2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE),
25-27 April, 2024, Gazipur, Bangladesh

Video Based Vehicle Detection and Tracking using Image Processing and Deep Learning

Sabbir Ahmed*, Farhana Akter Tumpa*,
Sabiha Benta Sayed Badhon*, Lamia Anjum*
*Ahsanullah University of Science and Technology, Bangladesh
Email: ahmedsabbiraust@gmail.com, Tumpafarhanaakter@gmail.com,
sabiha.sayed.badhon@gmail.com, lamianjum123@gmail.com

***Abstract*—With the number of vehicles on the road expanding rapidly, it is imperative to maintain sophisticated vehicle systems and traffic management. To decrease the frequency of accidents, traffic management strategies that promote highway safety and pinpoint the causes of collisions should use vehicle identification and monitoring. A new elevated highway vehicle data set containing around 8,000 images extracted from videos with proper annotation is made from the perspective of Bangladesh which provides a complete data foundation for vehicle detection and tracking based on image processing and deep learning. Classification of the vehicle is done into 10 categories. For detection purposes, the YOLO v5 model, MASK R-CNN model, and SSD(Single Shot Detection) have been used while YOLO v5 performs better among all. For tracking, we have used the YOLO v5 model, DeepSORT framework, and GOTURN method where YOLO v5 is ahead among all. Following identification and tracking, the Yolo V5 model is used to estimate speed and count the number of vehicles. To verify the suggested strategy, multiple highway-monitored recordings are taken in different areas of Bangladesh.**

***Index Terms*—Deep Learning, Yolo V5, Mask R-CNN, SSD, DeepSORT, GOTURN**

I. Introduction

Developing an efficient and risk-free transport, intelligent transportation system plays a significant role in today's world [1]. Effective detection and tracking systems are important to acquire the processed data from appropriate methods. With a view to monitoring and tracking, surveillance cameras have been set on the roads. From the recorded video footage data extraction is performed and then various image processing algorithms are applied to monitor the motion of the vehicles, humans, or any other objects that have been recorded on the video footage. Thus, real-time evaluation of vehicles from traffic data has been started [2] recently. Though the conventional vehicle systems detected vehicles, usually they failed when there were background obstacles, occlusion, and bad weather. Nevertheless, the primary intent of an efficient traffic image assessment attempt to detect, track, and classify the vehicles precisely. Recently computer vision has now served as an active field in the detection and tracking of moving vehicles. Identifying vehicles from a vision-based video monitoring system and tracking can be done remarkably with the help of computer vision techniques [3] as it also offers various advantages such as preventing accidents, avoiding congestion, counting vehicles, lane changes, etc. [4]. In [5], a survey of object motion detection and tracking methods, including Optical flow, Image Registration Technique, Adaptive Background Subtraction, and Enhanced Dynamic Bayesian Network for detection, as well as Region-Based, Contour, 3D Model-Based, Feature-based, and Color and Pattern-Based tracking techniques, is presented. In [6], shadow and partial occlusion challenges in vehicle detection and tracking are tackled through specific motion, camera calibration, and motion-based methods, aiming to improve traffic management. Additionally, [7] introduces Yolo, a deep learning algorithm achieving 155 frames per second prediction speed using a single network, surpassing systems like DPM and R-CNN. Moreover, [8] presents an IOU tracker with a remarkable speed of 100K fps, outperforming conventional tracking methods. Furthermore, [9] utilizes YoloV5 and VisDrone-2020 to enhance detection accuracy for smaller objects, achieving a significant increase in mAP value and demonstrating the effectiveness of deep learning algorithms and image processing techniques for multi-modal data processing and optimization in detection and tracking tasks. Moreover, other deep learning algorithms and image processing techniques have been very useful in processing multi-modal data [18], statistical processing [19], and optimization approach [20]. In this paper, we have conducted a classification of vehicles as a computation of the number of vehicles is important that use highways and roadways. From large metropolitans to small towns data should be stored about vehicle classes that use those streets. In an overcrowded country like Bangladesh where a variety of vehicles can be seen, this self-built data set can become an enriched database for performing research in this area. The inspiration of this study is to identify the cause of the accident, estimate the speed of the vehicles, as well as keep an eye on all vehicles in a specific time zone.

II. Methodology

We have tried to experiment ourselves with detecting the vehicles from a video using our custom dataset. Below we are presenting our experimental methodology as a flow chart diagram :

979-8-3503-8828-2/24 $31.00 © 2024 IEEE

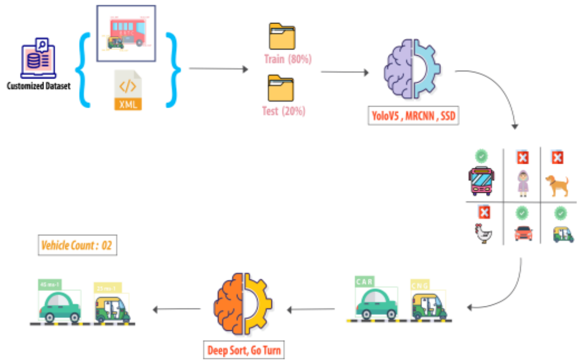


Fig. 1: Experimental Methodology Diagram.

A. Dataset Collection

We have made our custom dataset consist of local vehicle information of our country. As we are going to detect and track the local vehicles the dataset needs to be concrete and should have quality attributes. Therefore we collected data by taking video shots from different angles which have proper information regarding the vehicles.

B. Data Pre-processing

We have recorded videos of Traffic from different roads from which we first generated images from those videos. Images are selected given priority to vehicles in a frame, clear view, and not hazy objects. Images are strictly resized 64 px X 64 px. And for training efficiency size is reduced to under 40kb. For keeping track, images are sorted in numeric order.

C. Data Annotation

More than 8,000 frames are annotated with bounding boxes of vehicles are labeled. Our dataset contains videos with large variations in scale, pose, illumination, occlusion, and background clutters. We have introduced a total of 10 classes including Bus, Bike, Cycle, CNG, Car, Leguna, Rickshaw, Truck, Van, and Ambulance. At first, we have converted the videos into frames. The annotation has been done on the frames using MakeSense.AI [15], labeling [16], and VGG annotator tools [17]. As we used the models YOLOv5, SSD, and Mask R-CNN to validate our dataset, we had to create the dataset according to the format of each model. In YOLO labeling format, a .txt file with the same name is created for each image file in the same directory. Each .txt file contains the annotations for the corresponding image file, that is object class, object coordinates, height, and width. For each object, a new line is created. For SSD, the labeling format should be in Pascal VOC which stores annotation in XML file. For Mask R-CNN, we have used the VGG annotator for labeling purposes. Our dataset is divided into training and testing sets, with 8000+ and 1000+ sequences, respectively. We have training videos that are taken at different locations from the roads of Bangladesh as well as testing videos, but we ensure the training and testing videos have similar traffic conditions and attributes. This setting reduces the chances of detection or tracking methods to overfit particular scenarios.

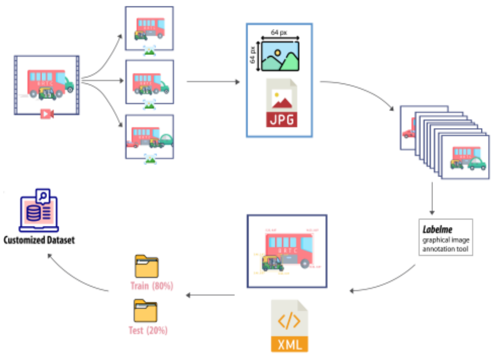


Fig. 2: Workflow of preparing custom dataset

III. Experiment And Analysis

In our work, For detection, we have used YoloV5, Mask R-CNN, and Single Shot Algorithm. Moreover, For tracking purposes, We have used DeepSort, GOTURN, and YoloV5 to train our model.

A. Analysis for Proposed Models in Vehicle Detection

We have trained the model using 8000+ data and measured the performance for epoch 60. The performance for YOLO v5 for 8000+ data is expressed in the table below:

In Table -I, the YoloV5 model is trained using 8000 data, and each classified vehicle such as - Bus, Bike, Cycle, etc. has illustrated their mAP, Precision, and Recall for the epoch of 60. Table - II shows that, YoloV5, Mask R-CNN, and SSD have been trained using 8000 data frames and comparison is performed based on their mAP, Precision, and Recall values for epoch. However, YoloV5 has outperformed the other two models.

TABLE I: Detection Result of YoloV5 using custom dataset for 8000+ frames

Classes	Epoch	mAP	Precision	Recall
Bike	60	0.382	0.378	0.542
Bus	60	0.529	0.495	0.495
Cycle	60	0.317	0.63	0.329
CNG	60	0.592	0.689	0.509
Car	60	0.744	0.81	0.687
Leguna	60	0.44	0.425	0.405
Rickshaw	60	0.384	0.718	0.289
Truck	60	0.31	0.406	0.424
Van	60	0.603	0.747	0.588

Our training experiments show that in Figure - 3 initially, the Mask R-CNN model achieves better accuracy (in terms of mAP) than the SSD model. Yolo V5 was ultimately selected as it gives higher accuracy (in terms of mAP)and runs faster while managing better performance in detecting small objects.

TABLE II: Comparison Result of Deep Learning Models using a Custom Dataset for 8000+ frames

Models	Epoch	mAP	Precision	Recall
Yolo V5	60	73.9	72.4	70.2
Mask R-CNN	60	55.4	37.9	54.2
SSD	60	56.4	70.1	32.9

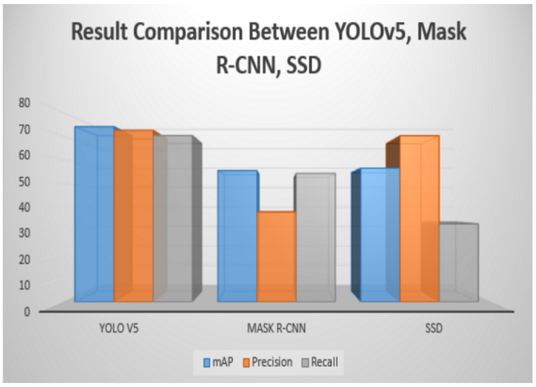


Fig. 3: Comparison Result of Deep Learning models using custom dataset for 8000+ frames

B. Analysis for Proposed Models in Vehicle Tracking

For tracking with DeepSORT, at first We used YOLO v5 for detection then DeepSort to track our model and we trained the model using around 8000+ data and measured the performance for epoch 60.

TABLE III: Tracking Result of YoloV5 using custom dataset for 8000 frames

Classes	Epoch	mAP	Precision	Recall
Bus	60	0.82	0.56	0.907
Bike	60	0.482	0.393	0.69
Cycle	60	0.592	0.61	0.63
CNG	60	0.763	0.793	0.665
Car	60	0.82	0.745	0.802
Leguna	60	0.865	0.76	0.905
Rickshaw	60	0.46	0.662	0.369
Truck	60	0.612	0.635	0.564
Van	60	0.452	0.50	0.550

In Table III and Table IV, tracking results of YoloV5 and DeepSORT are shown for each class. These two models track BUS with higher mAP values. In addition, YoloV5 has outperformed DeepSORT.
From the above figure 4, we can see the performance of the model by analyzing the Precision-Recall(PR) curve.

Here in Figure 5, The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers: higher recall and higher precision. Here

TABLE IV: Tracking Result of DeepSORT using a custom dataset for 8000 frames

Classes	Epoch	mAP	Precision	Recall
Bus	60	0.721	0.585	0.758
Bike	60	0.419	0.383	0.634
Cycle	60	0.613	0.724	0.513
CNG	60	0.709	0.659	0.622
Car	60	0.826	0.782	0.752
Leguna	60	0.703	0.828	0.524
Rickshaw	60	0.533	0.695	0.428
Truck	60	0.499	0.415	0.485
Van	60	0.541	0.704	0.588

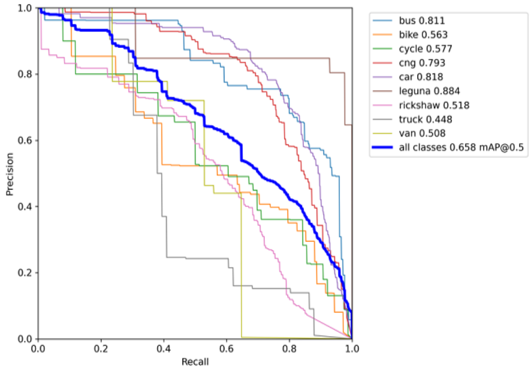


Fig. 4: PR curve of YOLOv5 using custom dataset for around 8000 frames

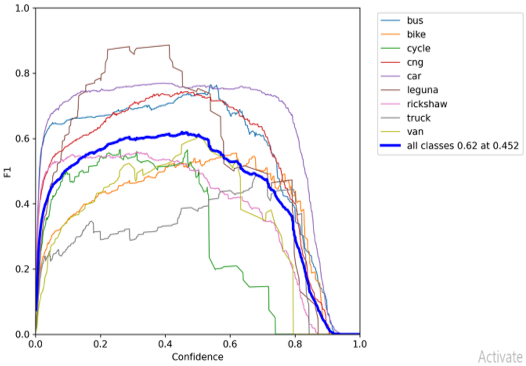


Fig. 5: F1 Score of YOLOv5 using custom dataset for around 8000 frames

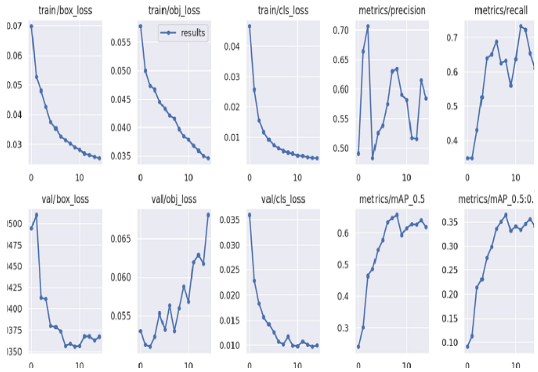


Fig. 6: Loss curve of YOLOv5 using custom dataset for around 8000 frames

Figure 6, we are showing the boxloss and classloss for both the training period and the validation period. By measuring box loss it can be identified how ”tight” the predicted bounding boxes are to the ground truth object, which usually refers to regression loss. Through class loss it can be measured the correctness of the classification of each predicted bounding box. Each box may contain an object class or a ”background”. This loss is usually called cross-entropy loss.

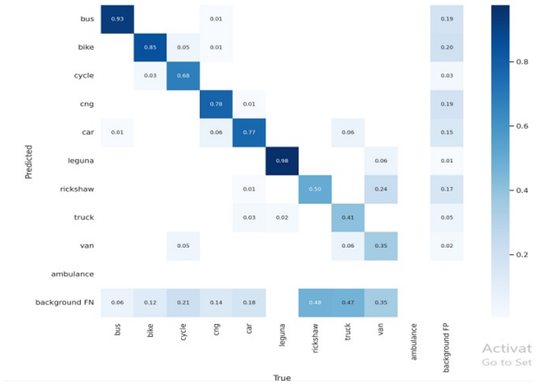


Fig. 7: Confusion Matrix of YOLOv5 using custom dataset

In Figure 7, the confusion matrix has been plotted to visualize important predictive analytics like recall, accuracy, and precision. It gives direct comparisons of values like True Positives, False Positives, True Negatives, and False Negatives. we found out that the actual true value with those predicted by YoloV5, Leguna was the highest. However, the van was among the lowest. From Figure 8 confusion matrix we found out that the actual true value with those predicted by the DeepSORT, car, and bus are among the highest. However, rickshaws and vans are among the lowest. Here Figure 9, we are showing the boxloss and classloss for both the training period and validation period. From this curve of Figure 10, we can conclude that the training loss continues to decrease until the end of training. With the increase of iteration and time elapsed the average loss is decreasing gradually.

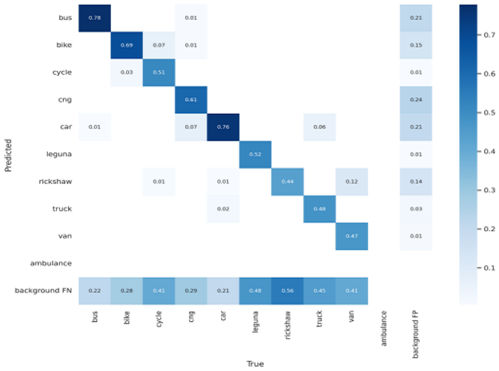


Fig. 8: Confusion Matrix of DeepSORT using custom dataset

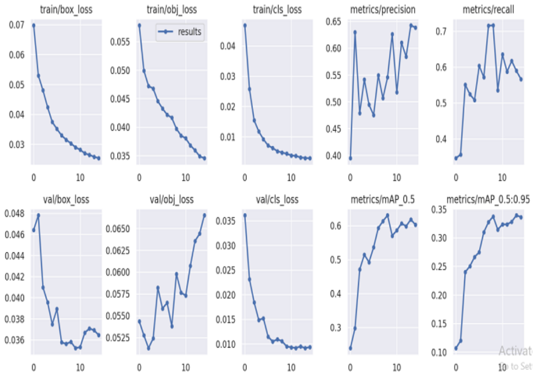


Fig. 9: Loss curve of DeepSORT using a custom dataset for around 8000 frames

C. Result of speed estimation

While using weights of YOLO v5, the model runs on 10 pre-defined classes. Based on that result, the counting vehicle part is shown. The Speed Estimation part is creating the frame converting pixels into calculating values and applying them to formulas. This gives the individual result showing a tracking ID, bounding box ID, lane position of the vehicle, and approximate speed. As the output depends on video input, the

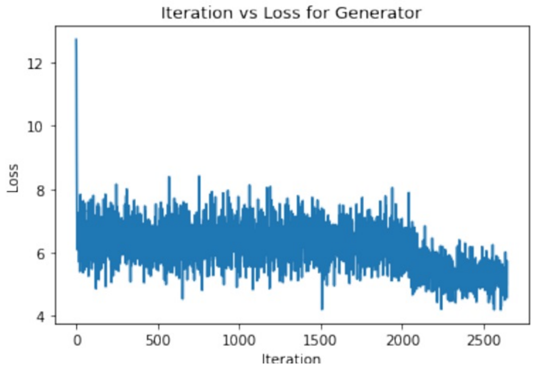
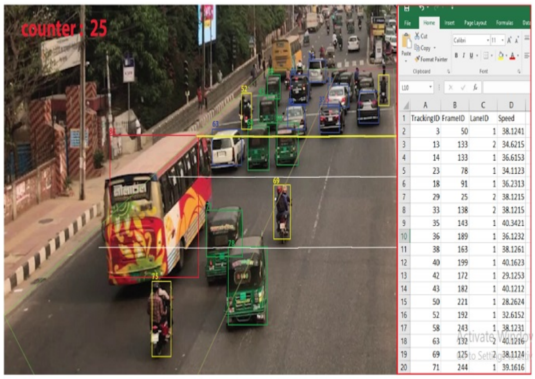


Fig. 10: Loss curve of GOTURN using a custom dataset for around 8000 frames

pixels, and other variables, as well as vehicle speed, may differ in some cases.



	A	B	C	D
1	TrackingID	FrameID	LaneID	Speed
2	3	50	1	38.1241
3	13	133	2	34.6215
4	14	133	1	36.6153
5	23	78	1	34.1123
6	18	91	1	36.2313
7	29	25	2	38.1215
8	33	138	2	38.1215
9	35	143	1	40.3421
10	36	189	1	36.1232
11	38	163	1	38.1261
12	40	199	1	40.1623
13	42	172	1	29.1253
14	43	182	1	40.1212
15	50	221	1	28.2624
16	52	192	1	32.6152
17	58	243	1	38.1231
18	63	132	2	40.1216
19	69	125	2	38.1124
20	71	244	1	39.1616

Fig. 11: Output of Speed Estimation and Counting

IV. Conclusion

In our work, the YoloV5 model has classified 10 types of vehicles – bus, car, bike, rickshaw, ambulance, laguna, van, cycle, CNG, and rickshaw in different scenarios and achieved a good result. From our experimental analysis, we find that SSD can make good detection with good speed whereas Mask R-CNN can make slow detection. We got different accuracy for them. YoloV5, Mask R-CNN, and SSD gave us mAP of 73.9%, 55.4%, and 56.4% respectively. As Go Turn only tracks a single object it is not that worthy of tracking. For tracking, YoloV5 and Deep Sort gave us mAP of 65.2% and 61.8% respectively. To identify vehicles, counting the number of vehicles and speed estimation we also used YoloV5. In addition, the methodology and results of the vehicle detection, tracking, and counting system provided in the analysis in our work will become important references for our local transport and for ensuring a safe transport system.

V. Referrences

[1] Sodhro, Ali Hassan, et al. ”Towards 5G-enabled self-adaptive green and reliable communication in intelligent transportation system.” IEEE Transactions on Intelligent Transportation Systems 22.8 (2020): 5223-5231.

[2] Mallikarjuna, Ch, A. Phanindra, and K. Ramachandra Rao. ”Traffic data collection under mixed traffic conditions using video image processing.” Journal of transportation engineering 135.4 (2009): 174-182.

[3] Spencer Jr, Billie F., Vedhus Hoskere, and Yasutaka Narazaki. ”Advances in computer vision-based civil infrastructure inspection and monitoring.” Engineering 5.2 (2019): 199-222.

[4] Shokravi, Hoofar, et al. ”A review on vehicle classification and potential use of smart vehicle-assisted techniques.” Sensors 20.11 (2020): 3274.

[5] S. Sri Jamiya and E. Rani, “A survey on vehicle detection and tracking algorithms in Real-time video surveillance

[6] R. A. Hadi, G. Sulong, and L. E. George, “Vehicle detection and tracking techniques: a concise review,” arXiv preprint arXiv:1410.5894, 2014.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE Conference on Computer Vision and pattern recognition, pp. 779–788, 2016.

[8] E. Bochinski, V. Eiselein and T. Sikora, ”High-Speed tracking-by-detection without using image information,” 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078516. keywords: Detectors;Computational efficiency;Object tracking;Face recognition;Visualization,

[9] W. Zhan, C. Sun, M. Wang, J. She, Y. Zhang, Z. Zhang, and Y. Sun, “An improved yolov5 real-time detection method for small objects captured by UAV,” Soft Computing, pp. 1–13, 2021

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, realtime object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, 2016.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE International conference on computer vision, pp. 2961–2969, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in European conference on computer vision, pp. 21–37, Springer, 2016.

[13] Dang, Tuan Linh, Gia Tuyen Nguyen, and Thang Cao. ”Object tracking using improved deep SORT YOLOv3 architecture.” ICIC Express Letters 14, no. 10 (2020): 961-969.

[14] Held, D., Thrun, S., Savarese, S. (2016). Learning to Track at 100 FPS with Deep Regression Networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9905. Springer, Cham.

[15] “Make Sense.” https://www.makesense.ai/. Accessed: 2021-05-10.

[16] “tzutalin/labelImg.” https://github.com/tzutalin/labelImg. Accessed: 2021-06-15.

[17] Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In Proceedings of the 27th ACM International Conference on Multimedia (MM ’19), October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3343031.3350535.

[18] M. S. Islam, M. S. Rahman and M. A. Amin, ”Beat Based Realistic Dance Video Generation using Deep Learning,” 2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON), Dhaka, Bangladesh, 2019, pp. 43-47, doi: 10.1109/RAAICON48939.2019.22.

[19] A. S. M Jahid Hasan, M. S. Rahman, M. S. Islam and J. Yusuf, ”Data Driven Energy Theft Localization in a Distribution Network,” 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2023, pp. 388-392, doi: 10.1109/ICICT4SD59951.2023.10303520.

[20] A. S. M. Jahid Hasan, J. Yusuf, M. S. Rahman and M. S. Islam, ”Electricity Cost Optimization for Large Loads through Energy Storage and Renewable Energy,” 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2023, pp. 46-50, doi: 10.1109/ICICT4SD59951.2023.10303409.