maps at multiple scales are developed.

3.2 Proposed Convolutional XGBoost Pyramid Fractional Network (ConvXPF-Net)

In order to identify the behaviour of the pedestrians, while effectively reducing the issues with higher false negatives, due to distracted walking, ConvXPF-Net classifier is employed and it involves Pyramid Net Technique (PNT) along with Fractional Calculus Technique (FCT) at its hidden layers, for reducing the Error of Omission and subtle behavioural changes. The structural illustration of the proposed work is illustrated in Figure 6.

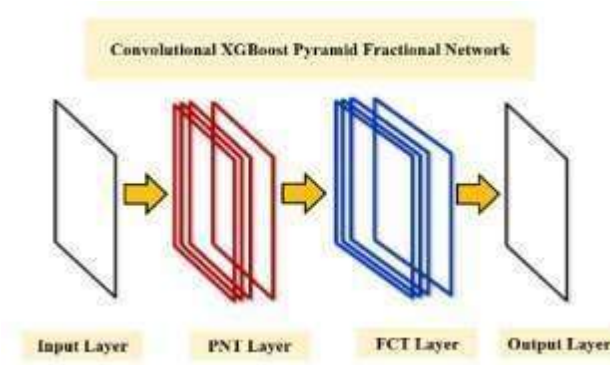


Figure 6: Convolutional XGBoost Pyramid Fractional Network

Input Layer: The output from the Y8-VCM is fed to the input layer and it assumes a training dataset U . If the dataset is in the specified format, it is forwarded to the Pyramid Net Technique (PNT) layer, in which, the feature at different spatial resolutions are captured, and it identifies the occluded part.

Pyramid Net Technique (PNT) in ConvXGB-Net: In order to deal with minimal Doppler shift, a PNT based ConvXGB-Net is developed, in which a single scale image is considered as input and is then featured as output maps at multiple levels. The PNT develops a top-down pathway along with lateral connections for combining the high-resolution spatial information from the lower levels that possess semantic rich features. The PNT has been developed over a back bone network, in which extracting feature maps at various levels, is the main function and is given in equation (3) [29].

$$E_l = cl_l(Im) \quad (3)$$

Where, Im denotes the input image, cl_l specifies the convolutional layers are various levels and E_l denotes the extracted feature maps at the specific level. With the aid of lateral connections, the higher-level semantic maps are up sampled, as well as merged along with lower level feature maps. The feature map for the next higher level N_l is given in equation (4).

$$N_l = Upsamp(N(l+1)) + Conv(E_l) \quad (4)$$

Here, $Upsamp(N(l+1))$ is employed for upsampling functioning and then the E_l is fed to align the total count of the channels before merging.

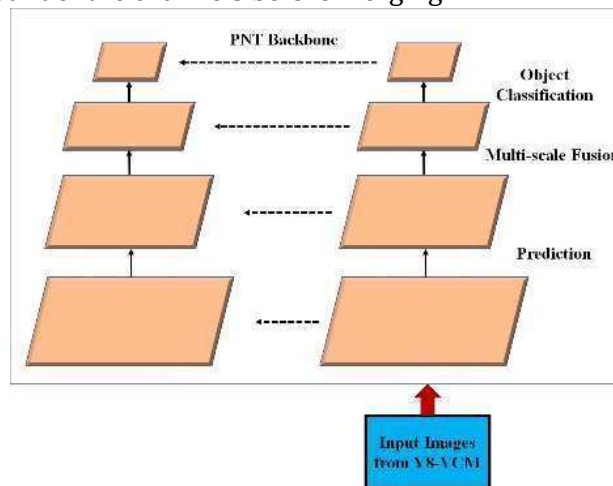


Figure 7: Pyramid Net Technique (PNT) in ConvXPF-Net

The Pyramid Net Technique (PNT) in ConvXGB-Net is shown in Figure 7. Finally, the final pyramid levels are developed with a stride convolution, in which every single N_l

maintains rich semantic over various scales that are capable for multi-scale pedestrian detection. Further, for identifying the subtle changes FC is employed.

Fractional Calculus (FC) in ConvXGB-Net: For recognizing the subtle behavioural changes, a powerful mathematical tool known as Fractional Calculus is employed and it aids in identifying the nuanced, as well as the subtle variations in the behaviour of the pedestrians. The FC includes three fractional derivatives namely Grunwald-Letnikov, Riemann-Liouville and Caputo. The fundamental theorem of FC includes, $fc : [i, j] \rightarrow \mathbb{R}$ is a continuous function and $FC : [i, j] \rightarrow \mathbb{R}$ is given in equation (5) [30].

$$FC(x) = \int_a^x fc(t)dt \quad (5)$$

In which FC is differentiable and is specified in equation (6),

$$F' = \frac{dF}{dt} = f \quad (6)$$

As the Fractional derivatives have long term memory, these are highly used in Neural Networks. In the defined work, the CNN, has been implemented using Grunwald-Letnikov derivative based back propagation method, in which an integer order gradient is employed amid two different layers and it aids in updating the parameters. The Grunwald-Letnikov based fractional derivative in the order of $\alpha(a^g d^q f)$ is specified in equation (7).

$$a^g d^q f(x) = \lim_{N \rightarrow \infty} \frac{\nabla_{h_N}^\alpha f(t)}{h_N^\alpha} \quad (7)$$

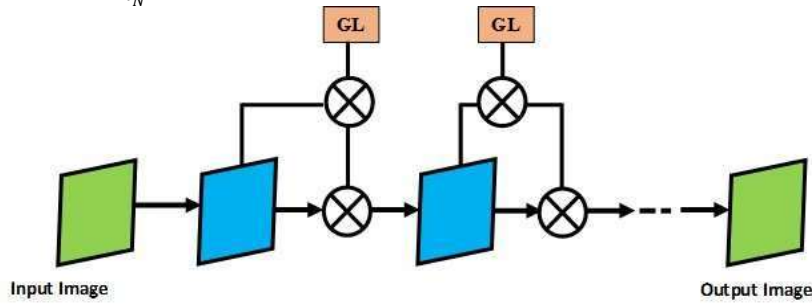


Figure 8: Architecture of Fractional Calculus (FC) in ConvXPF-Net

The Architecture of Fractional Calculus (FC) in ConvXPF-Net is illustrated in Figure 8. The Grunwald-Letnikov (GL) based fractional derivative assists CNN in optimizing the order of the mask for a specific image; the horizontal and vertical derivative of it is given in equation (8) and (9).

$$GL_x(s) = \begin{bmatrix} 0 & \frac{s-\alpha}{2} & 0 \\ 0 & -\alpha_s & 0 \\ 0 & 0 & 0 \end{bmatrix} * y(s) \quad (8)$$

$$GL_y(s) = \begin{bmatrix} 0 & \frac{s-\alpha}{2} & 0 \\ \frac{s-\alpha}{2} & -\alpha_s & 1 \\ 0 & 0 & 0 \end{bmatrix} * y(s) \quad (9)$$

The values of α_s for every single pixel in the CNN are estimated by the CNN with the learnable parameter. This thereby updates the efficiency of CNN, based on the weights and biases, which further aids in capturing the nuanced subtle behaviour of the pedestrians, while detecting the occluded or partially visible pedestrians.

Prediction of Class Labels using XGBoost: From the training data, the class labels are predicted by using feature learning in the convolutional layer and further using XGBoost the class labels are predicted. The XGBoost is the improved version of Gradient Boost Method and is known as Extreme Gradient Boost and is based on Decision Trees (DT). A DT is a single system that makes decisions based on all the image features, in which the

bagging technique combines with voting process for forming multiple decision trees; further Random Forest (RF) integrates with bagging and DT and takes decisions with respect to the sub-set of the randomly chosen features. Furthermore, on this basis, Gradient Boosting Method (GBM) develops a effective powerful tree-based learner with continuous iteration of gradient optimization algorithm. The XGBoost is an addition operation with k-tree functions and is specified in equation (10) [31].

$$xg_a^{(t)} = \sum_{k=1}^t f_k(x_a) = xg_a^{(t-1)} + f_t(x_a) \quad (10)$$

Where, $xg_a^{(t)}$ denotes the predicted output after adding one tree function, k specifies the number of additive trees, t denotes the number of iterations, $f_t(x_a)$ specifies the k th tree function. Further, it involves splitting, in which every single segment sorts the images based on the feature values, and if the gain split attained is positive, then the tree continues to split. Hence, with the ConvXPF-Net, the pedestrian behaviours like Walking, Looking, Standing and Crossing are effectively classified and then the motion of the Pedestrians are determined.

3.3 Determination of the movement of the Pedestrians

After effective classification of the behaviour of the pedestrians, the direction of motion of the pedestrians is determined. The defined work highly aids in detecting the behaviour of pedestrians beneath various lighting and weather conditions and the defined video-based analysis, gives out a detailed and systematic observation of even the microscopic movements of the pedestrians that includes the speed and placement. The proposed technique also has the capability to detect the behaviour of the pedestrians with respect to the obstacles, road users and the path width.

3.4 Speed and Break Automation

Once, the movement of the pedestrians are determined, the speed and direction of vehicle is automated. When certain potential collision is identified, then the system evaluates the speed, as well as the distance with the pedestrian, so as to identify the collision risk. In such cases, a warning or alert is given to the driver and if the risk is high, automatically the brake is engaged, which reduces the vehicle speed and prevents the collision.

3.5 Explainable AI (EXI)

In order to validate the significance of the proposed work, an Explainable AI is employed, as the ML and DL based AI techniques are often opaque, which hinders the limitations and trust of the proposed models. With the use of XAI, how the proposed model arrives to predict forecast and enables deep understanding about the factors that influence the output of the model is explained. The XAI in the proposed work employs LIME (Local Interpretable Model-agnostic Explanations) and it focuses on model explainability by perturbing the input. Besides, the LIME develops an interpretable surrogate model that optimizes the complex behaviour of the black-box model. LIME identifies the explainable model by decreasing the following objective functions given in equation (11) [31, 32].

$$\xi(L) = \arg \min_G \mathcal{L}(f, G, \pi_L) + \Omega(G) \quad (11)$$

In which the $\mathcal{L}(f, G, \pi_L)$ denotes the unfaithfulness of G in terms of f and $\Omega(G)$. LIME uses random generation of neighbours that provides the explanations related to very close instances, excluding the density of black-box outcomes. Hence, with XAI identifies the key risk factors, also quantitatively evaluates the degree of influence created by the proposed work.

Hence, with the proposed ConvXPF-Net effective pedestrian behaviour prediction is carried out with better reduction of Boundary discontinuity problem, missed detection, Error of Omission. In addition, with the significant classification carried out by

Convolutional XGBoost Pyramid Fractional Network, the superficial understanding due to subtle behaviour is addressed effectively, also with the assistance of XAI appropriate validation is carried out.

4. RESULTS AND DISCUSSION

The proposed research work ConvXPF-Net employs Y8-VCM, ConvXPF-Net and XAI for effective prediction of the behaviour of the pedestrians, which further aids in automatic driving system. In order to identify the significance of the proposed work, it is implemented in Python platform and to foreshow the significance of the defined work, it is analogized with conventional techniques and the corresponding performance metrics are analysed.

4.1 Dataset Description

The Dataset employed in the proposed pedestrian behaviour prediction is Joint Attention in Autonomous Driving (JAAD). The defined JAAD dataset mainly assists in the context of autonomous driving. The main concentration lies over the pedestrians crossing and the behaviour of the driver during the specified time, also its influence. The JAAD dataset involves 346 short video clips, each lasts for about 5-10 seconds, which are extracted from a 240 hours driving footage. Bounding boxes are provided for all the pedestrians, along with the occlusion tags that assists in pedestrian detection. The behaviour annotations intimates the behaviour of the pedestrians that necessitates the attention of the driver. Every single video clip possess, tags along with time-stamped behaviour labels. Besides, demographic attributes are also provided for each pedestrian and in addition, visible traffic scene elements are given for every single frame. With the JAAD dataset, an effective analysis is carried for the proposed model, so as to identify its significance.

4.2 System Configuration

For evaluating the significance of the proposed ConvXPF-Net, it is implemented in Python platform and the system configurations involved in the process are stated below.

Software Used	: Python
OS	: Windows 10 (64-bit)
Processor	: Intel i5
RAM	: 8GB

4.3 Simulation Results

The proposed research work for pedestrian behaviour detection assists in better identification and classification of behaviour with the aid of ConvXPF-Net model. At first, the videos from the dataset considered is fed as input to the Frame extractor, in which the frame necessary for classification are extracted.





Figure 9: Input Videos from Dataset

The sample input videos considered from the JAAD dataset is depicted in Figure 9 and initially it is pre-processed and the noises in it are removed and is then given to the feature extractor VFE, the frames essential for the pedestrian detection are extracted from the video file at specific interval of time.



Figure 10: Individual Video Frames Extracted using VFE

The Individual Video Frames Extracted using VFE is illustrated in Figure 10, which is then fed to Y8-VCM, in which the patterns of the frames are identified and the co-ordinates of the bounding-boxes are predicted. With effective reduction of the struggles faced by the regression head in identifying the body's orientation, the Y8-VCM gives better reduction of the Boundary discontinuity problem in angle prediction. The Y8-VCM mainly concentrates over the pedestrian cross walk, with respect to lighting, crowd density, and weather conditions, and is identified by the defined YOLOv8, while based these bounding boxes are developed with the aid of VCM, thereby reduces the discontinuity problem.



Figure 11: Bounding boxes formed by Y8-VCM

The Bounding boxes formed by Y8-VCM for pedestrian behaviour prediction is specified in Figure 11 and it gives out the likelihood of the pedestrians. Further, for decreasing the issues with missed detection and distracted walking ConvXPF-Net is employed, in which the PNT aids in multi-scale feature extraction. The PNT uses single input and gives out multiple level features with the aid of top-down path way and assists in detecting the occluded parts by capturing features at different intervals. Further for identifying the subtle behavioural changes that leads to superficial understanding FCT is employed and it uses fractional derivatives for identifying the unidentified behaviour. Since, the FCT possess long term memory, it is highly preferred in neural networks and in the defined work, the use of FCT improves the efficiency of the CNN. Then, from the training data, the XGBoost predicts the class labels, for which it employs decision trees that address the superficial understanding and is further classified using XGBoost.

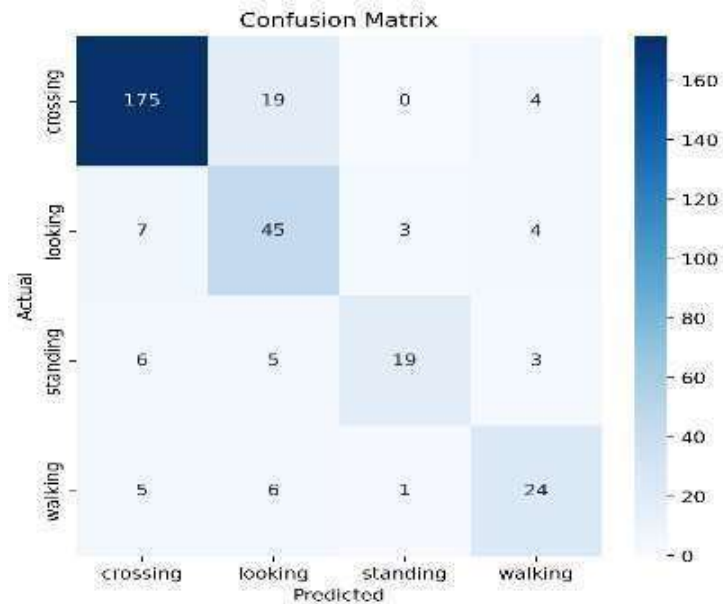


Figure 12: Confusion Matrix

The performance of the proposed ConvXPF-Net is evaluated using Confusion Matrix involved for Pedestrian Behaviour Detection is shown in Figure 12, which involves the actual and predicted values with respect to Walking, Standing, Looking and Crossing. In consideration with the confusion matrix, it involves 193 pedestrian crossing, in which 175 has been identified, while in case of Looking, 45 out of 75 has been identified. Moreover, out of 23 pedestrians standing, 19 has been detected and furthermore, with 35 Walking pedestrians, 24 is detected. The Confusion metrics mainly assists in detecting the frequently confused classes, which makes the CNN to generate error and this further aids in computing the performance metrics.



Figure 13: Classified Outputs

Further, based on the decision, the behaviour of the pedestrians like Walking, Standing, Looking and Crossing are classified using ConvXGBoost, which is depicted in Figure 13. Then, based on the classified output, the movement of the pedestrian is determined, which further aids in speed and brake automation (i.e.) if the pedestrian tries to cross the path, an intimation is given to the driver to slow down or to apply the brake.

**Case: 1****Case: 2****Case: 3**

Figure 14: Speed and Brake Automation Case 1: Pedestrian Standing, Case 2: Pedestrian Crossing and Case 3: Pedestrian crossed the path

The Speed and Brake Automation for Case 1: Pedestrian Standing, Case 2: Pedestrian Crossing and Case 3: Pedestrian Crossed the path is depicted in Figure 14.

Case 1: The Pedestrian is standing and it is not necessary for the driver to slow down or to apply brake.

Case 2: The pedestrian tries to cross the road and an information is given to the driver to slow down and stop the vehicle.

Case 3: The Pedestrian has crossed the path and the driver can speed up the vehicle. Further, the proposed work is validated using XAI that uses LIME for explaining the significance of the defined model, which further explains the effectiveness of the proposed model in Pedestrian Behaviour prediction.

4.4 Performance Metrics of the Proposed Model

The proposed work introduces a novel ConvXPF-Net for predicting and classifying the behaviour of the pedestrians, thereby giving out better reduction of boundary discontinuity problem, Error of Omission and also address the issues with superficial understanding. In order to analyse the importance of the defined work, the performance metrics namely Accuracy, Precision, Recall, F1 Score, AUC Score, False Positive Rate,

Specificity, Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are computed.

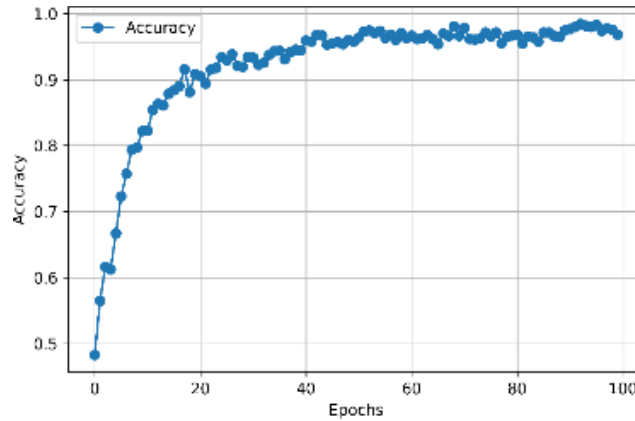


Figure 15: Accuracy of ConvXPF-Net

The Accuracy attained with the proposed ConvXPF-Net is depicted in Figure 15 and is plotted with respect to various epochs. From the figure, it is identified that the accuracy is about 90% at an epoch of 20 and steeply rises and reaches about 99.14% at an epoch of 100. As the Convolutional XGBoost Pyramid Fractional Network is employed for effective classification of pedestrian behaviour, the missed detections are rectified and rectifies the complexities with the erratic movements that leads to distracted behaviour, which further improves the accuracy of the model.

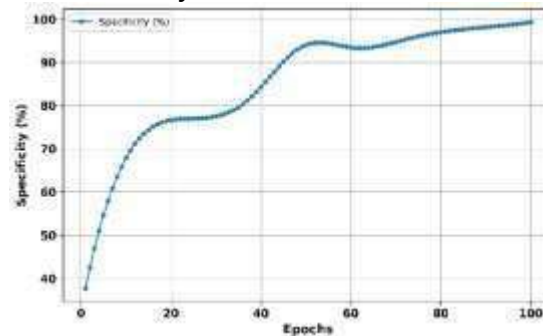


Figure 16: Specificity of ConvXPF-Net

Figure 16 illustrates the specificity of the Accuracy of ConvXPF-Net, in which the specificity is computed with respect to different epochs. The performance shows that the specificity stays at 88% at an epoch of 20 and increases to 99.11% at an epoch of 100. In case of pedestrian detection, the specificity aids in identifying the negative features that does not represent the pedestrians. The use of PNT detects the occluded parts and capturing the features at different resolutions and decreases the error of omission which further enhances the specificity and aids in recognizing the pedestrians.

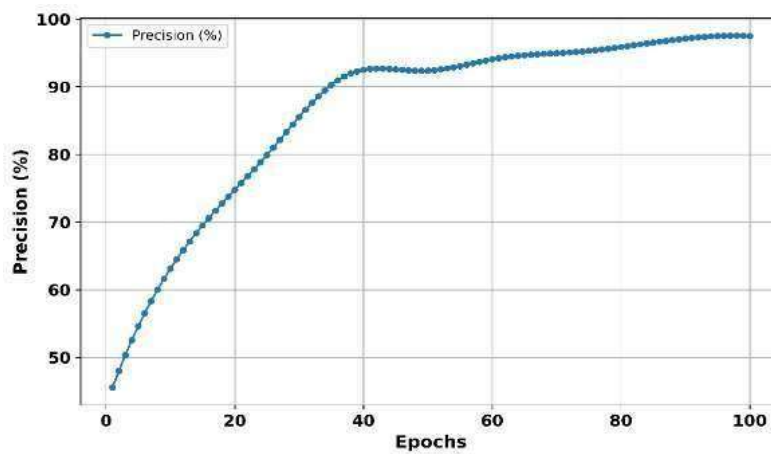


Figure 17: Precision of ConvXPF-Net

The Precision of ConvXPF-Net for the proposed work is depicted in Figure 17 and is computed with different epochs, in which the Precision is about 85% at an epoch of 20 and is 98.56% at an epoch of 100. The accuracy of the defined model in identifying the pedestrians from the image is measured as precision and the increase in precision possess less false positives and identifies the non-pedestrian as pedestrian. Since, FCT is employed in the defined work for PD, the subtle behaviour of the pedestrians, including the nuanced variations are identified accurately. Hence, the use of FCT in the classification stage improves the precision of the model.

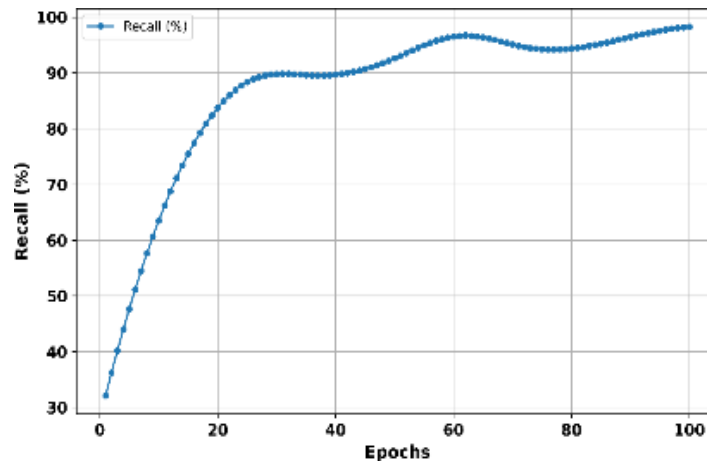


Figure 18: Recall of ConvXPF-Net

Figure 18 illustrates the Recall of ConvXPF-Net, which increases with increase of epochs. From the figure, it is noted that the Recall of the proposed work is about 82% at an epoch of 20 and reaches about 98.92% at an epoch of 100. The recall function specifies the ability of the defined model in identifying the pedestrians present in the video and it gives the ratio of total number of detected pedestrians to the total count of the actual pedestrians. For identifying the pedestrians in crosswalk from the video, the defined work employs Y8-VCN model, in which YOLOv8 appropriately identifies the pedestrians in the video frame, in consideration with different lighting and weather conditions, which improves the Recall of the defined model.

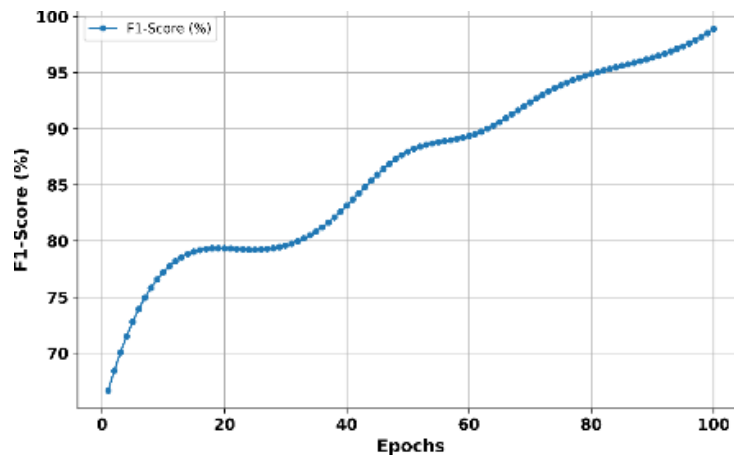


Figure 19: F1-Score of ConvXPF-Net

The performance representation of F1-Score of ConvXPF-Net is illustrated in Figure 19. At an epoch of 20, the F1-Score is identified to be 79% and at 100 epochs, it is about 98.89, which shows the enhancement if the performance. The F1-Score analyses the defined PD model by balancing the recall and precision function, thereby gives a single score that represents the effectiveness of the proposed model, while reducing the false positives and false negatives. In the defined work, a VCM is employed for analysing the relationship amid various features and the use of SGG in VCM highly aids in reducing the boundary discontinuity problem, which further improves the F1-Score of the model.

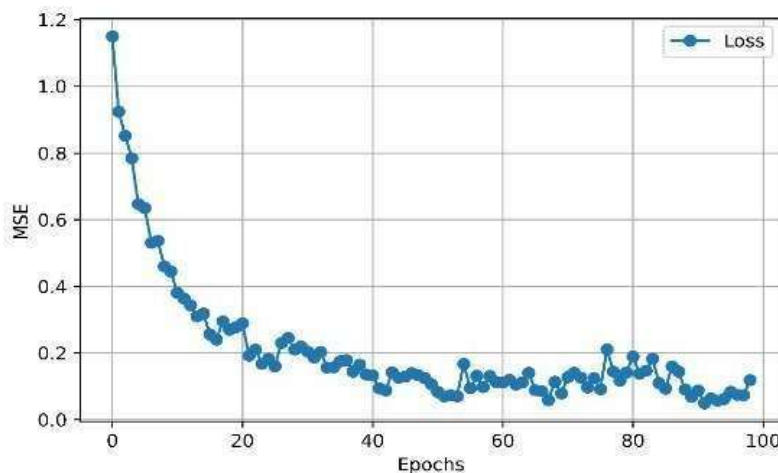


Figure 20: MSE of ConvXPF-Net

Figure 20 illustrates the MSE of ConvXPF-Net model, in which the MSE is plotted against different epochs. From the figure, it is identified that the MSE is about 0.3 at an epoch of 20 and is about 0.12 at an epoch of 100. The performance of the model is analysed in terms of loss function with the aid of MSE, which measures the square of difference amid the actual and the predicted values. The lesser the MSE, the better is the performance of the model, specifying accurate prediction of the model. The utilization of PNT brings multiscale extraction, reducing the error of omission, which lessens the MSE and thereby improves the performance of the model.

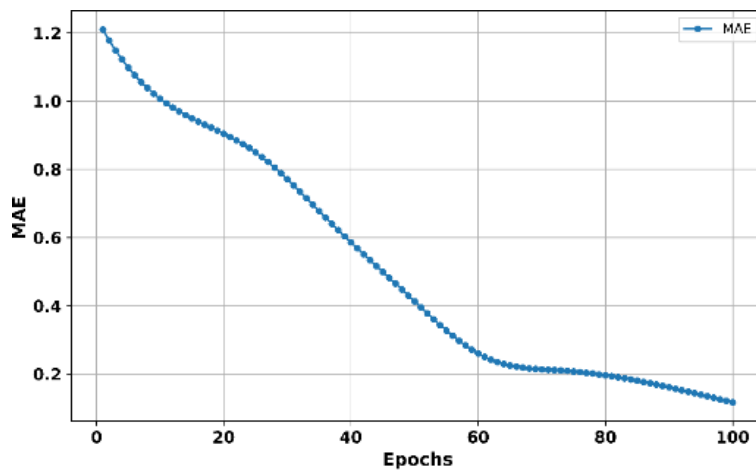


Figure 21: MAE of ConvXPF-Net

The MSE of ConvXPF-Net model for PD is illustrated in Figure 21, in which the MAE shows a drastic decrease. The figure clearly shows that the MAE is high at about 0.9 at an epoch of 20 and is less at about 0.13 at an epoch of 100. In a set of predictions performed, the MAE measures the average magnitude of errors. It is the ratio of the sum of absolute difference amid the predicted values and true values to that of the prediction counts. This shows the effectiveness of the model in predicting the positions of the bounding box. As Y8-VCM is employed in the defined work for capturing the full extent of the boundary boxes, reducing the issues with the boundary conditions, improving the MSE of the model.

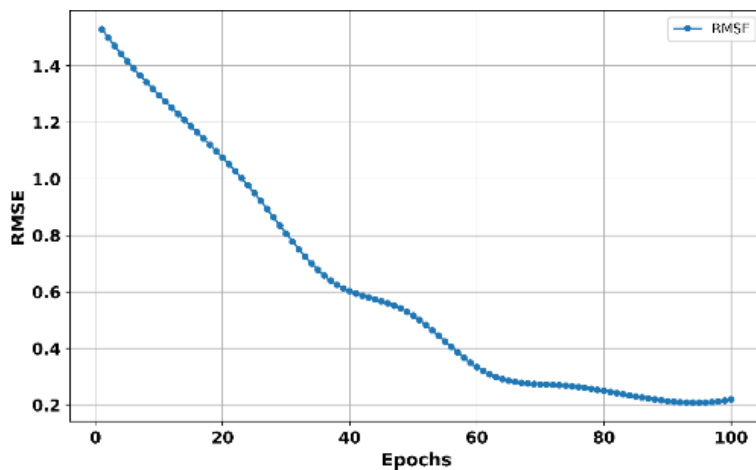


Figure 22: RMSE of ConvXPF-Net

Figure 22 illustrates the MSE of ConvXPF-Net model for PD, which depicts the plot amid the RMSE and epochs. The graph shows that at an epoch of 20, the RMSE is about 1.1 and at 100 epochs, the RMSE is about 0.21. The RMSE analyses the accuracy of predictions done with the pedestrian locations and it measures the average amid the models predicted and actual positions. The lesser the RMSE, higher is the accuracy of the model. The use of Y8-VCM in the defined model predicts the bounding boxes around the pedestrians, in which RMSE analyses the accuracy of these predictions. As VCM involves SGG along with it better bounding box formation is carried out, which gives out lesser RMSE.

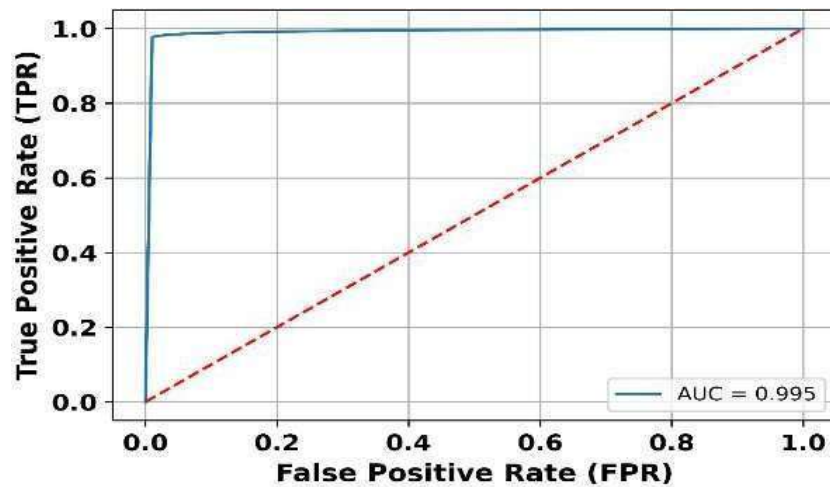


Figure 23: AUC of ConvXPF-Net

The AUC of ConvXPF-Net based PD is depicted in Figure 24 and it shows the plot amid the True Positive Rate (TPR) and The False Positive Rate (FPR). From the graph, it is identified that AUC value attained is about 0.995 and this because the AUC score summarizes the performance of the model beneath all classification thresholds. The plot in the figure represents the ROC that possess TPR and FPR at various threshold limits, giving out the relation amid the sensitivity and the specificity. The FCT of the proposed work, differentiates the relationship amid the pixels values of the image and then with the derivatives of the FCT, the pedestrian behaviour is classified.

4.1 Comparative Analysis

To foreshow the significance of the proposed work, it is analogized with conventional technique like Bayesian Ridge (BR) [3], Decision Tree (DT) [3], Multilayer Perceptron (MLP) [3], Support Vector Machine (SVM) [3], Variational Mode Decomposition (VMD) [16], (YOLOv3) [19], Bio-inspired Garra Rufa Optimization assisted Deep Learning Model (BGRODL-OC) [20], Adaptive Domain Matching Network (ADMN) [25], Multi-Description Training (MDT) [25], Suspicious Event Detection Network (SED-NET) [23], Deep Convolutional Neural Network based Crowd Density Monitoring- Intelligent Urban Planning (DCNNCDM-IUP) [33], YOLOv7-Global Mechanism (YOLOv7-GM) [34], First order Differential-Long Short Term Memory (Fod-LSTM) [35], Particle Swarm Optimization- long short Term Memory (PSO-5TM) [36], Hybrid Recurrent Network (HRN) [37], (YOLOv5) [38], Vision Transformer (VT) [39] and (YOLOv3) [40].

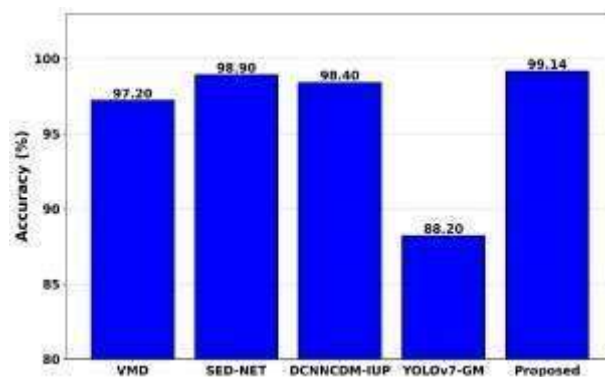


Figure 24: Accuracy Comparison of the proposed work

The Accuracy Comparison of the proposed work with conventional techniques like VMD, SED-NET, DCNNCDM-IUP and YOLOv7-GM is illustrated in Figure 24. From the figure, it is identified that the proposed work possess an accuracy of 99.14% compared to other technique like VMD, SED-NET, DCNNCDM-IUP and YOLOv7-GM that possess an accuracy of 97.20%, 98.90%, 98.40% and 88.20%. The analysis results shows that the proposed work performs better than the conventional techniques. The use of XGBoost along with Convolutional network improves the accuracy of the model, as it is a gradient enhanced decision tree algorithm. Besides, as it trains multiple decision trees, better accuracy is attained compared to other techniques.

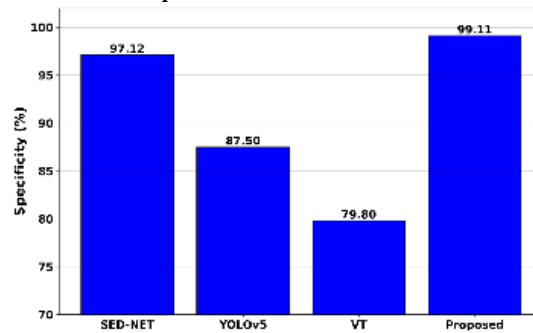


Figure 25: Specificity Comparison of the proposed work

Figure 25 depicts the Specificity Comparison of the proposed work with existing techniques like SED-NET, YOLOv5 and VT. It is identified that the conventional techniques namely SED-NET, YOLOv5 and VT has a specificity of about 97.12%, 87.50% and 79.80%, while the proposed possess a specificity of 99.11%. The comparative results illustrates that the conventional techniques has less specificity than the proposed work. The PNT technique used effectively address the scale variation complexities, by developing a feature pyramid having higher level semantics, this further improves the specificity of the model, compared to other conventional techniques.

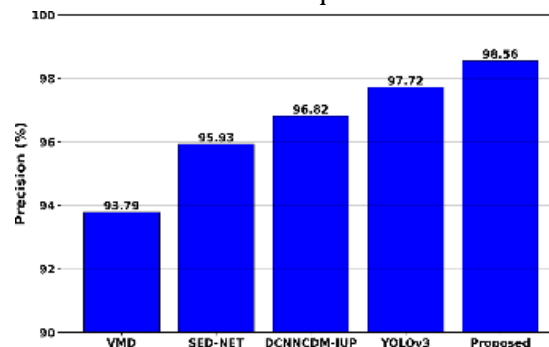


Figure 26: Precision Comparison of the proposed work

The Precision Comparison of the proposed work is shown in Figure 26, in which the proposed work is compared with conventional techniques like VMD, SED-NET, DCNNCDM-IUP and YOLOv3. The analogization results shows that the defined work incorporates a precision value of about 98.56% and the existing techniques namely VMD, SED-NET, DCNNCDM-IUP and YOLOv3 has a precision of about 93.79%, 95.93%, 96.82% and 97.72%, from which it is identified that the proposed work gives out better Precision value than the conventional techniques. Since, FCT is employed in the defined work, unlike other integer order functioning, the FCT uses fractional calculus that enables integration and differentiation of the orders, giving out effective precision than other techniques.

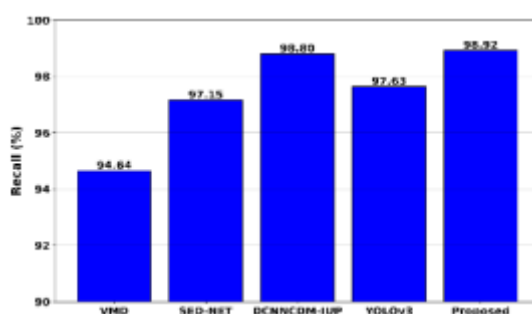


Figure 27: Recall Comparison of the proposed work

The Comparison of Recall metrics of the proposed work is shown in Figure 27, in which the defined work is analogized with existing techniques namely VMD, SED-NET, DCNNCDM-IUP and YOLOv3. From the comparison, it is identified that the proposed work has a recall metrics of about 98.92%, while the techniques like VMD, SED-NET, DCNNCDM-IUP and YOLOv3 has a recall value of about 94.64%, 97.15%, 98.80% and 97.63%, which shows higher value of the proposed work compared to others. The proposed work employs YOLOv8 for real time pedestrian detection, which predicts both the class labels, as well as the pedestrian locations in the image that further improves the Recall function effectively than other techniques.

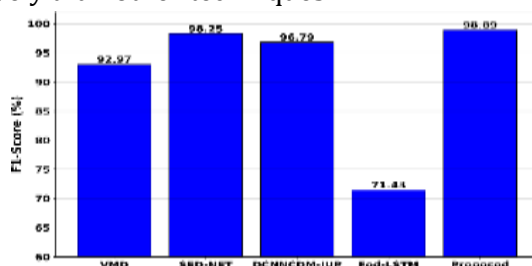


Figure 28: F1-Score Comparison of the proposed work

The F1-Score Comparison of the proposed work with prevailing techniques like VMD, SED-NET, DCNNCDM-IUP and Fod-LSTM is shown in Figure 28. From the figure, it is noted that the proposed work gives out a F1-Score value of about 98.89%, whereas the existing techniques namely VMD, SED-NET, DCNNCDM-IUP gives a F1-Score value of about 92.97%, 98.25%, 96.79% and 71.43%. The results attained accurately show that the proposed work performs better with respect to conventional techniques, in terms of F1-Score metrics. This is mainly because, the relationship amid different features are effectively analysed by the VCM and gives better F1-Score value than other techniques.

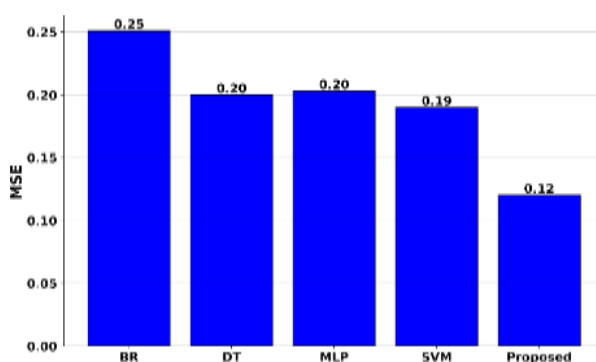
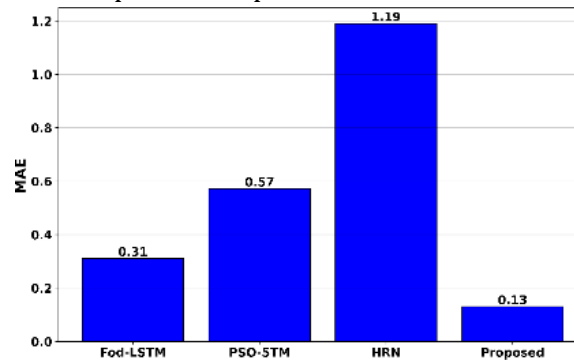
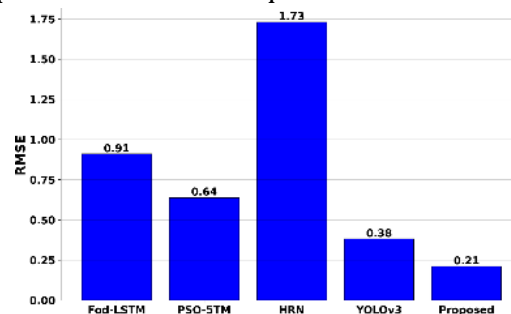


Figure 29: MSE Comparison of the proposed work

In the MSE Comparison, the defined work is compared with the prevailing techniques like BR, DT, MLP and SVM, which is illustrated in Figure 29. The comparison shown in the figure denotes that the proposed work has a MSE value of about 0.12, while other techniques like BR, DT, MLP and SVM has an MSE of about 0.25, 0.20, 0.20 and 0.19. The comparison results show that the proposed work performs better compared to other techniques. The proposed work shows reduced MSE than other techniques, because the PNT technique used significantly decreases the error of omission, which leads to higher false positives and further improves the performance of the model.

**Figure 30: MAE Comparison of the proposed work**

The MAE comparison of the defined work, which involves the analogization amid proposed work and conventional techniques like Fod-LSTM, PSO-5TM and HRN is depicted in Figure 30. The results denotes that the proposed work has an MAE value of about 0.13, while the other techniques like Fod-LSTM, PSO-5TM and HRN has an MAE value of about 0.31, 0.57 and 1.19. This intimates that the proposed work performs better due to Y8-VCM that brings proper alignment of the grid's axis and orientation of the body. This results in effective elimination of the boundary discontinuity probes, which gives out lesser MAE value, compared to other techniques.

**Figure 31: RMSE Comparison of the proposed work**

The Comparison of RMSE of the proposed work is illustrated in Figure 31, in which the proposed work is analogized with conventional techniques like Fod-LSTM, PSO-5TM, HRN and YOLOv3. From the analysis, it is concluded that the proposed work gives out a RMSE vale of about 0.21, while other techniques like Fod-LSTM, PSO-5TM, HRN and YOLOv3 gives out an RMSE value of about 0.91, 0.64, 1.73 and 0.38. The results shows the better performance of the proposed work, compared to other techniques, since effective capturing of full extend of bounding box around the pedestrian is performed by VCM.

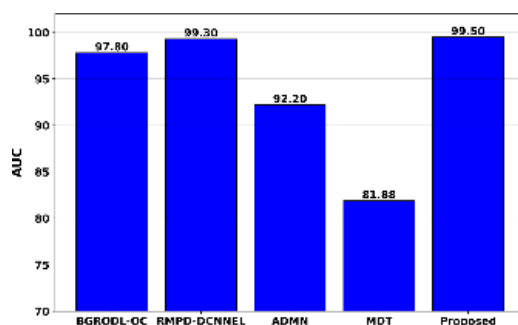


Figure 32: AUC Comparison of the proposed work

Figure 32 illustrates the AUC Comparison of the proposed work, which involves the comparison of conventional techniques like BGRODL-OC, RMPD-DCNNEL, ADMN and MDT, with the proposed work. From the comparison it is identified that the proposed work possess a RMSE value of about 99.50%, while the other techniques like BGRODL-OC, RMPD-DCNNEL, ADMN and MDT gives out a AUC value of about 97.80%, 99.30%, 92.20% and 81.88%. This concludes that the use of FCT in the classification layer aids in identifying the excluded images, which aids the proposed work performs better compared to other techniques.

From the discussion, it is concluded that the proposed ConvXPF-Net possess an Accuracy of about 99.14%, Specificity of about 99.11%, Precision of about 98.56%, Recall of about 98.92%, F1-Score of about 98.89%, MSE of about 0.12, MAE of about 0.13, RMSE value of about 0.21 and AUC value of 0.995, which shows the better performance of the proposed work compared to other techniques, thus aiding in effective pedestrian behaviour detection.

5. CONCLUSION

In this research work, a novel ConvXPF-Net with XAI is presented for Pedestrian Behaviour Detection, in which the necessary video frames are extracted from the input video. Then, using YOLOv8 based Visual Casual Model, the full extent of the boundary box is captured and further, the pedestrian behaviour like Walking, Standing, Looking and Crossing at the traffic are classified with the aid of Convolutional XGBoost Pyramid Fractional Network. Then, based on the classified behaviour, the movement of the pedestrian is estimated, followed by the speed and break automation. Further, the defined work is validated using the XAI, which analyses the features of the model, which uses the most for predicting the pedestrian behaviour. In order to analyse the significance of the proposed work, it is implemented in Python and gives better results with respect to performance metrics namely Accuracy of about 99.14%, Specificity of about 99.11%, Precision of about 98.56%, Recall of about 98.92%, F1-Score of about 98.89%, MSE of about 0.12, MAE of about 0.13, RMSE value of about 0.21 and AUC value of 0.995. From the analogization, it is concluded that the proposed model performs better compared to other conventional techniques in pedestrian behaviour detection.

Future Scope: In future, it is possible to improve the performance of the defined work by employing Hybrid technique in feature extraction stage. In addition, by optimizing the ConXGBoost network with different nature-inspired Metaheuristic algorithms, better hyper parameter tuning is attained, which thereby improves the classification accuracy of the pedestrian behaviour detection.

REFERENCES

- [1]. Zahedi, H., Mohammadinia, L., Dehghani, S.L., Habibzadeh, S. and Kheibar, N., 2024. The association between health literacy and pedestrian safety behavior among adults: a cross-sectional study. *BMC Public Health*, 24(1), p.1110.
- [2]. Acela, L.M.B., Sousa, E., Faria, S., Almeida, R. and Freitas, E.F., 2025. Pedestrian crossing behavior in social groups: Exploring cultural contexts through a comparative study. *Case Studies on Transport Policy*, 19, p.101333.
- [3]. Cai, J., Wang, M. and Wu, Y., 2024. Research on pedestrian crossing decision models and predictions based on machine learning. *Sensors*, 24(1), p.258.
- [4]. de Winkel, K.N. and Christoph, M., 2025. Rethinking Advanced Driver Assistance System taxonomies: A framework and inventory of real-world safety performance. *Transportation Research Interdisciplinary Perspectives*, 29, p.101336.
- [5]. Hussien, M.M., Melo, A.N., Ballardini, A.L., Maldonado, C.S., Izquierdo, R. and Sotelo, M.A., 2025. Rag-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models. *Expert Systems with Applications*, 265, p.125914.
- [6]. Guo, Y., Ma, S., Wei, F., Lu, L., Sun, F. and Wang, J., 2022. Analysis of behavior characteristics for pedestrian twice-crossing at signalized intersections based on an improved social force model. *Sustainability*, 14(4), p.2003.
- [7]. Reyes-Norambuena, P., Pinto, A.A., Martínez, J., Karbassi Yazdi, A. and Tan, Y., 2024. The Application of Machine Learning and Deep Learning with a Multi-Criteria Decision Analysis for Pedestrian Modeling: A Systematic Literature Review (1999–2023). *Sustainability*, 17(1), p.41.
- [8]. Liang, E. and Stamp, M., 2023. Predicting pedestrian crosswalk behavior using Convolutional Neural Networks. *Traffic injury prevention*, 24(4), pp.338-343.
- [9]. Alghodhaifi, H. and Lakshmanan, S., 2023. Holistic Spatio-Temporal graph attention for trajectory prediction in Vehicle–Pedestrian interactions. *Sensors*, 23(17), p.7361.
- [10]. Ham, J.S., Kim, D.H. and Moon, J., 2025. WatchoutPed: A dataset and model for Vulnerable Pedestrian Anticipation in surveillance videos. *Knowledge-Based Systems*, p.113352.
- [11]. Iaboni, C., Kelly, T. and Abichandani, P., 2025. NU-AIR: A Neuromorphic Urban Aerial Dataset for Detection and Localization of Pedestrians and Vehicles. *International Journal of Computer Vision*, pp.1-19.
- [12]. Liu, S., Zhang, L., Ge, J., Li, W. and Long, Y., 2025. Framework for timely perception of spatiotemporal crowd congestion risk in public spaces based on video pedestrian tracking and geographic mapping. *GIScience & Remote Sensing*, 62(1), p.2480416.
- [13]. Nomoto, K. and Yifan, C., 2025. Analysis of Increase in Accidents Between Pedestrians and Vehicles During Twilight: A Perspective of Pedestrians. *IEEE Access*.
- [14]. Pipkorn, L., Domeyer, J., Mehler, B., Reimer, B. and Gershon, P., 2025. Decoding the silent dialogue: Unveiling driver-pedestrian communication dynamics with a hidden Markov model. *Transportation Research Part F: Traffic Psychology and Behaviour*, 109, pp.965-976.
- [15]. Ahsan, M.J., Abdel-Aty, M., Anwari, N. and Barbour, N., 2025. Evaluating drivers' braking behavior at mid-block pedestrian crosswalks using video data and a mixed logit model with heterogeneity in the means. *Transportation Research Interdisciplinary Perspectives*, 31, p.101380.
- [16]. Hassani, S., Mustapha, S., Li, J., Mousavi, M. and Dackermann, U., 2025. Next-generation coupled structure-human sensing technology: Enhanced pedestrian-bridge interaction analysis using data fusion and machine learning. *Information Fusion*, p.102983.

- [17]. Han, C., Seshadri, P., Ding, Y., Posner, N., Koo, B.W., Agrawal, A., Lerch, A. and Guhathakurta, S., 2024. Understanding pedestrian movement using urban sensing technologies: the promise of audio-based sensors. *Urban Informatics*, 3(1), p.22.
- [18]. BaniSalman, M., Aljaidi, M., Elgeberi, N., Alsarhan, A. and Al Mamlook, R.E., 2025. VRDeepSafety: A Scalable VR Simulation Platform with V2X Communication for Enhanced Accident Prediction in Autonomous Vehicles. *World Electric Vehicle Journal*, 16(2), p.82.
- [19]. Iftikhar, S., Asim, M., Zhang, Z. and El-Latif, A.A.A., 2022. Advance generalization technique through 3D CNN to overcome the false positives pedestrian in autonomous vehicles. *Telecommunication Systems*, 80(4), pp.545-557.
- [20]. Yang, E., Shankar, K., Kumar, S. and Seo, C., 2023. Bioinspired Garra Rufa optimization-assisted deep learning model for object classification on pedestrian walkways. *Biomimetics*, 8(7), p.541.
- [21]. Aung, N.H.H., Sangwongngam, P., Jintamethasawat, R. and Wuttisittikulkij, L., 2024. Fully Synthetic Pedestrian Anomaly Behavior Dataset Generation in Metaverse for Enhancing Autonomous Driving Object Detection. *IEEE Access*.
- [22]. Kilcarslan, M. and Zheng, J.Y., 2022. DeepStep: Direct detection of walking pedestrian from motion by a vehicle camera. *IEEE Transactions on Intelligent Vehicles*, 8(2), pp.1652-1663.
- [23]. Siva Senthil, D. and Sivarani, T.S., 2025. SED-NET: Real-Time Suspicious Event Detection via Deep Learning-Based Di-Stream Neural Network. *International Journal of Computational Intelligence Systems*, 18(1), pp.1-19.
- [24]. Xu, J., Xu, D., Wu, J. and Shi, X., 2025. Modeling the collective behavior of pedestrians with the spontaneous loose leader-follower structure in public spaces. *Computer-Aided Civil and Infrastructure Engineering*.
- [25]. Jain, D.K., Zhao, X., Garcia, S. and Neelakandan, S., 2024. Robust multi-modal pedestrian detection using deep convolutional neural network with ensemble learning model. *Expert Systems with Applications*, 249, p.123527.
- [26]. Kaur, S., Kaur, L. and Lal, M., 2024. An effective Key Frame Extraction technique based on Feature Fusion and Fuzzy-C means clustering with Artificial Hummingbird. *Scientific Reports*, 14(1), p.26651.
- [27]. Yaseen, M., 2024. What is yolov9: An in-depth exploration of the internal features of the next-generation object detector. *arXiv preprint arXiv:2409.07813*.
- [28]. Liu, L., Sun, S., Zhi, S., Shi, F., Liu, Z., Heikkilä, J. and Liu, Y., 2025. A Causal Adjustment Module for Debiasing Scene Graph Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [29]. Shrivastava, A. and Poonkuntran, S.H.A.N.M.U.G.A.M., 2025. Multi-Scale Feature Pyramid Network with Camera Artifact Augmentation for Pedestrian Detection. *IEEE Access*.
- [30]. Coelho, C., Costa, M.F.P. and Ferrás, L.L., 2024. Fractional Calculus meets neural networks for Computer Vision: a survey. *AI*, 5(3), pp.1391-1426.
- [31]. Galli, A., La Gatta, V., Moscato, V., Postiglione, M. and Sperli, G., 2024. Explainability in AI-based behavioral malware detection systems. *Computers & Security*, 141, p.103842.
- [32]. Panduru, K. and Walsh, J., 2025. Exploring the Unseen: A Survey of Multi-Sensor Fusion and the Role of Explainable AI (XAI) in Autonomous Vehicles. *Sensors (Basel, Switzerland)*, 25(3), p.856.
- [33]. Mansouri, W., Alohal, M.A., Alqahtani, H., Alruwais, N., Alshammeri, M. and Mahmud, A., 2025. Deep convolutional neural network-based enhanced crowd density monitoring for intelligent urban planning on smart cities. *Scientific Reports*, 15(1), p.5759.

- [34]. Su, J., Wang, F. and Zhuang, W., 2025. An Improved YOLOv7 Tiny Algorithm for Vehicle and Pedestrian Detection with Occlusion in Autonomous Driving. *Chinese Journal of Electronics*, 34(1), pp.282-294.
- [35]. Zhang, H., Deng, J., Xu, Y., Deng, Y. and Lin, J.R., 2025. An Adaptive Pedestrian Flow Prediction Model Based on First-Order Differential Error Adjustment and Hidden Markov Model. *Buildings*, 15(6), p.902.
- [36]. Yan, S., Luo, B., Sun, X., Xiao, J., Ji, Y. and Ghazali, K.H.B., 2025. Accurate Localization Method Combining Optimized Hybrid Neural Networks for Geomagnetic Localization with Multi-Feature Dead Reckoning. *Sensors*, 25(5), p.1304.
- [37]. Zuo, Y., Jiang, J. and Yada, K., 2025. Application of hybrid gate recurrent unit for in-store trajectory prediction based on indoor location system. *Scientific Reports*, 15(1), p.1055.
- [38]. Sivalingan, H., 2024. Pedestrian Detection In Video Surveillance Using Yolo V5 With Light Perception Fusion. *ICTACT Journal on Soft Computing*, 14(4).
- [39]. Tahir, M. and Anwar, S., 2021. Transformers in pedestrian image retrieval and person re-identification in a multi-camera surveillance system. *Applied Sciences*, 11(19), p.9197.
- [40]. Haggui, O., Bayd, H. and Magnier, B., 2024. Centroid human tracking via oriented detection in overhead fisheye sequences. *The Visual Computer*, 40(1), pp.407-425.
- [41]. https://data.nvision2.eecs.yorku.ca/JAAD_dataset/
- [42]. Astha Modak, Samruddhi Paradkar, Shruti Manwatkar, Amol R. Madane, and Ashwini M. Deshpande. "Human head pose and eye state-based driver distraction monitoring system." In *Proceedings of 3rd International Conference on Computer Vision and Image Processing: CVIP 2018*, Volume 1, pp. 403-421. Singapore: Springer Singapore, 2019.
- [43]. Dinesh Kadam, Amol R. Madane, Krishnan Kutty, S. V. Bonde, "Rain Streaks Elimination using Image Processing Algorithms", *Signal & Image Processing: An International Journal (SIPIJ)*, Vol. 10, No. 3, pp. 21-32, Jun-2019
- [44]. Apeksha Kamble, Akanksha Singh, Damini Thusoo, Amol R. Madane, "Design and Implementation of Surround View Monitoring System in View of Autonomous Vehicle", *6th International Conference on Data Management, Analytics and Innovation 2022 (ICDMAI 2022)*, India, pp. 509-519, 14-16 Jan 2022.
- [45]. Ashish Kathe, Pranav Teli, Amol R. Madane, "COVID-19 Regulations Check: Social Distancing, People Counting and Mask Wear Check", *6th International Conference on Data Management, Analytics and Innovation 2022 (ICDMAI 2022)*, India, pp. 467-484, 14-16 Jan 2022.
- [46]. Priyojit Chakraborty, Amol R. Madane, "Vehicle Queue Management System at Fuel Station", *NIELIT's International Conference on Communication, Electronics and Digital Technologies (NICE-DT'23)*, New Delhi, India, 10-11 Feb 2023.ss