In the 21st century, the world is experiencing a significant population explosion, which contributes to significant increase in crime rates. The rise of crime demands a continuous observation and investigation. In such case CCTV and IP cameras has provide necessary surveillance; however, when it comes to large volumes of footages we experience significant challenges. A crime scene investigation demands a manual cataloguing of evidences by an expert investigator. This helps in rightful identification of the victim as well as the evidence.

Historically the images and videos collected from the crime scene is crucial for forensic investigation to retrieve key evidences. The modern world shift to digitisation has helped transform the traditional forensic investigation. The advent of artificial intelligence further has helped revolutionize the digital forensic as well as data collection.

This review of literature explores existing digital forensic investigation and data collection models. This chapter aims to identify gaps in current models and highlight the need for improvement and innovation. It focuses to present current findings to discuss the shortcoming and scope of improvement for the new proposed model. Deep Neural Networks (DNNs) has certainly shows a significant contribution in the improving image classification.

Erhan et al. (2014) while working in object localization in images using DNN-based object mask regression focuses on specific objects such as fire arms and knives.

O'Reilly et al. (2012) in their research use signal processing techniques for concealed weapon detection with the help of neural networks, displaying amazing results with the multimeter wave radar effectively detects concealed weapons.

He, Zhang, and others (2016) concluded that automatic feature representations, unsupervised approaches, multiple instance learning significantly surpass manual feature representation, supervised approaches and supervised learning performance.

Gerga et al. (2016) proposed an algorithm to detect firearms and knives using OpenCV. This proposed algorithm reduces weapon detection efficacy to 35% minimizing false alarms, sacrifice sensitivity as it is vital to not miss even single weapon detection in real world.

Pandey et al. (2016) found that mobile devices, low cost cameras, AI and Machine Learning have enhanced forensic analysis. These have also pioneered research focused on video validation to detect forgery attempts.

Kamenicky et al. (2016) explored and introduced various video analysis methods and tools, addressing the challenge of source of images and videos in criminal investigations.

Amerini et al. (2017) introduced a method to extract composite fingerprint from cell phone cameras using PRNU to identify source from videos shared on social media.

Horsman (2018) proposed a forensic method for detecting and reconstructing cached online video stream data from platforms such as YouTube, Twitter, and Facebook.

Senan (2017) highlighted the issue of over-enhancement in CCTV footage using HE-based methods, which can result in unnatural and washed-out appearances, particularly in low-light conditions with narrow dynamic ranges.

Ayyavoo and Suseela (2018) introduced a color video enhancement technique using Discrete Wavelet Transform and CLAHE, named 'DWT E-CLANE,' which improves facial image recognition.

Hendrawan and Asmiatun (2018) discussed the effectiveness of CLAHE in overcoming the over-enhancement problems associated with HE, noting that various CLAHE variants have demonstrated effectiveness in specific scenarios.

He et al *(2014).* proposed SPP-net based on Spatial Pyramid pooling that improved the detection and classification time by pooling region features instead of sending each region into the CNN.

Ren *et al (2015)* introduced Faster R-CNN, a faster version of Fast R-CNN, which replaces the previous region proposal method with RPN (Region proposal Network), which simultaneously predicts object bounds and scores.

Girshick et al (2014) proposed a R-CNN algorithm, which was one of the first real target detection model based on convolutional neural networks. The improved R-CNN model gave 66% mAP score making the initial Selective Search to extract approximately 2000 region proposals of each image. Finally, a linear regression model is trained to perform the regression operation of the bounding box for each extracted image that is processed into SVM classifier. Resulting in improved R-CNN. As a limitation this model had large computation and low efficacy. Also, directly scaling the region proposal to a fixed-length feature vector may cause object distortion.

He, K.M , et al (2015), The ineffectiveness and poor detection issues are resolved by the Spatial Pyramid Pooling (SPP) model. This eliminates the requirement for R-CNN picture blocks with fixed input sizes. The suggested method extracts the features map and only needs to do the convolution computations once. To extract the feature vector of a fixed size, the spatial pyramid pooling layer is added to the final convolutional layer and passed through. In contrast to the R-CNN, Spp-Net only needs to execute feature extraction once, saving on several computations. It still has the same drawbacks as R-CNN, though: 1) Training steps with several steps are complex. 2) More regressors are needed, and separate SVM classifiers must be trained.

Man Ro et al. (2021) introduce a novel method for object classification and localization that employs attentive layer separation. Using ResNet-101 as the backbone network, their approach separates less semantic and semantic layers to handle object classification and localization independently. This technique enhances the effectiveness of object detection by leveraging distinct semantic layers for improved accuracy.

Yu et al. (2021) develop an object detection algorithm that utilizes Region-based Convolutional Neural Networks (RCNN) combined with selective search. Their approach involves extracting around two thousand candidate regions from the initial image, normalizing these regions, and applying Support Vector Machines (SVM) for feature extraction. Non-Maximum Suppression (NMS) is used to select the highest scoring regions, refining the detection process.

Kanimozhi et al. (2021) proposed a lightweight object detection network using MobileNet and Single Shot MultiBox Detector (SSD), termed MobileDet. Their model demonstrated a detection time of 3-5 seconds. In contrast the accuracy decreases when objects are distant than 30 meters from camera. This study is highlights the comparison aspect between detection speed and accuracy at varying distances

Zahisham et al. (2021) resized the images to 224x224 pixels and applied convolutional filters for feature extraction. By fine-tuning a pre-trained ResNet-50 model on various food datasets, including ETHZ-FOOD101, UECFOOD100, and UECFOOD256, they achieve superior performance compared to existing methods, demonstrating fast training and high accuracy.

Deepa et al. (2021) trained their model for real-time tennis ball tracking using YOLO, SSD, and Faster RCNN. The dataset had images captured from multiple angles and lighting conditions. In their research they found SSD providing minimum detecting time with high efficacy.

Garg et al. (2021) Using a 448x448 pixel input image they trained a model for face detection on a YOLO-based architecture. The model predicts bounding box coordinates and class probabilities using Non-Maximum Suppression (NMS). The outcome of the study showed consistent accuracy of 92.2% and improved frames per second (FPS) with lower resolution images, proving YOLO's effectiveness for face detection.

Zhang et al. (2021) using TensorFlow and OpenCV, and training on the WIDER FACE dataset presented a Multi-task Cascaded Convolutional Networks (MTCNN) approach for face detection. As a result, an average precision (mAP) of 85.7% was achieved when compared MTCNN with YOLOv3. MTCNN's performance was superior in detecting multiple faces, especially in complex scenes.

Oumina et al. (2021) uses pre-trained deep learning models such as MobileNetV2, VGG19, and Xception to detect face masks. The method of combining models with classifiers like Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN), they achieve a 97.1% presents high classified performance for face mask detection. They combine these accuracies with a small dataset.

Girshick, (2010) advanced the field with the Deformable Part-Based Model (DPM), extending HOG by using a part-based approach to object detection, which decomposes objects into parts for detection. Girshick later enhanced DPM with mixture models to address real-world variations, influencing subsequent object detectors.

Girshick et al., (2012) The field underwent a major transformation post-2010 with the resurgence of convolutional neural networks (CNNs), which provided robust feature representations and marked a new era in object detection.

Liu et al., (2015) introduced the Single Shot MultiBox Detector (SSD), which improved detection accuracy and speed by employing multi-reference and multi-resolution techniques. SSD detects objects of varying scales on different network layers, significantly enhancing its performance for small objects and achieving a COCO mAP@.5 of 46.5% with a fast version running at 59fps.

Lin et al., (2017) proposed RetinaNet, addressing the accuracy limitations of one-stage detectors by introducing "focal loss," which reshapes the standard cross-entropy loss to focus on hard-to-classify examples, achieving a COCO mAP@.5 of 59.1%.

Law et al., (2018), CornerNet, developed by Law et al. in 2018, shifted the paradigm from anchor boxes to key point prediction, improving performance by predicting object corners and forming bounding boxes from them. This approach resulted in a COCO mAP@.5 of 57.8%.

Zhou et al., (2019) put forth CenterNet achieving a COCO mAP@.5 of 61.1%. It is a key point-based detection model to simplify the detection process. It focuses on object centers and integrates multiple tasks into a single framework.

Carion et al., (2020) introduced DETR, that uses Transformers for object detection. This eliminated the need for anchor boxes and a new level of performance was achieved.

Zhu et al., (2021) proposed Deformable DETR to address DETR's time and performance issues in detection of small objects. This model achieved a COCO mAP@.5 of 71.9%.

Negi et al. (2021) proposed a simplified a Neural network using Keras-Surgeon for efficient face mask detection. The study signifies the efficacy and improved performance of pruning while reducing complexity.

Girshick, (2015) introduced the Fast R-CNN model, which significantly improved upon the original R-CNN by enhancing detection speed and accuracy. On the joint VOC2007 and VOC2012 dataset. The Fast R-CNN achieved a mean Average

Precision (mAP) of 70.0%. The model incorporates three key innovations: (1) it replaces the Support Vector Machine (SVM) used in R-CNN with a softmax function for classification, (2) it introduces the Region of Interest (RoI) pooling layer, derived from the pyramid pooling layer in SPP-Net, to convert candidate box features into a fixed-size feature map suitable for the fully connected layer, and (3) it employs two parallel fully connected layers instead of a single softmax classification layer. Despite these advancements, Fast R-CNN does not fully meet the requirements for real-time detection.

Ren et al., (2016) proposed the Faster R-CNN model, which further advances object detection by replacing the Selective Search method with Region Proposal Networks (RPN) to generate region proposals. The model comprises two main modules: (1) a fully convolutional neural network that generates region proposals, and (2) the Fast R-CNN detection algorithm. These modules share a set of convolutional layers, with the input image passing through the CNN to reach the final shared convolutional layer. This setup allows the network to generate feature maps for both the RPN and the Fast R-CNN detection algorithm. While Faster R-CNN excels in detection accuracy, it still falls short of achieving real-time detection capabilities.

LIDAR, (2023) Object detection has evolved significantly with various technologies and methodologies introduced over the years. Early systems employed LIDAR sensors for vehicle detection and non-intrusive methods like Adaptive Spatial Feature Fusion (ASFF) and Radar Sensors (ASFF, 2023; Radar Sensors, 2023).

Phillips (2021) developed a vehicle distance estimation system using monocular cameras, though accuracy declined with distance. Wang (2021) demonstrated edge detection for vehicle identification, including use in drones. Sokalski (2021) combined edge detection with color identification, while Kanistras (2021) proposed an edge detection method using angle vectors.

Xiao and Kang (2021) emphasized the importance of diverse datasets for training algorithms, and Zoph (2021) highlighted data augmentation strategies to improve accuracy.

Lin (2021) created a YOLO-based traffic counting system, achieving 95% accuracy during the day with improvements for night detection. Tao (2021) optimized YOLO with a pooling layer and pre-processing for night images, reaching 80.1% accuracy on custom datasets. Corovic (2021) demonstrated YOLO's effectiveness in real-time detection, though occluded objects reduced accuracy. Salarpour (2021) employed Kalman filters and background subtraction for multi-vehicle tracking, achieving 96% accuracy. Phan (2021) introduced an occlusion reduction method with background subtraction, improving detection accuracy to 85% in high traffic. Lu (2021) modified the Region Proposal Network (RPN) to address scale variability, achieving precision scores of 64.1% and 84.8% for different scales.

Redmon et al., (2016) introduced the YOLOv1 object detection model, which marked a significant departure from previous methods by eliminating the need for region proposal extraction. Instead, YOLOv1 utilizes a simple convolutional neural network (CNN) structure. The model processes the entire image as input and directly predicts bounding box locations and categories at the output layer. Specifically, it divides an image into an S × S grid, where each grid cell predicts B bounding boxes along with their confidence scores. Each cell predicts a total of B*(4+1) values. YOLOv1 achieves real-time detection with a speed of 45 frames per second (fps) on a Titan X GPU. Despite its fast processing, YOLOv1 has been noted for producing fewer background errors but struggles with recognizing objects in groups.

Redmon et al., (2016) In 2016, Redmon proposed YOLOv2 to enhance both recall and localization while maintaining classification accuracy. YOLOv2 integrates several improvements, including the use of a new feature extraction network, Darknet-19, which consists of 19 convolutional layers and 5 max pooling layers. The model incorporates batch normalization, removes dropout, introduces an anchor box mechanism, and employs k-means clustering on training set bounding boxes. These modifications significantly boost recall and accuracy. However, challenges remain in detecting highly overlapping and small targets.

Redmon & Farhadi, (2018) developed YOLOv3, which is noted for its balanced performance in terms of detection speed and accuracy. YOLOv3 introduces multi-label classification by replacing the original softmax layer with a logistic regression layer for multi-label classification. The model uses a multi-scale prediction approach with upsampling and fusion similar to Feature Pyramid Networks (FPN). YOLOv3 employs a deeper feature extraction network, Darknet-53. Although YOLOv3 improves the detection of small targets and overall speed, it does not significantly enhance detection accuracy, particularly when Intersection over Union (IoU) exceeds 0.5.

Brindha et al. (2021) propose an enhancement to the YOLOv3 algorithm by integrating edge detection for boundary box construction. Their method involves scaling the input image to 416x416 pixels, applying CNN processing, and utilizing edge detection techniques to construct bounding boxes based on threshold values. This approach aims to improve the precision of object detection.

Bhuiyan et al. (2021) apply YOLOv3 for mask detection, using a dataset of 600 images with annotations for mask-wearing and non-mask-wearing individuals. Their method detects bounding box coordinates and determines mask presence, contributing to improved face mask detection accuracy.

Liu and Zhang (2021) improved the YOLOv3 model for traffic conditions, achieving a Mean Average Precision (mAP) of 91.12% with their F-YOLOv3 algorithm, surpassing Faster R-CNN and YOLOv3. Redmon (2018) integrated classification and localization into a single convolutional neural network, though it struggled with smaller objects and varying aspect ratios.

Chandan (2021) used OpenCV and SSD for vehicle detection, providing a benchmark against YOLO models. Chen (2021) compared YOLOv3 and SSD, finding YOLOv3 more effective for high-resolution traffic detection, while Kim (2021) reported YOLOv4 achieving 98.1% precision compared to SSD and Faster R-CNN.

(Liu et al., 2016) proposed the SSD (Single Shot MultiBox Detector) model, which combines the regression idea from YOLO with the anchor box concept from Faster R-CNN. SSD improves multi-scale object detection by using both lower and higher-level feature maps. Its base architecture is VGG, with the last two fully connected layers replaced by convolutional layers. SSD incorporates the anchor mechanism from Region Proposal Networks (RPN). It achieves a mean Average Precision (mAP) of 74.3% on VOC2007 at 59 fps on a Nvidia Titan X. However, SSD's performance diminishes for small targets, and it struggles with redundant detections due to independent feature maps at different scales.

Ahmed et al. (2021) use SSD with MobileNet for pedestrian detection, focusing on real-time accuracy with the COCO dataset. Their model excels in detecting overlapping pedestrians, providing a balance between speed and accuracy.

Bochkovskiy, (2020) In 2020, Bochkovskiy introduced YOLOv4, which set new benchmarks for the balance of speed and accuracy (Bochkovskiy, 2020). YOLOv4 builds on the original YOLO framework by incorporating several innovations, including Weighted Residual Connections, Cross Stage Partial connections, Cross Mini-Batch Normalization, Self-Adversarial Training, Mish activation, Mosaic data augmentation, DropBlock, and CIoU loss. The model uses CSPDarknet53 as its backbone network and includes an SPP module to expand the receptive field for better feature separation. YOLOv4 replaces FPN with PANet for path aggregation and retains the head structure from YOLOv3. Compared to YOLOv3, YOLOv4 improves accuracy and speed by 10% and 20%, respectively.

Bhambani et al. (2021) propose a YOLOv4-based model that classifies and detects objects in three categories: people, masked faces, and unmasked faces. With components like CPSDarknet53 and SPP, and calibration for social distancing, their model achieves a mean average precision (mAP) of 95% and a frame rate of 38 FPS on NVIDIA Tesla P100 GPU. The YOLOv4 model has shown high accuracy with a mean average precision of 98.1%. In contrast the YOLOv5 has shown further improvements (YOLOv5, 2023). This research enhanced an object detection capability in South Asia using a robust system to process and execute a diverse dataset of 21 vehicle classes.

Li, Lin, Shen, Brandt, and Hua (2015) introduced a CNN cascade for face detection. Their proposed model balances high discriminative capability and efficiency, by operating at multiple resolutions while only evaluating high-resolution candidates only. A CNN-based calibration stage is integrated after each detection stage to enhance efficacy and lowers number of subsequent stages.

RoyChowdhury, Lin, Maji, and Learned-Miller (2015) proposed a Bilinear CNN (B-CNN) method that bridges texture models and part-based CNN models. The B-CNN consists of two CNNs whose convolutional-layer outputs are multiplied using an outer product to create a bilinear feature descriptor. This approach models part-based representations and resembles Fisher vectors, integrating local features with cluster center membership through outer products.

The B-CNN architecture is trained by back-propagating gradients from a task-specific loss function, starting with pre-trained networks (e.g., AlexNet) and fine-tuning on face images. The bilinear layer, similar to a quadratic polynomial kernel in SVMs, improves recognition performance. The B-CNN showed substantial performance gains on face recognition benchmarks with pre-trained networks.

Chaudhari et al. (2021) proposed VGG-16 for identifying tree species from WorldView-3 satellite images. They analysed their approach with Random Forest and Gradient Boosting classifiers, using data from eight visible infrared bands to enhance the detection of the tree species.

Liu (2021) enhanced Faster R-CNN by integrating hard negative sample strategies and feature sharing training. This method retrains the model with negative samples and improves detection speed and precision. The study analyses YOLO, SSD, and RFCN, signifying the strengths of each in terms of detection speed and precision.

Rezaee et al. (2021) use WorldView-3 satellite imagery and VGG-16 for detecting individual tree species. Their pipeline leverages eight visible infrared bands and compares results with Random Forest and Gradient Boosting classifiers, showcasing advancements in forest monitoring.

Han, C., Liu et al., (2018) suggested conducting research on a cutting-edge method to determine the clustered image's shape. The main concept in the image's metamorphosis is the shift from a cluttered image to a clear shape. The pairs between the local shape picture and the requested shape are identified by the point-based descriptive PAD (Pyramid of arc length descriptor). Wavelet wave transforms and Fourier transforms were used to convert the shape into domains in order to measure it. Afterwards, a number of shape descriptors were put forth to gauge the degree of shape similarity. Triangle regions are used as a set of reference points to describe the shapes. Nonetheless, the current methods and shape descriptors for matching shapes are made for matching shapes.

He, K., Zhang et al. (2016) proposed a research on the classical and deep learning methods in object detection. The model’s approach is centered on operation and real-time performance, and precise detection. The difficulties in object detection using deep learning methods. Three processes make up the subtraction method: object identification, background updating, and background modeling. The backdrop subtraction approach updates in real time and works similarly to the frame difference method. The deep learning approach-based object detection model. The models later put out a novel concept.

Numerous researchers have addressed the benefits and drawbacks of the literature reviews that have been covered, but there are still certain problems with object detection and facial recognition. In order to overcome the problem of uncontrolled environmental problem, face direction recognition where the more Headshot and real-time captures help to verify the effectiveness of the Object detection model, enabling higher accuracy.