

## Performance investigation of two-stage detection techniques using traffic light detection dataset

Sunday Adeola Ajagbe<sup>1,4</sup>, Adekanmi Adeyinka Adegun<sup>2</sup>, Ahmed Babajide Olanrewaju<sup>3</sup>, John Babalola Oladosu<sup>4</sup>, Matthew Olusegun Adigun<sup>5</sup>

<sup>1</sup>Department of Computer Engineering, First Technical University, Ibadan, Nigeria

<sup>2</sup>School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, South Africa

<sup>3</sup>Department of Computer Science, University of Ibadan, Ibadan, Nigeria

<sup>4</sup>Department of Computer Engineering, Ladoko Akintola University of Technology LAUTECH, Ogbomoso, Nigeria

<sup>5</sup>Department of Information Technology, Cape Peninsula University of Technology, Cape Town, South Africa

### Article Info

#### Article history:

Received Feb 7, 2023

Revised May 5, 2023

Accepted May 7, 2023

#### Keywords:

Artificial intelligence  
Convolutional neural network (CNN)  
Deep learning  
Faster-CNN  
Object detection  
Two-stage detection

### ABSTRACT

Using a camera to monitor an object or a group of objects over time is the process of object detection. It can be used for a variety of things, including security and surveillance, video communication, traffic light detection (TLD), object detection from compressed video in public places. In recent times, object tracking has become a popular topic in computer science particularly, the data science community, thanks to the usage of deep learning (DL) in artificial intelligence (AI). DL which convolutional neural network (CNN) as one of its techniques usually used two-stage detection methods in TLD. Despite all successes recorded in TLD through the use of two-stage detection methods, there is no study that has analyzed these methods in experimental research, studying the strength and witnesses by the researchers. Based on the needs this study analyses the applications of DL techniques in TLD. We implemented object detection for TLD using 5 two-stage detection methods with the traffic light dataset using a Jupyter notebook and the sklearn libraries. We present the achievements of two-stage detection methods in TLD, going by standard performance metrics used, FASTER-CNN was the best in detection accuracy, F1-score, precision and recall with 0.89, 0.93, 0.83 and 0.90 respectively.

*This is an open access article under the [CC BY-SA](#) license.*



### Corresponding Author:

Sunday Adeola AJAGBE

Department of Computer Engineering, First Technical University, Ibadan

Km. 15, Lagos Express Way, Ibadan, Nigeria

Email: Sunday.ajagbe@tech-u.edu.ng

## 1. INTRODUCTION

The goal of deep learning (DL), a subfield of machine learning (ML) and artificial intelligence (AI) is to recreate and imitate in computers the architecture through which the brain conducts vision. DL has many advantages in image processing and computer vision generally, these advantages include; there is no manual setting of a learning rate, the automatic setting of a learning rate helps the accuracy of DL-based models, over gradient descent, minimal computation is referred, suitable for both local and distributed environments and robust in the face of large gradients, noise, and architecture selection [1]. Using a camera, object tracking involves following a specific object or set of objects over time. It is an intelligence-based system because a non-living things object is performing the roles that are meant to perform by living things objects. The surveillance cameras have a wide range of applications, including security and surveillance, lane detection (LD), pedestrian detection (PD), traffic light detection (TLD), traffic sign detection (TSD), and vehicle

detection (VD), object detection from complete video (ODFCV), object detection from compressed video public areas such as airports and subway stations [2], [3]. Computer vision (CV) has been particularly successful in the areas of object detection and tracking. High-level, intricate abstractions are extracted as data representations by DL models. Robots, self-driving cars, scene interpretation, and video surveillance are a few examples of how DL transforms object detection and tracking into an intelligence system [4].

The capacity to recognize and track significant things in near real-time is highly wanted by manufacturers who are striving toward streamlining manufacturing in today's competitive production environment. It was almost impossible to keep track of every object in a complicated manufacturing plant by hand; thus, an automation system with that capacity is desperately needed [5]. In a related study, [6] opined that sharing the roads with human drivers, autonomous terrestrial vehicles must be able to detect traffic lights and recognize their current states, which has always been an issue. The majority of the time, human drivers were able to recognize the appropriate traffic lights. Integration of recognition with past maps has a frequent technique for autonomous cars to deal with this problem. Although, for the purpose of identifying and detecting traffic lights, an extra solution is necessary. DL algorithms have demonstrated excellent performance and generalization power in a variety of issues, including traffic congestion.

Methods for recognizing traffic lights that use onboard sensors in vehicles have been actively researched [7]. The advancement in image processing methods that have been largely employed to sequence images acquired by in-vehicle cameras in many manners seems helpful. The successes might be because of their superior categorization performance, and learning-based approaches have grown in popularity [8]. Meanwhile, due to the presence of many disturbance variables in outdoor contexts, like incomplete light forms, dark light conditions, and partial occlusion, detection accuracy remains inadequate [9]. These issues are difficult to solve using computer vision and image processing methods, the difficulties require an investigation. Although vision-based autonomous driving has shown great promises, there has still the issue of analyzing the complex traffic scenario using the data collected. Recently, autonomous driving has been broken down into multiple tasks utilizing various models, such as object detection and intention identification [10].

Although DL models have very powerful object detectors, especially the generic object detector (one-stage and two-stage detection technique), the effectiveness of the DL-based models in TLD is crucial to the users (most importantly, the motorist, traffic operators, and pedestrians) for the upcoming vehicle identification, which is a critical task for self-driving cars. DL-based models are not generalized to new and unseen traffic light scenarios, including different types of traffic lights and intersections. Therefore, analysis of the usage of DL models in TLD to detect the environment, and appraise the applications of the known DL models in the TLD projects have not been duly attended to in the academic community. Based on this requirement, this study has the following objectives; i) Carry out an experimental comparison of convolutional neural network (CNN) two-stage algorithms for TLD; ii) Evaluation of the two-stage detection methods implemented in (iii) using standard evaluation metrics (detection accuracy (DA), F1-score, precision and recall in TLD. The remaining parts of this paper are organized in the following ways: Section 2 is the review of the related works. In section 3, the traffic light detection algorithm using pre-trained CNN is explained, and in section 4 analysis of the experimental results and discussion are compiled. The authors conclude the study in section 5.

## 2. RELATED STUDIES

Salari *et al.* (2022) [11] provided a detailed analysis of datasets in the highly researched areas of object recognition. About 160 datasets were examined using statistics and descriptions. In addition, the paper provided a discussion of the metrics frequently used for evaluation in the CV community, together with an overview of the most well-known object recognition benchmarks and contests. Li *et al.* (2020) [12] enumerated detection tracking and classification of the traffic light that was implemented in DL. The traffic jam affects business activities as well as other social lives of people living in the urban centers. It also affects the role of social workers and security operatives in saving lives and properties. These issues and other related issues informed the TLD system that is priority-based [13]. It's become essential as numerous government and private organizations gather enormous amounts of domain-specific data, which can offer insightful data on topics like marketing, national intelligence, cyber security, and fraud detection.

Ornek *et al.* (2021) [14] demonstrated the application of the deep neural network (DNN) method, (a DL-based method) in face images in order to classify them into the mask and non-mask classes using the last convolutional layers of DNN. The ResNet-18 DNN was chosen, and the technique was trained on 18,600 balanced facial images from two classes and tested with 4540 non-training face images. When the study was asked why an image was classified into the mask class, it responded by indicating the location where the mask image was present. This was another intelligence-based system showcasing object detection means and DL technique. For counting persons in films, In´acio *et al.* (2021) [15] presented a DL method that paid particular

attention to gender and age information. Counting individuals in movies is an important task with applications in surveillance, commerce, health care, and many other areas. Gender and age were extracted from faces found in films, the system created uses customized Deep Neural Networks (DNN). The study project makes an effort to manage occlusions, prevent double-counting, and lessen the detrimental effects of background information. The hardest part of the work was identifying and tracking faces in angles and handling occlusions.

With their work, Zhang *et al.* (2019) [16] enhanced the functionality of the selective refinement network (SRN) algorithm. The research project combined existing methods, such as a decoupled classification module, to create an SRN face detection. On the benchmark WIDER FACE dataset for face detection, the system performed superbly. Based on the three assessment measures utilized, the resulting approach performed admirably, particularly on the hard subset that contains a sizable number of small faces. This was another landmark achievement of DL-based application, it gave further confidence in to TLD, an innovation that aids smart city development. Also, high-performance face identification, according to [17], [18] was a difficult problem to solve, especially when there were many little faces. The researchers found that the current face detection algorithms prioritize a high recall rate while neglecting the problem of too many false positives. They think there is room for improvement in recall effectiveness and accuracy of location, particularly in terms of lowering the proportion of false positives at high recall rates and optimizing the position of the bounding box. To reduce false positives and increase the accuracy of the location at the same time, they suggested a revolutionary single-shot face detection termed an SRN. To generate a more diverse receptive field and aid in data collection in extreme positions, a model with a receptive field improvement was designed.

Another line of study that was similar was done by [19] was the creation of the repulsion loss for pedestrian detection, which significantly improved detection performance, particularly in circumstances with crowds. The underlying justification for the repulsion loss was that repulsion-by-surrounding can be incredibly helpful and attraction-by-target loss alone may not be sufficient to train an optimal detection. To implement the repulsion energy, they offered two repLoss forms. On the two large datasets, City Persons and Caltech, they have the highest reported performance. Sun *et al.* (2018) [20] offered a new DL-based face detection strategy that achieved an improved detection performance through a well-known Fddb face detection benchmark evaluation. To enhance the Faster Region-based Convolutional Neural Network (RCNN) architecture, they integrated a number of strategies, including meticulous calibration of crucial parameters. The WIDER FACE dataset was initially used to train the RCNN algorithm of CNN. The pre-trained model was later put to the test using the data set in order to provide hard negatives. In the following round of training, the hard negatives were input into the network to produce fewer false positives. The CNN network's backbone, VGG-16, served as the installation and training platform for the DL approach. The encouraging results demonstrate the efficacy of the suggested DL-based system for face detection as an intelligence-based system.

The method of zero watermarking was used by [21] to authenticate medical images. The image is transformed into four bands using the Discrete Wavelet Transform method. Following the discrete wavelet transform (DWT), each block was handled with SVD to remove selected features, and a master share was created by conducting a logical operation between the extracted unique feature and the watermark. The technique produces superior experimental findings and can be used in the field of telemedicine to safeguard shared photos from unauthorized use. It was another breakthrough in the medical field.

Zhou *et al.* (2017) [22] demonstrates the importance of DL technology applications and the influence of dataset for DL technique through the use of the quicker R-CNN on new image datasets of a football game, which identifies four (4) categories of objects, and the corner flag. Experiments indicate that DL technology was an efficient method for moving a human-made feature that relied on the drive of experience to learning that relied on the drive of data. The challenge was in the quality of the data fed, they recommend the use of some synthetic data in order to increase the amount of data in future research. Huge data is the foundation for DL's success, just as large data was the fuel for DL's rocket. A DNN fusion architecture for fast and reliable pedestrian detection was introduced by [23]. For speed, the proposed network fusion algorithm allows many networks to be processed simultaneously. As an object detector, a single shot Deep convolutional neural network (DCNN) was all potential pedestrian candidates through training of various sizes and occlusions. Also, for further improvement of these pedestrian candidates, multiple DNNs were utilized in tandem. Additionally, as a reinforcement to pedestrian detection, they present a method for incorporating a pixel-by-pixel network fusion architecture that incorporates the semantic segmentation network. The technique performs better than most of the existing techniques while also outperformed other in terms of speed.

Angelova *et al.* (2015) [24] designed an intelligence-based system for real-time pedestrian detection architecture using a cascade of DCNN. The tiny model was used to reject a huge DNN was used to classify the hard proposals of a large number of easy negatives were identified. Girshick (2017) [25] introduced an object detection models named the Fast RCNN. The Fast R-CNN improved on the advantages and disadvantaged of Spatial pyramid pooling networks (SPPnet) and Region-based Convolutional R-CNN. It efficiently classifies objects with improved speed and accuracy using DCNNs.

Sermanet *et al.*, (2013) [26] proposed another learning model for pedestrian detection, an area in traffic zones. Unlike common ways where the low-level features were manually designed, the proposed technique learns all features at each and every level of a hierarchy. By combining high- and low-resolution features in the methods and learning features from the input's color channels, they improved the convolutional filter bank method, which they had previously used. Using the INRIA dataset, the method demonstrated that these improvements offer obvious performance advantages. On most measures of all publicly accessible datasets, the generated framework offers a competitive result. Karim *et al.* (2019) [7] suggested Mask RCNN technique for the segmentation and zones in plants, the technique was trained via transfer learning, which started with a neural network (NN) that had been pre-trained using the Microsoft Common Objects in Context (MS-COCO) dataset then fine-tuned it with a small number of annotated photos images. The Mask RCNN technique was then tweaked to produce consistent video detection outcomes, which was accomplished by employing a two-staged detection threshold and analyzing the temporal coherence information of discovered items. The system now includes an object tracking capability for detecting object misplacement. Analyzing a sample of video footages confirmed the usefulness and efficiency of the suggested approach. This has not been employed or experimented using TL dataset.

Ajagbe *et al.* (2021) [27] implemented multi-classification of alzheimer disease using magnetic resonance images (MRI) and DCNN techniques, the research particularly used CNN and transfer learning techniques Visual Geometry Group (VGG-16 and VGG-19). The VGG-19 outperformed other techniques used in the study based on the results obtained from the four (4) different evaluation metrics in the research. It is observed from the reviewed literature that there are limited studies that accounted for the applications of DL in object detection with specific attention to TLD. Studies that outlined that of DL-based models that is efficient in TLD are rarely found in scholarly databases. Research investigating TDL based on CNN two-stage detection methods is limited to the best of researchers' knowledge and based on the requirement and contribution of this method. Hence, this study was informed.

### 3. METHOD

This section reports our DL experimental approach towards TLD on Representative two-stage detectors, a DL framework. Representative two-stage detection technique comprising three CNN methods, namely; FASTER-CNN (FASTER-CNN), FAST-CNN (FAST-CNN), R-CNN, Region-based fully convolutional network (R-FCN) and Spatial Pyramid Pooling networks (SPPNet). Basically, the experimental approach in this chapter entails: 1) TLD data collection, 2) pre-processing of the dataset, 3) Data partitioning, 4) Feature section 5) DL-based TLD Analysis, and 6) Evaluation. Figure 1 presents the architecture for this study named DL-based (two-stage detection methods) analysis using LaRA Traffic Lights Recognition. Algorithm 1 was proposed for the investigation of two-stage detectors in a TLD environment and thus presented.

---

#### ***Algorithm 1: Investigation of Two-stage Detectors in TLD Algorithm***

---

***Input:*** LaRA TL dataset; Training and testing, testing: 0

---

***Output:*** TLD result with forecast and accomplishment, DA, F1-score, precision, recall and RT in TLD.

---

```

1. Dataset = LaRA TL dataset
2. Train = Train FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet using LaRA dataset
3. analyzeNetwork (FaceNet)
4. TrainingOption:
    OptimizationAlgorithm = rmsprop
    InitialLearningRate = 0.00002
    MaxEpochs = 25
    MiniBatchSize = 32,
5. Load Dataset
6. Resize Dataset to [227,227, 3]
7. Dataset Split [TrainingData (80%), TestingData (20%)]
8. Train Load Dataset
9. Test Trained models/* using TestingData */
10. Return TestResults
11. TestResults = Validation DA, F1-score, precision and recall
15. If TestResults = Satisfactory, then
    (1) Save the Trained models and TestResults /* for transfer
        learning purpose */
    (2) End
16. Else
17. Adjust the TrainingOption, then
18. Repeat the process

```

---

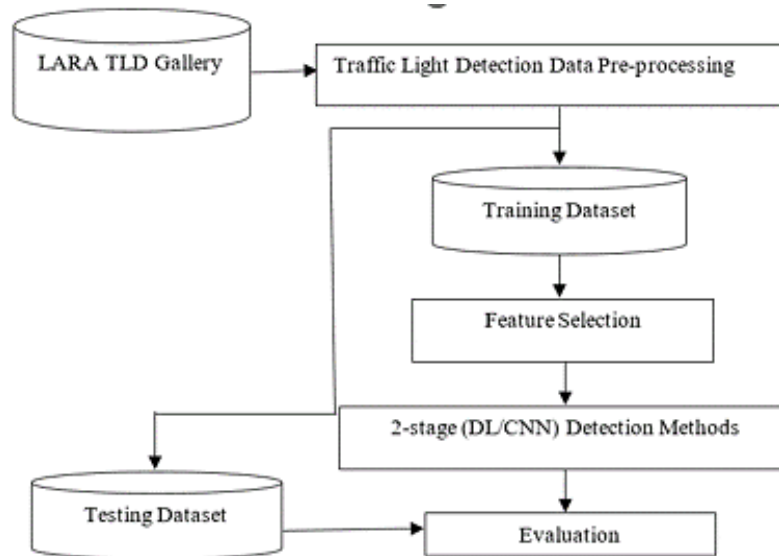


Figure 1. Architecture of two-stage detection technique based on TLD

### 3.1. Data collection

The pre-trained dataset used for a benchmark of the experiment was acquired online and used by many scholar including [28]. The LaRA traffic lights recognition public benchmarks [28] video data was collected. There are 11,179 frames in total, with a resolution of 640\*480. This dataset includes a sequence file and a ground truth file containing 9,168 instances of traffic lights classified into four categories (ambiguous, green, red, and orange). Labelled traffic lights in LaRA are 5 pixels larger.

#### 3.1.1. Pre-processing of TLD dataset

The TLD dataset was automatically resized, cropped and unwanted noises were removed from the background of the data. The data pre-processing was done using the instrumentality of the python programming environment. The data were also flattened, and converted to greyscale. This was done to gain better and good insight of the dataset for analysis and better performance of the models.

#### 3.1.2. Selection of features

After the preprocessing of the TLD dataset that terminated at the conversion of the images to grayscale images, suitable features were selected for the experiment. The feature vectors were partitioned into a 20: 80 percent test-train split ratio of the datasets. This was done to gain better and good insight of the dataset for analysis and better performance of the models. All the irrelevant and unwanted features were removed.

#### 3.1.3. Representative two-stage detectors

The generic object detector is a CNN framework that uses for object mainstreaming using a DL-based approach, they are typically divided into two categories; one-stage and two-stage methods. While FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet are examples of two-stage detectors, you only look once (YOLO), Single-Shot object Detection (SSD), YOLOv2 and YOLOv3 are popular examples of one-stage. Detection in two-stage method is based on region proposals. Proposals in R-CNN and Fast R-CNN were generated from images selected from the original image, whereas proposals in Faster R-CNN are also generated directly from feature maps. Classification and bounding box regression was applied in two stages to each region proposal to achieve Faster R-CNN. In this section, the typical two-stage detection series are introduced in detail, focusing on network structure and performance. Other representative two-stage detectors that will be discussed and used include spatial pyramid pooling [29] and region-based fully convolutional network [30].

**R-CNN:** Girshick *et al.* (2014) [31] developed R-CNN as a breakthrough study in the development of applying DL-based methods to detection. In contrast to traditional methods that use a sliding window, R-CNN obtains proposals using the selective search (SS) technique [32]. CNN and support vector machine (SVM) were responsible for feature extraction and classification, respectively. Such a significant improvement demonstrated the significant benefit of the region proposal with CNN. Meanwhile, this ground-breaking framework was not without flaws [25] (1) multi-step training with time-consuming steps; and (2) slow object detection. SPPNet provided a good solution to the second problem.

**SPPNet:** In R-CNN, there were 2k proposals, and each proposal was fed to a CNN and an SVM classifier separately. The feature extraction and detection were repeated 2,000 times for each image, which was the main cause of the slow detection. 2,000 proposed were also generated in SPPNet [30]. Meanwhile, CNN was used to extract the entire feature map of the original image, and the feature map of each proposal is generated by mapping. Consequently, feature maps are only computed once. Although the input length for the FC layer is fixed, the feature maps for the various approaches are not all the same size. An SPP [33] layer was added before the FC layers. In conclusion, using SPPNet greatly accelerates object detection when compared to R-CNN, but issues such as multi-stage training remain [25]. Furthermore, when fine-tuning the network, the CONV layers are fixed, and only the FC layer is fine-tuned, on the other hand, very deep networks' expressiveness was constrained. Fast R-CNN, was an efficient and quick object detection method that Girshick (2015) [25] presented to address these two issues.

**Fast R-CNN:** The feature extraction sections of various proposals are also shared. Pooling by RoI was used to generate proposals of the same size. As previously stated, the training phase of SPPNet and R-CNN are a multi-stage pipeline. Softmax classifier is used instead of SVM to perform classification in Fast R-CNN. Furthermore, the loss function, which includes regression task loss, makes the entire training process end-to-end. To summarize, Fast R-CNN simplifies the training and testing of the entire network. However, SS for region extraction is time-consuming, accounting for a significant portion of the running time, and cannot meet the requirements of a real-time application. Because SS was used to extract region proposals first, in fact, Fast R-CNN does not achieve true end-to-end [34].

**Faster R-CNN:** It requires only 0.32 s to feed-forward an image; the bottleneck is that generating proposals takes 2 s. Ren *et al.* (2015) [35] proposed a faster R-CNN that replaced SS with RPN. In comparison to SS, which slides on the original image, RPN employs a similar sliding policy but on smaller feature maps, resulting in fewer proposals. End-to-end training was achieved by combining RPN with the entire network, and the test time for Faster R-CNN was 0.2 s, which was close to real-time.

**R-FCN:** After RoI pooling, quicker for each proposal, R-CNN must carry out two-branch prediction independently. Long *et al.* 2015 [36] designed R-FCN to allow almost all computing to be shared in order to increase speed in the network. Following the shared CONV layers, the position-sensitive score maps are generated using one more CONV layer. When performing RoI pooling, each proposal was divided into 3 \* 3 grids, and R-FCN assigns each grid to a different channel of the score maps. To generate a final feature map, average pooling is applied to each grid.

In summary, R-CNN, SPPNet, Fast R-CNN, Faster R-CNN, and R-FCN methods evolving increase the amount of image-level calculations in a network over time while decreasing the proportion of regional-level calculations. Almost all R-CNN calculations were at the regional level, whereas almost all R-FCN calculations were at the image level. Therefore, this research is extending the frontier of research on the DL-based CNN two-stage detection methods to be studied using the TLD dataset.

### 3.1.4. Implementation environment

The implementation environment for DL-based analysis of the TLD was carried out on the Anaconda platform using a Jupyter notebook and the sklearn libraries [37], [38]. All experiments were carried out on an Intel® core™ i5-7200 Pentium Windows computer with 8GB RAM and an Intel® core™ i5-7200 CPU running at 2.50 GHz to 2.70 GHz. In this section, an analysis of DL works in TLD was implemented according to Representative two-stage detection methods. Unlike the one-stage detection methods, these methods are advanced and evolving relatively used in object detection projects. It also allows direct use of the framework with or without changes.

### 3.1.5. Evaluation method

The analysis of TLD in this chapter was based on five metrics, they are detection accuracy (DA), F1-score, precision and recall. The five (CNN) two-stage detection methods used for TLD analysis were; FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet are based on the five metrics. The DL-based TLD analysis in this research is experimentally based, the four (4) distinct state-of-the-art evaluation metrics have been selected to examine the strength and weaknesses of these CNN methods techniques. The metrics of evaluation in the two-stage detection DL-based methods for TLD investigation presents in this sub-section:

$$\text{Detection Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (1)$$

$$\text{Precision} = \frac{Tp}{Tp+Fp} \quad (2)$$

$$\text{Recall} = \frac{Tp}{Tp+Fn} \quad (3)$$

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

#### 4. RESULTS AND DISCUSSION

After a successful experimental investigation of two-stage detectors models LaRA dataset (traffic light dataset). The results and discussion of the experimental analysis of TLD in this study are presented in this section. The section presents the detail analysis of the five standards metrics for the evaluation of TLD in this study viz-a-viz; detection accuracy (DA), F1-score, precision and recall. The DL-based two-stage detection methods evaluated and discussed are FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet.

##### 4.1. Result

The result of the evaluation was based on the state-of-the-art performance metric in DL and AI generally, the metrics are detection accuracy (DA), precision, F1-score and recall [39], all the experimental parameters were kept on default for the five methods used. The suitable and pre-trained dataset meant for TLD from LaRA was used [28]. As discussed earlier in Sections 3, the comparison of the five different two-stage TLD detection algorithms' outcomes on the LaRA dataset were presented. The values for all other performance metrics range from between 0 and 1. The closer the value of the metrics to 1, the better the method. Table 1 summarizes the Performance Analysis of the two-stage detection methods on the TLD Dataset in this research.

Table 1. Performance Analysis of the two-stage detection methods on TLD Dataset

Methods	R-CNN	SPPNet	Fast-CNN	Faster-CNN	R-FCN
Detection Accuracy (DA)	0.86	0.73	0.88	0.89	0.79
F1-Score	0.88	0.84	0.89	0.93	0.85
Precision	0.79	0.75	0.83	0.83	0.79
Recall	0.81	0.75	0.84	0.90	0.81

##### 4.2. Discussion

###### 4.2.1. Detection accuracy

Figure 2 presents the detection accuracy (DA) of the two-stage detection methods using the TLD dataset in this research; FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet with 0.89, 0.88, 0.86, 0.79 and 0.73 respectively. Since the method that has a rate closer to 1 is the best. Hence, FASTER-CNN is better than the others.

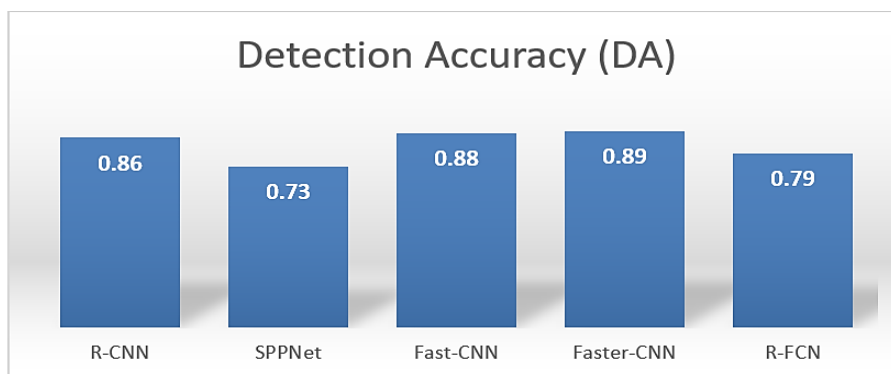


Figure 2. Two-stage TLD detection accuracy

###### 4.2.2. F1-score

Figure 3 presents the F1-score of the two-stage detection methods using the TLD dataset in this research; FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet with 0.93, 0.89, 0.88, 0.85 and 0.84 respectively. Since the method that has a rate closer to 1 is the best. Hence, FASTER-CNN is better than the others.

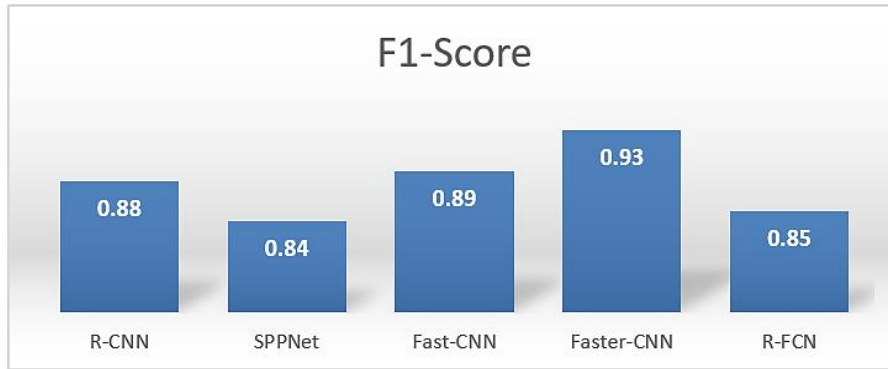


Figure 3. Two-stage TLD F1-score

#### 4.2.3. Precision

Figure 4 presents the precision of the two-stage detection methods using the TLD dataset in this research; FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet with 0.83, 0.83, 0.79, 0.79 and 0.75 respectively. Since the method that has a rate closer to 1 is the best. Hence, FASTER-CNN and FAST-CNN are better than the others.

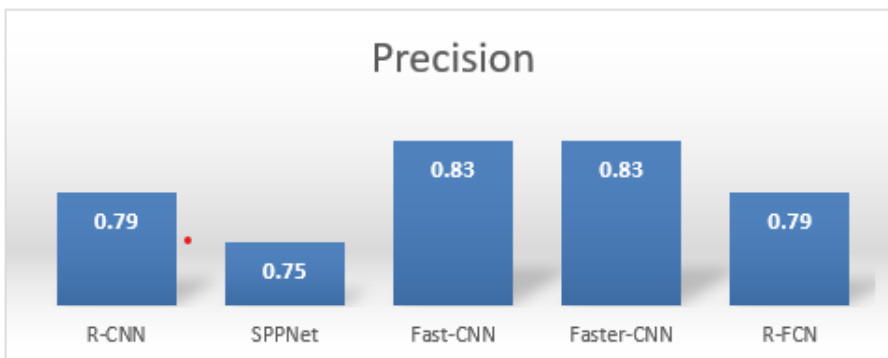


Figure 4. Two-stage TLD precision

#### 4.2.4. Recall

Figure 5 presents the recall report of the two-stage detection methods using the TLD dataset in this research; FASTER-CNN, FAST-CNN, R-CNN, R-FCN and SPPNet with 0.9, 0.84, 0.81, 0.81 and 0.75 respectively. Since the method that has a rate closer to 1 is the best. Hence, FASTER-CNN is better than the others.

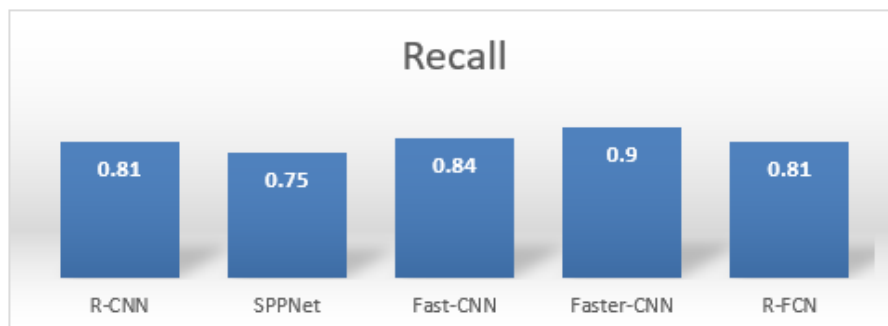


Figure 5. Two-stage TLD recall



Summarily, among the four standard and state-of-the-art performance metrics used for the evaluation of the five distinct two-stage detection methods that have been judged to be more popular in object detection and tracking, FASTER-CNN performed best in all. This was followed by FAST-CNN which was second in four metrics and alongside FASTER-CNN, FAST-CNNout performed others techniques in precision.

## 5. CONCLUSION

The DL-based method is a hot research field, its roles in object detection particularly TLD booming due to its cutting-edge solutions in autonomous driving and implementation of the smart city. The powerful CNN is one of the DL-based methods that made DL-based research more popular. This chapter has exhaustively reviewed object detection applications viz-a-viz architecture, techniques, and DL-CNN in TLD. We further implemented object detection using DL/CNN two-stage detection methods in a TLD environment. FASTER-CNN was the best method based on the standard evaluation metrics used. This chapter addresses some of the concerns in TLD using DL methods which is one of the popular AI methods in TLD, we also classified detection according to different applications works such as pedestrian detection, lane detection and so forth. To the extent that we are aware, this is the first experimental study that analyzed DL-based models focusing on two-stage detection in the TLD environment. Considering the performance improvement of two-stage detection methods in this study, this will also bring advantages of DL/CNN in TLD to both researchers and users.

## ACKNOWLEDGEMENTS

Authors acknowledge the support of the Cape Peninsula University of Technology, South Africa and the First Technical University, Ibadan, Nigeria for their support on this project.





## REFERENCES

- [1] S. Ojha and S. Sakhare, "Image processing techniques for object tracking in video surveillance- A survey," Jan. 2015, doi: 10.1109/PERVASIVE.2015.7087180.
- [2] R. Verschae and J. Ruiz-del-Solar, "Object detection: Current and future directions," *Frontiers Robotics AI*, vol. 2, no. NOV, Nov. 2015, doi: 10.3389/frobt.2015.00029.
- [3] S. A. Ajagbe and I. Bamimore, "Design and implementation of smart home nodes for security using radio frequency modules," *International Journal of Digital Signals and Smart Systems*, vol. 4, no. 4, p. 286, 2020, doi: 10.1504/ijdsss.2020.10032471.
- [4] T. P. Olalere and O. D. Adeniji, "An artificial intelligent video assistant invigilator to curb examination malpractice," 2021.
- [5] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *ACM International Conference Proceeding Series*, 2006, vol. 148, pp. 233–240. doi: 10.1145/1143844.1143874.
- [6] M. Enzweiler and D. M. Gavrilu, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009, doi: 10.1109/TPAMI.2008.260.
- [7] M. M. Karim, D. Doell, R. Lingard, Z. Yin, M. C. Leu, and R. Qin, "A region-based deep learning algorithm for detecting and tracking objects in manufacturing plants," *Procedia Manufacturing*, vol. 39, pp. 168–177, 2019, doi: 10.1016/j.promfg.2020.01.289.
- [8] L. C. Possatti *et al.*, "Traffic light recognition using deep learning and prior maps for autonomous cars," in *Proceedings of the International Joint Conference on Neural Networks*, Jul. 2019, vol. 2019-July. doi: 10.1109/IJCNN.2019.8851927.
- [9] M. B. Jensen, M. P. Philipsen, A. Møgelmoose, T. B. Moeslund, and M. M. Trivedi, "Vision for looking at traffic lights: issues, survey, and perspectives," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 1800–1815, Jul. 2016, doi: 10.1109/TITS.2015.2509509.
- [10] V. Ramanishka, Y. T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: a dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7699–7707. doi: 10.1109/CVPR.2018.00803.
- [11] A. Salari, A. Djavadifar, X. Liu, and H. Najjaran, "Object recognition datasets and challenges: A review," *Neurocomputing*, vol. 495, pp. 129–152, Jul. 2022, doi: 10.1016/j.neucom.2022.01.022.
- [12] Y. Li *et al.*, "A deep learning-based hybrid framework for object detection and recognition in autonomous driving," *IEEE Access*, vol. 8, pp. 194228–194239, 2020, doi: 10.1109/ACCESS.2020.3033289.
- [13] A. Mousavian, D. Anguelov, J. Košecká, and J. Flynn, "3D bounding box estimation using deep learning and geometry," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Jul. 2017, vol. 2017-January, pp. 5632–5640. doi: 10.1109/CVPR.2017.597.
- [14] A. H. Ornek, M. Celik, H. T. R. Center, and M. Ceylan, "Explainable artificial intelligence: How face masks are detected via deep neural networks," *International Journal of Innovative Science and Research Technology*, vol. 6, no. 9, pp. 1104–1112, 2021.
- [15] A. de S. Inácio, R. H. Ramos, and H. S. Lopes, "Deep learning for people counting in videos by age and gender," in *Anais do 15. Congresso Brasileiro de Inteligência Computacional*, Jan. 2021, pp. 1–6. doi: 10.21528/cbic2021-53.
- [16] S. Zhang *et al.*, "Improved selective refinement network for face detection," Jan. 2019, Accessed: May 24, 2023. [Online]. Available: <http://arxiv.org/abs/1901.06651>
- [17] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, vol. 33, no. 01, pp. 8231–8238, Jul. 2019, doi: 10.1609/aaai.v33i01.33018231.
- [18] J. B. Awotunde *et al.*, "An improved machine learnings diagnosis technique for COVID-19 pandemic using chest x-ray images," in *Computer and Information Science*, vol. 1455 CCIS, Springer International Publishing, 2021, pp. 319–330. doi: 10.1007/978-3-030-89654-6\_23.




- [19] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7774–7783. doi: 10.1109/CVPR.2018.00811.
- [20] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018, doi: 10.1016/j.neucom.2018.03.030.
- [21] S. Sinha, A. Singh, R. Gupta, and S. Singh, "Authentication and tamper detection in tele-medicine using zero watermarking," *Procedia Computer Science*, vol. 132, pp. 557–562, 2018, doi: 10.1016/j.procs.2018.05.009.
- [22] X. Zhou, W. Gong, W. Fu, and F. Du, "Application of deep learning in object detection," in *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, May 2017, pp. 631–634. doi: 10.1109/ICIS.2017.7960069.
- [23] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, Mar. 2017, pp. 953–961. doi: 10.1109/WACV.2017.111.
- [24] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proceedings of the British Machine Vision Conference 2015*, 2015, pp. 32.1-32.12. doi: 10.5244/c.29.32.
- [25] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2015, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.
- [26] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3626–3633. doi: 10.1109/CVPR.2013.465.
- [27] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, O. F. AFE, and K. I. Okesola, "Multi-classification of alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches," *International Journal of Advanced Computer Research*, vol. 11, no. 53, pp. 51–60, Mar. 2021, doi: 10.19101/ijacr.2021.1152001.
- [28] Q. Wang, Q. Zhang, X. Liang, Y. Wang, C. Zhou, and V. I. Mikulovich, "Traffic lights detection and recognition method based on the improved yolov4 algorithm," *Sensors*, vol. 22, no. 1, p. 200, Dec. 2022, doi: 10.3390/s22010200.
- [29] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Jul. 2017, vol. 2017-Janua, pp. 6517–6525. doi: 10.1109/CVPR.2017.690.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.
- [31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.
- [32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.
- [33] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2169–2178. doi: 10.1109/CVPR.2006.68.
- [34] A. Boukerche and Z. Hou, "Object detection using deep learning methods in traffic scenarios," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–35, Mar. 2021, doi: 10.1145/3434398.
- [35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, vol. 07-12-June-2015, pp. 431–440. doi: 10.1109/CVPR.2015.7298965.
- [37] R. Rajmohan *et al.*, "G-Sep: A deep learning algorithm for detection of long-term sepsis using bidirectional gated recurrent unit," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 30, no. Supp01, pp. 1–29, May 2022, doi: 10.1142/S0218488522400013.
- [38] S. Devadharshini, R. KalaiPriya, R. Rajmohan, M. Pavithra, and T. Ananthkumar, "Performance investigation of hybrid YOLO-VGG16 based ship detection framework using SAR images," Jul. 2020. doi: 10.1109/ICSCAN49426.2020.9262440.
- [39] S. Patel and A. Patel, "Deep leaning architectures and its applications a survey," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 6, pp. 1177–1183, Jun. 2018, doi: 10.26438/ijcse/v6i6.11771183.

## BIOGRAPHIES OF AUTHORS






**Sunday Adeola Ajagbe**     is a Ph.D. candidate at the Department of Computer Engineering, Ladoko Akintola University of Technology (LAUTECH), Ogbomosho, Nigeria and a Lecturer, a First Technical University, Ibadan, Nigeria. He obtained MSc and BSc in Information Technology and Communication Technology respectively at the National Open University of Nigeria (NOUN). His specialization includes Artificial Intelligence (AI), Natural language processing (NLP), Information Security, Communication, and Internet of Things (IoT). He has many publications to his credit in reputable academic databases. He can be contacted at email: Sunday.ajagbe@tech-u.edu.ng






**Adekanmi Adeyinka Adegun**    received the B.Tech., M.Sc., and Ph.D. degrees in computer science. He has close to ten years lecturing experience in Universities. He has also co-supervised M.Sc. and Ph.D. candidates in ML fields. He has published extensively in several artificial intelligence and computer vision-related accredited journals and international and national conference proceedings. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and NLP. He currently serves as a reviewer for some machine learning and computer vision-related journals. He can be contacted at email: [adegunadekanmi@gmail.com](mailto:adegunadekanmi@gmail.com)






**Ahmed Babajide Olanrewaju**    is a System Analyst at University of Ibadan, his interest in the area of NLP with special focus on social media, women access to health services. He has been the lead organiser of the AI community in Ibadan which provides opportunity for women and youth through organising local content resources to aid AI-based learning and research. He can be contacted at email: [a.olanrewaju@ui.edu.ng](mailto:a.olanrewaju@ui.edu.ng)



**Prof. John Babalola Oladosu**    is a Professor of Computer Engineering and the current Head of the Department, Computer Engineering, LAUTECH, Ogbomoso, Nigeria. He is licensed by The Council Regulating Engineering in Nigeria (COREN) as a professional Computer Engineer and a member of the International Association of Engineers (IAENG). He has over 20 years of Teaching and Research experience in the university system. He can be contacted at email: [jboladosu@lautech.edu.ng](mailto:jboladosu@lautech.edu.ng)



**Prof. Matthew Olusegun Adigun**    retired in 2020 as a Senior Professor of Computer Science at the University of Zululand. He obtained his doctorate degree in 1989 from Obafemi Awolowo University, Nigeria; having previously received both Masters in Computer Science (1984) and a Combined Honours degree in Computer Science and Economics (1979) from the same University (when it was known as University of Ife, Nigeria). A very active researcher in Software Engineering of the Wireless Internet, he has published widely in the specialised areas of reusability, software product lines, and the engineering of on-demand grid computing-based applications in Mobile Computing, Mobile Internet and ad hoc Mobile Clouds. Recently, his interest in the Wireless Internet has extended to Wireless. He has received both research and teaching recognitions for raising the flag of Excellence in Historically Disadvantaged South African Universities as well as being awarded a 2020 SAICSIT Pioneer of the Year in the Computing Discipline. Currently, he works as a Temporary Senior Professor at the University of Zululand to pursue his recent interest in AI-enabled Pandemic Response and Preparedness. He can be contacted at email: [profmatthewo@gmail.com](mailto:profmatthewo@gmail.com)