# PATIENT MONITORING BASED ON HUMAN ACTIVITY USING YOLOV7

**S.A. Hamddi*1, M.B.Omer2 & M.A. Ahmed3**
*1College of Computer Science and Mathematics, Department of computer science, Tikrit University, Saladin, (Tikrit), 34001, Iraq, shihab.a.hamddi@tu.edu.iq
2College of Computer Science and Mathematics, Department of computer science, Tikrit University, Saladin, (Tikrit), 34001, Iraq, mahammed.b.omer35524@st.tu.edu.iq
3College of Computer Science and Mathematics, Department of computer science, Tikrit University, Saladin, (Tikrit), 34001, Iraq

**Abstract:**
A patient in a state of coma requires close observation and interest. however, hospitals now experience a serious shortage of intensive care and emergency nurses. ICU patients must be constantly observed by proactive observers as part of the standard monitoring protocol. One drawback of previous approaches is that one observer can only monitor one patient at a time and must be available around the clock. The goal of this study is to create a system that fixes the problems that traditional approaches have. The suggested method uses deep learning to monitor patients in the critical care unit and send reports to the nurse or doctor. The suggestion system can instantly recognize the intended target (patient), identify any patient actions, and alert the nurse or doctor.

**Keywords:** AI, artificial intelligence, deep learning, Patient monitoring system, YOLOV7.
**DOI:** 10.24297/j.cims.2023.4.3

## 1. Introduction

The need for analysis of activity is growing in the research and development sector. In contemporary times, it has caught the attention of numerous developers. The technique of analyzing various actions carried out by the target is called activity analysis. Activity analysis can be defined as the process of identifying activity, determining if a certain activity is completed or not, defining the type of activity, etc. The term "target" refers to an object that is executing an action. The human being itself is one type of target. The process of analyzing human activity is known as human activity analysis (HAA)[1]. Simple actions like standing still or sleeping are examples of human activity, while more complicated actions include dancing, fighting, playing games, etc. Analysis of human activity is a complex problem for many different reasons. The following are a few difficulties with vision-based human activity analysis: (a) Human behaviors are intricate and incredibly varied, (b) there are many factors to take into account when analyzing human action based on vision shown through pictures or videos, including the lighting position, the brightness, and the video quality, the frames per second, the camera shaking, the obstruction, the unstable frames in videos, etc. (c) Interclass variation—The complication of vision-based activity analysis is posed by the appearance of the human whose activity is being examined. People might vary in age, gender, attire, body type, and height, (d) One of the complexities is the presence of several people inside the analysis's purview. Identification of the target person and the separation of all other people in the frame becomes quite challenging, (e) Analysis of human activity More complexity is required for harsh real-time applications, which are time- and accuracy-sensitive and need the production of results in a short amount of time with the maximum level of precision. Results that are inaccurate or delayed may result in the loss of that same environment or human life, (f) A changing background is another difficulty. When the background moves, it gives the impression that the target object is moving, which could result in inaccurate activity analysis, (g) camouflage the target object under analysis's physical characteristics can match the background, making it hard to identify the target object from the remainder of the background, (h) The target's shadow item may provide the appearance that other objects are present, which causes an inaccurate activity analysis. Different forms of activity analysis can be categorized using a variety of techniques[2].

**Figure 1** patient in ICU

## 2. Related Work

(**Ahmed, Jeon, and Piccialli 2021**) [3] Explain a non-invasive, automated, deep learning-based algorithm-based IoT patient discomfort monitoring/detecting technology. The study's 94% TP (true positive) rate and 7% FP (false positive) rate demonstrated the effectiveness of the suggested system.

**(Qayyum et al. 2019)** [4] used a deep learning system to analyze the RGB camera video and rPPG signal to determine the vital signs. The findings demonstrate that the suggested method delivered precise SNR values with good efficiency for applications involving remote health monitoring.

**(Qiu et al. 2018**) [5] introduced a new framework that combines spatial and temporal filtering with a (CNN) convolutional neural network to remotely estimate the heart rate under realistic conditions, When the trained model is tested against the testing dataset, the results show that 74.13% of the data are accurately predicted, while the remaining data have a considerable mistake that is brought on by the lack of high- and low-frequency components in the dataset.

**(Song, Nho, and Kwon 2017)** [6] suggesting a monitoring system to prevent falls by sounding an alarm when monitored patients are in a risky position; they used the SVM (Support Vector Machine) to create an optimum decision boundary and attained a 99.70 accuracy rate.

**(Chaichulee et al. 2017)** [7] outlines a multi-task convolutional neural network (CNN) model that can automatically determine a patient's presence or absence and, if one is found in front of the camera, segment the patient's skin regions. Combining the fully convolutional network for skin segmentation and global average pooling for patient detection led to high performance.

**(Ruba et al. 2020)** [8] proposes a method to calculate the pulse rate from face videos using CNN models and the EVM methodology. With the EVM approach, an average accuracy of 96.35% and an average execution time of 19s could be attained (for a video of 5 seconds).

**(Bodilovskyi and Popov 2017)** [9] Explain a novel technique for the estimate of complex time domain parameters for camera-based systems. By analyzing the lengths of the exhale and inhale phases and their ratio, the suggested method offers clinicians more options for evaluating and monitoring the patient's condition. The average relative error was 22.6% for the ratio of their lengths, 9.7% for the length of the inhale, and 18.3% for the length of the exhale.

## 3. Materials and Methods

### Acquisition of eye Images

We first collected data from 5 different objects by continually altering the camera's distance and angle to prepare the dataset. Following that, we went over and removed some high-frequency data as well as some redundant data that had been gathered as a result of several exceptional.

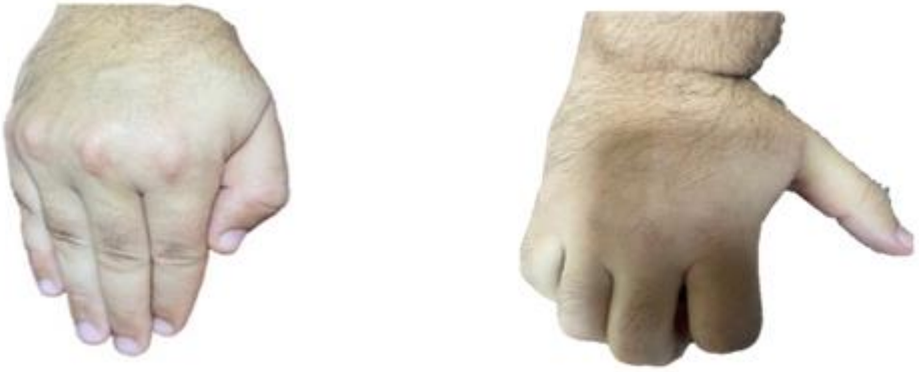


Figure 2 normal and abnormal hand classification

### Image Preprocessing and Dataset Partitioning

From a total of 421 photos, 382 (172 normal and 210 aberrant) were chosen after removing the images of outliers. The eye's placement in the image and the lighting were both taken into account. The 382 images were divided at random, 8 to 2, into a training set (302 images) and a validation set (80 images). the tool for annotating image data "LabelImg" was employed to draw. the target's outer rectangle in each image, concluding the manual labeling of the hand action. Images were tagged with the smallest rectangle around the hand to guarantee that there is the least amount of background in the rectangle. Examples of tagged eye images are seen in Figure 3. When the annotations were saved, TXT format files were created.

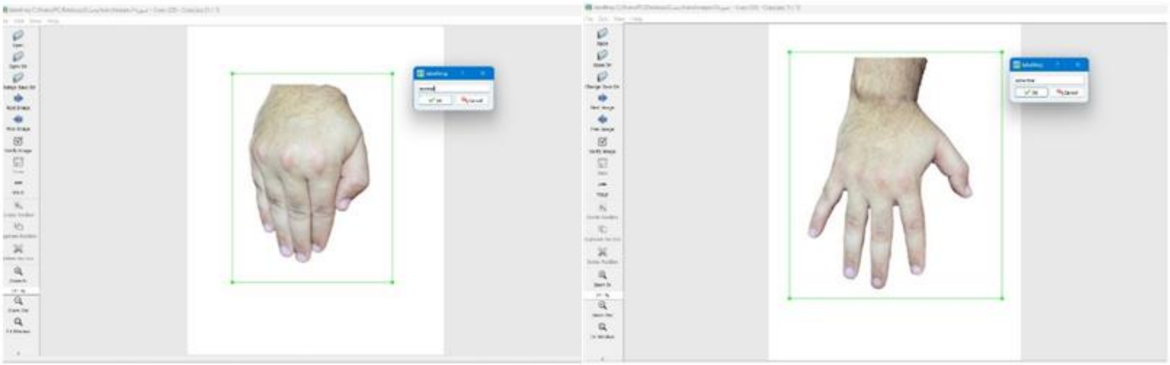
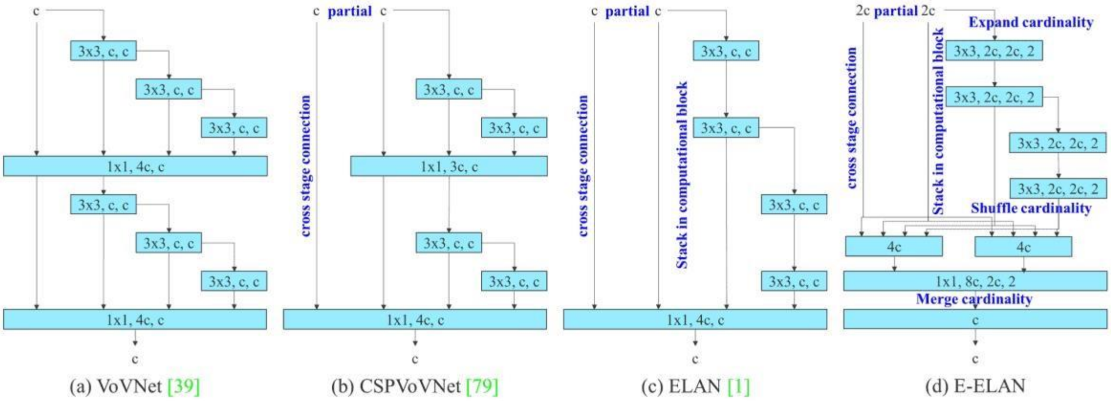


Figure 3 labelling images

### YOLOv7 (You Only Look Only Once)

YOLOv7 is a neural network-developed technique that improves the speed and accuracy of object detection [10]. The network enhances currently employed object detectors with 5-160 FPS accuracy and speed. Compared to YOLOv5, the network is 120 percent faster in the same volume (FPS). The YOLOv5 detector performs worse than the MS COCO dataset test results [11]. YOLOv7 divides the image into several fixed-size grids. The task of finding items within a grid's boundaries belongs to each grid. Each grid cell predicts the bounding boxes and confidence levels for each box. The model's level of certainty and the accuracy with which it claims that the box contains the given object are both indicated by these confidence ratings. If the box contains nothing, the confidence score is zero[12]. Because cells from the photo sustain simultaneous detection and recognition, this strategy significantly decreases computation. However, because multiple cells may anticipate the same object with various bounding boxes, it leads to a significant number of duplicate predictions. YOLOv7 uses non-maximal suppression as a tactic to deal with this issue. With this approach, Lower probability bounding boxes are suppressed or disregarded. for this, YOLOv7 chooses the enclosing box with the highest probability score. Then, with the current high probability bounding box, The bounding boxes with the greatest Intersection over Union (IoU) are eliminated. Repeat the procedure until the desired bounding box with precise object detection is attained[13, 14].



(a) VoVNet [39] (b) CSPVoVNet [79] (c) ELAN [1] (d) E-ELAN

**Figure 2** YOLOv7 Architecture

### Experimental Environment

The PyTorch deep learning framework was used for training and testing on a laptop (MSI GL65 Leopard 10SCSR) running Windows 11 Home 64-bit and equipped with an Intel Core i7-10750H processor, 16GB of RAM, and NVIDIA GeForce GTX 1650Ti graphics, which include 8GB of video RAM. Python 3.9.12 was employed as the programming language. The software bundles contained Visual Studio 2022, OpenCV 4.6.0, CUDNN 8.5, and CUDA 11.3.

### Training Parameters

The training parameters applied in the experiment are shown in Table 1

**Table 1.** Training Parameters

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Image Size | 640 × 640 | Batch Size | 16 |
| Learning Rate | 0.01 | Epochs | 100 |
| Momentum | 0.937 | Weight Decay | 0.0005 |

## 4. Establishment and Evaluation Indicators of Model

The training set was used to train the YOLOv7 model, and the assessment results were validated using the validation set. The model with the best performance weight was chosen in the end to serve serving as the initial model for object detection for hand activity[15]. In order to ensure that an application could select machines in the future, the prediction outputs of the models applied to new data were reviewed. The procedure is depicted in motion in Figure 5. The detection box for the recognized ocular item and the likelihood (confidence) The neural network's final outputs shows whether an object belongs to a specific category.

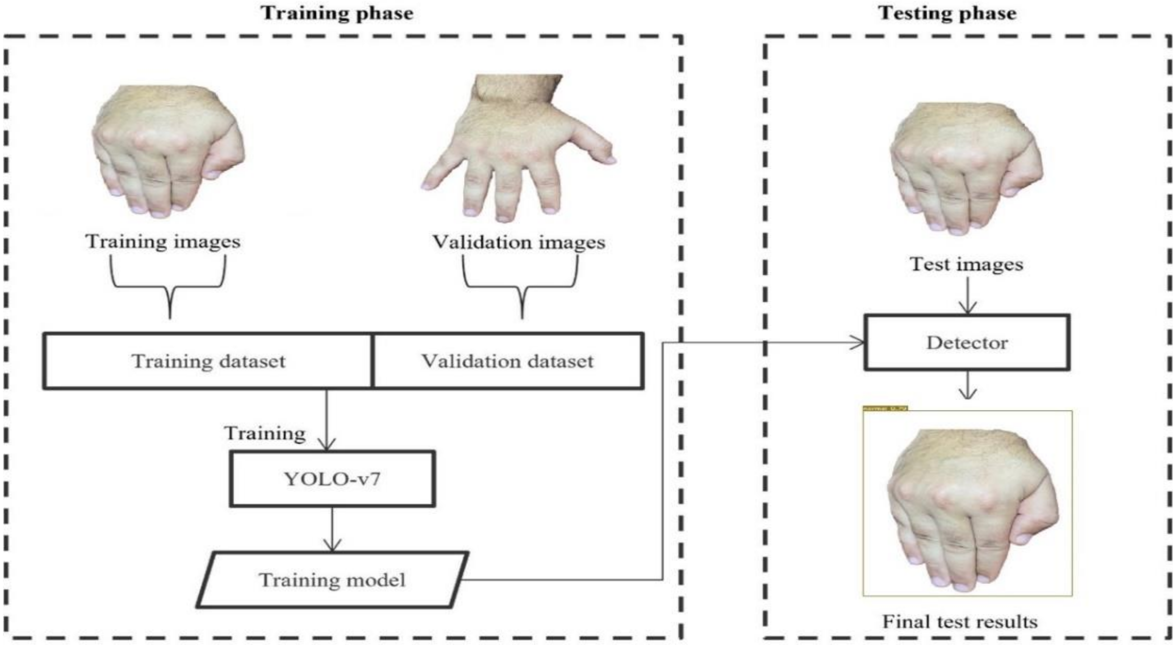


**Figure 5** Workflow of the proposed study

### Establishment of Model

### *Evaluation Indicators of Model*

In this study, the metrics Precision, Recall, F1 score, and Mean Average Precision (mAP) was employed to evaluate the model's performance precisely and objectively. Precision[16], which is defined as the ratio of correct targets to detected targets, is the most widely used evaluation statistic. however, It occasionally misses some details. Thus, for in-depth research, F1score, Recall, and mAP was introduced. The computations for recall, precision, F1 score, and mAP are as follows [16-19]:

Precision: $p = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \times 100\%$ .........(1)

Recall: $R = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100\%$ .........(2)

Average Precision: $AP = \int_0^1 P(r)dr$ .........(3)

Mean Average Precision: $MAP=\frac{1}{n}\sum_{i=1}^{n} \mathrm{API}$ .........(4)

F1 score: $$F1 = 2 \times \frac{P\ X\ R}{P\ +\ R} \quad \text{.........(5)}$$

where FN (False Negative) is the number of things that are undiscovered or missing, TP (True Positive) is the number of correctly detected objects, FP (False Positive) is the number of additional objects found[20].

## 5. Results

Hand detection model is trained using the YOLOv7 model with a custom dataset. Annotated hand image dataset of 382.jpg pictures and 382.txt files that indicate the confidence score was used to train the model. the results of using the YOLOv7 are shown in Table 2. mAP0.5 of 88.3%, precision of 93.2%, recall of 84.5%, and F1 score of 87.5% values were achieved by YOLOv7 for the average of all classes. For the normal class, YOLOv7 achieved the mAP0.5 of 86.4%, a precision of 97%, a recall of 69%, and an F1 score of 80.6%. For the abnormal class, YOLOv7 achieved the mAP0.5 of 90.2%, the precision of 89.5%, the recall of 100%, and the F1 score of 94.4%.

**Table 2.** Results obtained by YOLOv7

| Classes | mAP0.5 | precision | Recall | F1 |
|---|---|---|---|---|
| normal | 86.4% | 97% | 69% | 80.6% |
| abnormal | 90.2% | 89.5% | 100% | 94.4% |
| Average | 88.3% | 93.2% | 84.5% | 87.5% |

From Figure 6 For the normal and abnormal classes, the classification accuracy values are 0.95 and 0.86, respectively, based on the confusion matrix values.

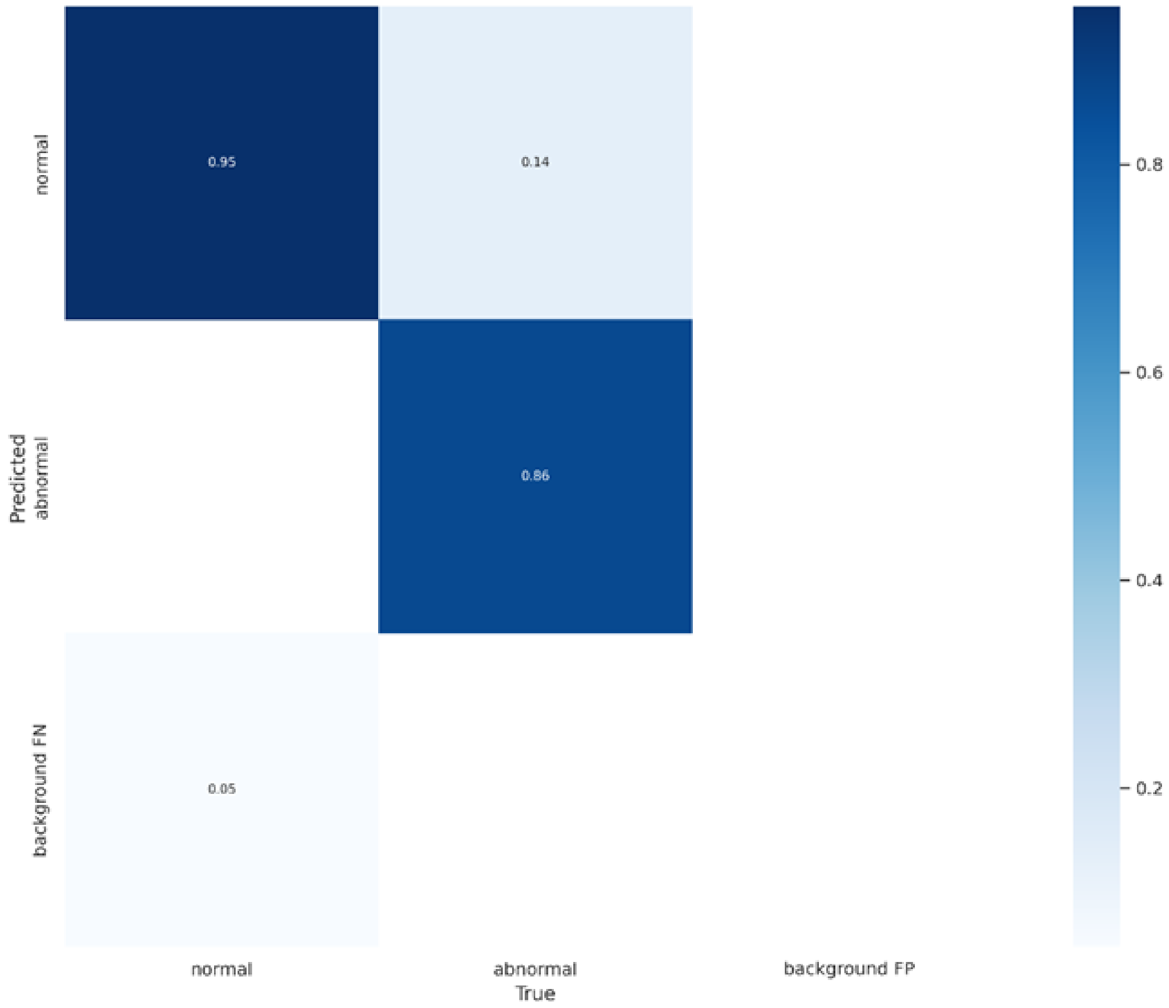


**Figure 6** Obtained confusion matrix by using YOLOv7 architecture design

How many truly relevant results were returned makes a difference between recall and accuracy. Measurement of outcome significance is precision[21]. According to Figure 7, the normal class and abnormal class average precision-recall values were 0.864 and 0.902, respectively.

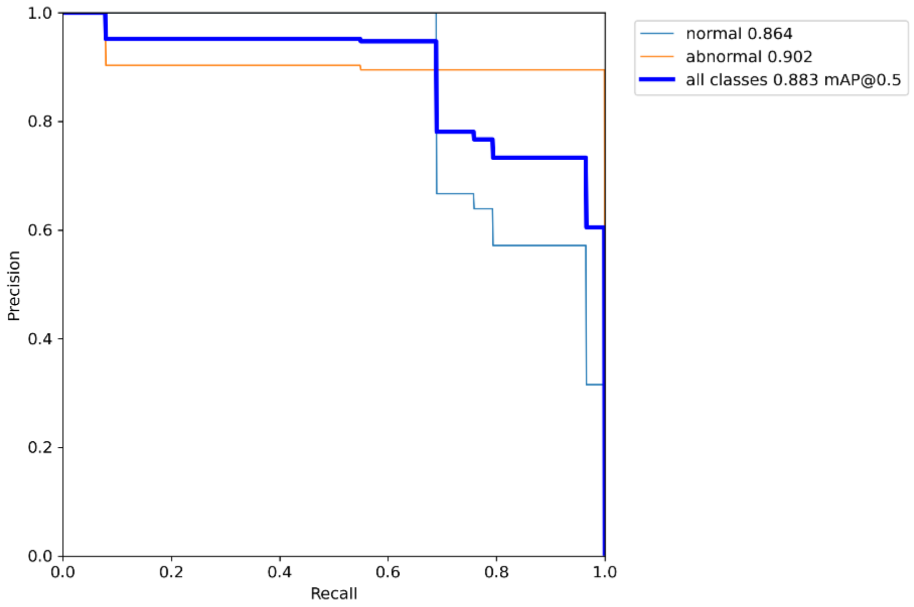


**Figure 7.** Obtained Precision-Recall curve by using YOLOv7x architecture design

## 6. Conclusion

A patient who is in a coma requires close observation and treatment. A nurse is always responsible for the patient's care according to the traditional technique of ICU patient monitoring, although this system has several drawbacks. This paper describes an autonomous vision-based system that keeps track of ICU patients using live video from the ICU room and notifies hospital authorities right once whenever a patient exhibits any activity. Deep learning became famous because it can learn by itself. The hand detection model locates the patient's hands and sends the coordinated location of the patient's hands to the activity detection process' next stage. Finally, patient activity is detected using the Yolov7 model using a bespoke dataset. The implementation of activity detection uses a straightforward picture classification approach. As long as the patient's hand is deemed to be "normal," no activity is said to have been discovered. When a patient's hand is labeled as "abnormal," it is indicated that activity has been noticed. The activity is considered to have been discovered once the hand classification model labels it as "abnormal," at which point hospital officials are promptly notified so that the patient can begin treatment as soon as feasible. Compared to conventional techniques, this automatic ICU patient monitoring system uses less staff and is more precise, economical, and trustworthy.

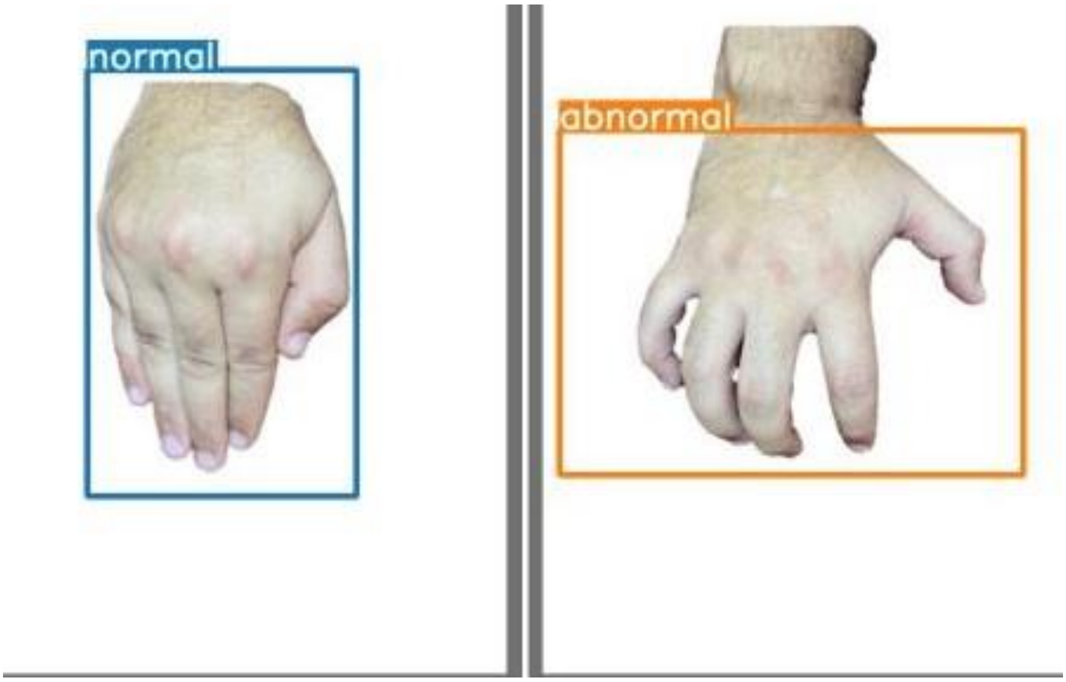


**Figure 3** normal and abnormal hand detection

## References

1. J. K. Aggarwal, and M. S. Ryoo, "Human activity analysis: A review," Acm Computing Surveys (Csur), vol. 43, no. 3, pp. 1-43, 2011.
2. A. H. Alrubayi, M. A. Ahmed, A. Zaidan, A. S. Albahri, B. Zaidan, O. S. Albahri, A. H. Alamoodi, and M. Alazab, "A pattern recognition model for static gestures in malaysian sign language based on machine learning techniques," Computers and Electrical Engineering, vol. 95, pp. 107383, 2021.
3. I. Ahmed, G. Jeon, and F. Piccialli, "A deep-learning-based smart healthcare system for patient's discomfort detection at the edge of Internet of things," IEEE Internet of Things Journal, vol. 8, no. 13, pp. 10318-10326, 2021.
4. A. Qayyum, M. A. Khan, M. Mazher, M. Suresh, D. N. Jamal, and J. T. D. Chung, "Convolutional neural network approach for estimating physiological states involving face analytics." pp. 68-72.
5. Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, and A. El Saddik, "EVM-CNN: Real-time contactless heart rate estimation from facial video," IEEE transactions on multimedia, vol. 21, no. 7, pp. 1778-1787, 2018.
6. K.-S. Song, Y.-H. Nho, and D.-S. Kwon, "Histogram based fall prediction of patients using a thermal imagery camera." pp. 161-164.
7. S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman, and L. Tarassenko, "Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring." pp. 266-272.
8. M. Ruba, V. Jeyakumar, M. Gurucharan, V. Kousika, and S. Viveka, "NON-CONTACT PULSE RATE MEASUREMENT USING FACIAL VIDEOS." pp. 1-6.

9. O. Bodilovskyi, and A. Popov, "Estimation of time domain parameters for camera-based respiration monitoring." pp. 1-4.
10. D. Wu, S. Jiang, E. Zhao, Y. Liu, H. Zhu, W. Wang, and R. Wang, "Detection of Camellia oleifera Fruit in Complex Scenes by Using YOLOv7 and Data Augmentation," Applied Sciences, vol. 12, no. 22, pp. 11318, 2022.
11. I. Ahmad, Y. Yang, Y. Yue, C. Ye, M. Hassan, X. Cheng, Y. Wu, and Y. Zhang, "Deep Learning Based Detector YOLOv5 for Identifying Insect Pests," Applied Sciences, vol. 12, no. 19, pp. 10167, 2022.
12. S. Shinde, A. Kothari, and V. Gupta, "YOLO based human action recognition and localization," Procedia computer science, vol. 133, pp. 831-838, 2018.
13. F. Yang, X. Zhang, and B. Liu, "Video object tracking based on YOLOv7 and DeepSORT," arXiv preprint arXiv:2207.12202, 2022.
14. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022.
15. M. S. Al-Samarraay, M. M. Salih, M. A. Ahmed, A. Zaidan, O. S. Albahri, D. Pamucar, H. AlSattar, A. H. Alamoodi, B. Zaidan, and K. Dawood, "A new extension of FDOSM based on Pythagorean fuzzy environment for evaluating and benchmarking sign language recognition systems," Neural Computing and Applications, pp. 1-19, 2022.
16. D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020.
17. D. Chicco, and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC genomics, vol. 21, no. 1, pp. 1-13, 2020.
18. G. Zebele, "Computer Vision-based detection and tracking of fish in aquaculture environments," 2022.
19. S. K. Abdulateef, T.-A. N. Abdali, M. D. S. Alroomi, and M. A. A. Altaha, "An optimise ELM by league championship algorithm based on food images," Indonesian journal of electrical engineering and computer science, vol. 20, no. 1, pp. 132-137, 2020.
20. J.-E. Park, and K.-A. Park, "Application of deep learning for speckle removal in goci chlorophyll-a concentration images (2012–2017)," Remote Sensing, vol. 13, no. 4, pp. 585, 2021.
21. Z. Kareem, A. Zaidan, M. Ahmed, B. Zaidan, O. Albahri, A. Alamoodi, R. Malik, A. Albahri, H. Ameen, and S. Garfan, "An approach to pedestrian walking behaviour classification in wireless communication and network failure contexts," Complex & Intelligent Systems, pp. 1-23, 2022.