Human Activity Recognition Using Various YOLO Models

Om Subrato Dey

School of computer Science
and Engineering (SCOPE)
Vellore Institute of Technology,
Chennai, India
omsubrato.dey2021@vitstudent.ac.in

DevaryaLakhanpal

School of computer Science
and Engineering (SCOPE)
Vellore Institute of Technology,
Chennai, India
devarya.lakhanpal2021@vitstudent.ac.in

Priyam Chowdhury

School of computer Science
and Engineering (SCOPE)
Vellore Institute of Technology,
Chennai, India
pryam.chowdhury2021@vitstudent.ac.in

Abstract: Activity recognition in the context of computer vision is a very important problem with a wide range of uses in a variety of areas including healthcare, security, and smart homes. Conventional approaches do not offer the capability for interpreting the results, which poses a problem when it comes to trust and implementation for important uses. In this work, we study the applicability of YOLOv5, YOLOv7 and the very recent YOLOv8 versions for real-time HA detection, with a stronger focus on the Explainable AI (XAI). Real time object detection is possible with YOLO (You Only Look Once) models because they are highly efficient. With the high-level object detection feature of YOLOv5, YOLOv7 and YOLOv8, we are able to refine human activity in real-time and incorporate Grad CAM XAI technique for visual and model interpretation. These explanations contribute towards reviewing the determination of the model, so that, the finalized system does not only provide correct outcomes but it is also comprehensible by the end clients. Thus, evaluating our experimental results we prove high relevancy of this approach in the field of human activity detection with high precision, as well as the explainability of the model's predictions, allowing to bridge the performance/interpretability gap in AI-based human activity recognition.

Keywords: Video analysis, actions recognition, XAI, YOLOv5, YOLOv7, YOLOv8, real-time detection, Grad-CAM, model understanding, AI transparency, human actions recognition, object detection, explainability of AI, AI reliability, high-demand applications, based-model explanations, visualization of methods, AI descriptions, accurate detection, detection accuracy vs. model interpretability.

1. Introduction

Human activity detection is an important and urgent problem in the area of computer vision and is widely used in various fields such as health care, security, smart homes and HCI. Real-time identification of human actions also helps prevent human and computer interaction in response to users 'actions, improving automation and safety. Nevertheless, the recent deep learning model including YOLO (You Only Look Once) has greatly improved object detection results in various fields However, these models enclosure the "black-box" problem and lack of interpretability and reliability in certain application scenarios.YOLOv5, YOLOv7 and YOLOv8, which is the latest version in this series, can provide high-speed and high-precision detection of human activities. However, the problem is on how to translate such models to something that the users can understand and perhaps, interact with in cases that understanding of how the model arrived at the decision is essential in safety and compliance or even the debugging process. Explainable AI(XAI) is a solution to the challenge of turning the AI systems 'decision into a form that users can understand. With the help of XAI methods, including Grad-CAM and SHAP, it becomes possible to consider which parts of the input data influence the model and in what way, and thus get insights about the model's behavior.This paper looks at the effectiveness of YOLOv5, YOLOv7, and YOLOv8 in HAD and improves HAD models based on explainability. Our desire is to integrate the state-of-the-art in object detection with a reliable explainability paradigm that will support a high performing and easily explicable system for the identification of human activities.

2. Literature Survey

XAI has emerged as a significant subdiscipline of deep learning to improve clear and comprehensible model interpretation in applications, like human activity identification, where model choices should be explainable. Some of the notable XAI techniques the following techniques are; Grad-CAM (Gradient-weighted Class Activation Mapping), LIME (Local Interpretable Model-agnostic Explanations), and SHAP (SHapley Additive exPlanations). Even though all these methods offer great insight, Grad-CAM offers better results for tasks that involve using convolutional neural networks (CNNs) such as human activity detection with models like YOLOv5, YOLOv7 and YOLOv8.

2.1. **Gradient-weighted Class Activation Mapping (Grad-CAM):** More specifically, Grad-CAM is intended for producing visualization for models based on the CNN. Grad-CAM, which stands for gradient-weighted class activation mapping, uses gradients of the target class flowing down to the final convolutional layer to generate a heatmap that points out the input image regions for which the model pays attention to when making its prediction. According to various researches, Grad-CAM has been proven to excel in tasks such as image recognition and object detection, and thus can seamlessly integrate in models that perform a lot of convolutional computations such as YOLOv5, YOLOv7 and YOLOv8.

Advantages of Grad-CAM:

2.1.1. **Model-Specific Insights:** While model-agnostic methods that are explained above rely on the fact the network gave a certain output, Grad-CAM uses more information about the internal functioning of CNNs and therefore provides more specific and detailed explanations in tasks that operate on spatial data.

2.1.2. **Real-Time Interpretability:** However, in cases where an explanation is needed, Grad-CAM can provide these in near real time to complement and support the real time human activity detection tasks using the YOLO models.

2.1.3. **Visual Explanations:** Grad-CAM, since it generates the heat map, can provide an idea of what regions in the picture have been important for the decision and thus, make the decision reliable.

2.2. **Shapley Additive exPlanations (SHAP):** SHAP is another post-processing model-agnostic technique derived from cooperative game theory. It computes Shapley values – values which describe the value of each feature to the prediction. However, the use of SHAP for feature attribution is highly effective and provides consistent and theoretically sound explanations whereas the extension of this technique to image data is not very clear.
Limitations of SHAP:

Complexity and Scalability: SHAP's computation is high and time-consuming, which is unsuitable for real-time detection of activities since it is scaling well with the depth of input data which may be high-dimensional data like image.2.2.1. Non-Visual Explanations: SHAP is better suited for tabular data and does not naturally provide visual explanations like heatmaps, which are essential for understanding image-based model decisions.

2.3. **Local Interpretable Model-agnostic Explanations (LIME):** LIME provides the explanation for the result provided by any black-box model by surrogating the result by an easily understandable model within the localized vicinity of that result. This is done in a way that the small changes occurring to the input data are monitored in order to understand the corresponding changes in the output. What it means, however, is that LIME is versatile from the model's perspective, and does not depend on the type of model used; this is an advantage but can prove challenging in explaining complex image models such as CNNs, especially when the model is used for real time, applications such as YOLO.

Limitations of LIME:

2.3.1. High Computational Cost: When using LIME, the number of perturbed samples and fitting a local interpretable model make it inefficient and not suitable for realtime applications.

2.3.2. Limited Spatial Context: For image data, LIME can have some issues with identifying spatial understanding of the images because LIME is intended to explain the model at the feature level rather than giving the region based knowledge like Grad-CAM.

2.4. **Why XAI models like Grad-CAM and suitable variants chosen for this task?** In as complex tasks such as human activity detection using YOLOv5, YOLOv7 and YOLOv8, Grad-CAM and its higher version will be of great benefits when compared to LIME and SHAP. Its property of producing real-time heat maps in space makes it more appropriate for the explanation of CNN-based object detection models than the other three models. Grad-CAM and it's higher variants rely heavily on visuals, hence are easy to interpret in real time this makes it easier to determine where the model is focusing in applications such as monitoring human activities, that require precise analysis. Moreover, Grad-CAM's and it's variants has considerably lower time complexity compared to LIME & SHAP and therefore applicable for real time use, to make sure that explainability does not hamper the processing capability of the detection system.

2.4.1 Grad-CAM（Gradient-weighted Class Activation Mapping):

GradCAM is among the best ways of making a lot of logical and visual sense in understanding CNNs architectures. Working via employing of the gradients from the last convolutional layer related to the target class and the map of the

heat density on the input image. This heatmap is used to justify which parts of the image contributed most toward the construction of the model.

GradCAM employs the gradients that interact with the feature maps of the last layer in relation to a loss function. These gradients are accumulative and are then used to give an indication of the relative importance of the feature maps under consideration. The fused feature maps are followed by ReLU activation to generate an output of Class-Discriminative heat map. The good thing as to this is that the solution is very simple to implement and does not include any architectural modifications. It provides class-wise localization so that the user can easily see that how much area of the input the model considered. GradCAM may produce little messy heat maps, so it may not be very precise if the areas to be regarded as important are small or a few. Another one, it can find several areas of interest in the image given that such areas are few and clear in the image it receives only.

2.4.2 Grad-CAM++:

GradCAM++ is an upgrade of GradCAM where the interpretation of visualizations are augmented with Ga for image with regions of interests or cases of model failures. GradCAM++ modifies how gradients are passed through a network to fashion a better and dissimilar heatmap.

It enhances the weighting of the feature maps by employing positive part derivatives of output class score with respect to the maps. This makes it produce more detailed heat maps than GradCAM.It is most appropriate in the contexts where, there are many objects or if the objects within an image are numerous but small in size and spread out all over the image. GradCAM++ aid in providing better localization of the regions that contain information for a given class prediction.It also takes more time to generate than GradCAM because the latter requires more gradients to be computed than in GradCAM.

2.4.3 Score-CAM (Score-weighted Class Activation Mapping):

Moreover, with ScoreCAM it fully provides an opportunity to avoid calculations of gradients using the output score of the model as the measure of region importance of the input image. This leads to below heatmap which shows areas which should be very important for image detection except with using gradients.

This amount computes different regions of the image and then tracks how the model alters the prediction score. It also then substracts the second prediction value from the first to use these as weights to overlay the feature maps and come to the heatmap. Since ScoreCAM does not rely on the gradients then it is less affected by noisy gradients which appear in deep networks. This makes it produce less noisy and more accurate heatmaps most of the time.However, unlike GradCAM and GradCAM++, it requires several forward passes through the model – which can be very time-consuming. However, it probably takes longer time to complete because in some occasion, the model score need to calculate multiple times.

3. Summary of findings

The authors of the paper "Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications," are A. M.. F et al..". In this they discuss the computer vision-based Human Activity Recognition (HAR) systems and especially on how accuracy and response time of the HAR systems can be enhanced using deep learing and certain key application such as real-time surveillance.

Y. Hu elaborately presents from his work: ''Research on Human Activity Recognition Algorithm Based on Metric Learning', that the proposal to use the learning method for enhancing the accuracy of Human Activity Recognition (HAR) in Computer Vision as conventional approaches lack efficiency and introduce potential complications in intelligent monitoring and security applications. It underlines Harmony Agricultural Robot's (HAR) relevant role in the intelligent period for multiple domains including computer vision, pattern recognition, and even machine learning.

In this paper by R. Maurya et al .[3] the authors have proposed a 1D CNN model for complex Human Activity Recognition using tri-axis accelerometer sensor data from a smartwatch and achieved high level of accuracy of 98.28% to recognize complex activities such as studying , playing games and mobile scrolling etc. The model is to be effective for tracking of complex activities and enhance well-being and health; the model become the foundation for future work on HAR.

In the paper of Twinkle et al.[4], they put this kind of research to raise the HAR system by reviewing the past studies and benchmark datasets, and introduce a practical CNNs ensemble framework to lift the precision and stability of one real-world HAR application. The applicability of CNNs is supported by a bibliographic example to shed light on the prospects of improving the HAR outcomes.

V. K. Fukace et al[5] in their paper “Integrating Multiple Public Datasets for Human Activity Recognition using Machine Learning”, explain how it is possible to merge multi-datasets of public HAR and test different machine learning algorithms and among which the authors discovered that the Random Forest, marked the highest results of accuracy and F-score. The findings further show that this integrated kind of analysis provides significant enhancements over prior research applying separate databases.

In a paper, titled, Human Activity Recognition using a Multi-branched CNN-BiLSTM-BiGRU model, by Pooja Lalwani et al., propose a deep learning model that incorporates CNN, BiGRU and BiLSTM to improve the recognition of human activity using wearable sensors to overcome the challenges facing when analyzing sequences and features extracted from such data. The truth is that the model achieves very high accuracy of 99.33% on the WISDM dataset and shows better results than other approaches do have.

In this report, Kumari Priyanka Sinha et al.[7] present their paper titled Preventing Human Activity Recognition from UAV videos, through the conceptualization of the Diminutive Multi- Dimensional Locality Coding based Convolutional Neural Network (DMLC-CNN) model to mitigate challenges of UAV based HAR. This proposed method enhances efficiency through mechanisms such as RoLSA-KNN clustering, LR-SPS segmentation together with DMHG and LOFSC feature extraction.

Deqin Xiao et al.[8] in their paper, “DHSW-YOLO: Based on this, that is “Establishing a duck flock daily behavior recognition model adaptable to bright and dark conditions”, a deep learning model, DHSW-YOLO is designed and developed to precisely detect the daily behaviors of duck flocks, meeting real time detection necessity under bright as well as dark surroundings. It increases correct detection probability, decreases model dimensionality, and accelerates inference time, the subject that provides a technical solution for automated animal movements observation.

This work by Shubham Shinde et al.[9], in their paper, ““YOLO based Human Action Recognition and Localization” Here, the authors propose an approach based on YOLO for human action detection, localization and recognition from disjoint frames of video streams in real or near real time while striving for the highest possible accuracy as individuals perceive it. The method is also tested on the fake Liris Human Activities dataset to demonstrate that it works well in video analysis.

In this research, Shubham Shinde[10], in the paper, “Simple to Complex, Single to Concurrent Sensor-Based Human Activity Recognition: Including both open problems and perception, they offer modern review of literature on Human Activity Recognition (HAR) that concerns sensors, extracting features and classifying by ML and DL, with analyses of limitations in recognizing multiple complicated activities simultaneously. From 190 articles, it examines the open challenges that require investigation and improvements that are expected in the sector in the following years.

The study explained by Quach, Luyl-Da, et al in [11] illustrates the GradCam++ and YoloV8 implementation and establishes the practical hardware avid compatibility in case of THMS dataset with reasonable explainability.

In Kuroki, et al.[12] ,the author has explained the following visualizations and explanations that can be given to help look into the Shapley Explainable Model at the local and global levels.

This paper by Garcia-Garcia et al.[13] presents the comparative study of the outcome of applying different versions of the Yolo Model on the Human Activity Recognition Dataset and demonstrates various probable performance metrics and their corresponding classification on the dataset fed as the input to the system depending on a particular yolo model architecture.

In Dalabehera, Aditya Ranjan, et al.[14] want to define how the aspects known as Human Activities, Emotions and Motions would be recognized and analyzed by the model represented in [14] the respective research analysis. The data set included information regarding human activity recognition and multiple features were represented.

The work by Hsiao, Tzu-Shan, et al in [15] indicates that how the integration of Yolo models and DG-STGCN is quite distinct is explained in the specific research work very systematically is described and the evaluation of the corresponding visualization of it along with the overall performance of the total model is also assessed.

4. Proposed Methodology

The methodology is formulated based on employing appropriate Yolo v5, v7 and v8 algorithm for human activity recognition in the corresponding dataset. Yolo v5 and v7 links with XAI methods such as GradCAM, GradCAM++, and ScoreCAM with the aim of providing visualization in order to provide an explanation about the contents in the model. These versions will be trained on the dataset, and accuracy, fidelity, ambiguity and interoperability of the model will be performed on them before applying the XAI models to them. Yolo v8 is still under development up to the time of writing this paper; thus, we cannot assess the explainability of Yolo v8. These outputs will be then compared in the next versions 5 and 7 in order to achieve best architecture and then using available XAI to explain the decision made. The actual versions lower than v5 are even excluded from testing because they are no longer supporting the most current tools for XAI because of their architecture and XAI model libraries in python that are still in experimental phase or have compatibility issues out of it.

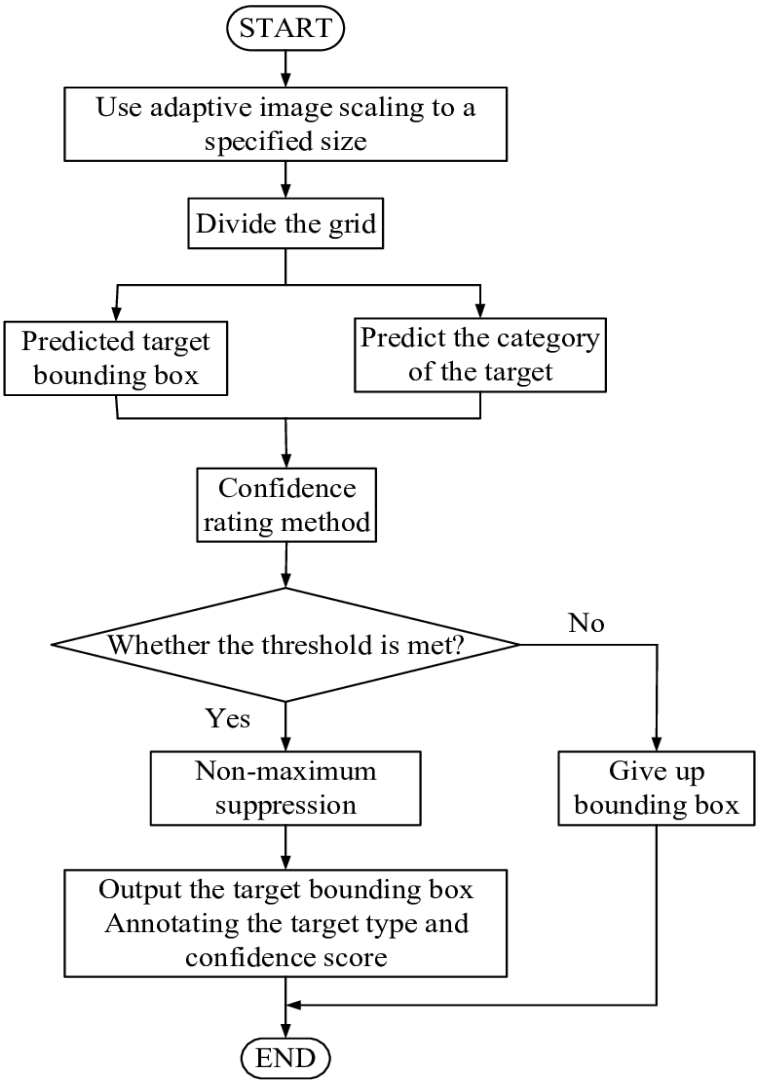

Figure 1. The core steps involved in YOLOv5 detection.

Figure 1. illustrates the flowchart for the YOLOv5 detection strategy: First of all, the input image has its size proportionally submitted given size to have a fixed network size. There is then used a division of the image into cells where each of them is supposed to identify possible locations of the objects. It's used to predict the target bounding box for all objects within the grid cells within the image and in the same instance, it predicts the category or class of the object as well. In addition to the prediction, the network assigns each predict a confidence rating using what is called a confidence rating. In this case, if the provided score prevails the set limit, the process continues otherwise the bounding box is omitted. For the last four points of the IoU threshold with a value greater than or equal to 0.5, the strategy of non-maximum suppression is applied to remove the redundant box while retaining the most accurate one. Finally, the system then then draws the bounding box around the object regarding to label class and confidence level. Last is the representation of the locations of the objects in the image to denote their position.

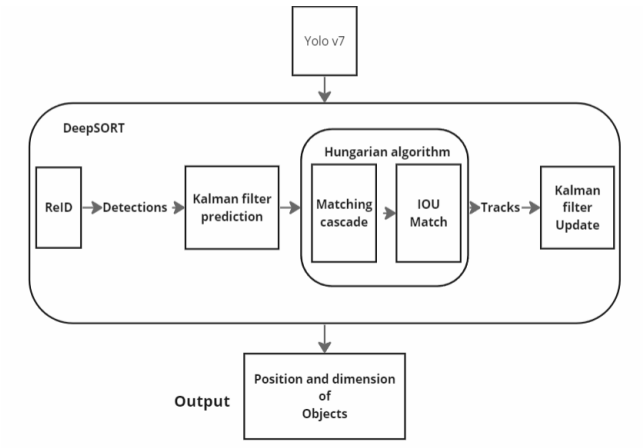

Figure 2. The layer architecture for YOLOv7

Figure 2. represents the approach used in YOLOv7. The system starts with input. Here, YOLOv7 is used on this. These detections are then passed to DeepSORT as the system which has the object class in its responsibilities. It then process initiation: DeepSORT consists of two steps; the first one is ReID which identifies the class of the object. These steps will enable the right classification of each activity in regards to the predicted or detected class of the object. After that the classes are matched depending on the assignment the output shows the class and dimension of classified activity.The flow of methods proposed for v5 and v7 classes is shown below.

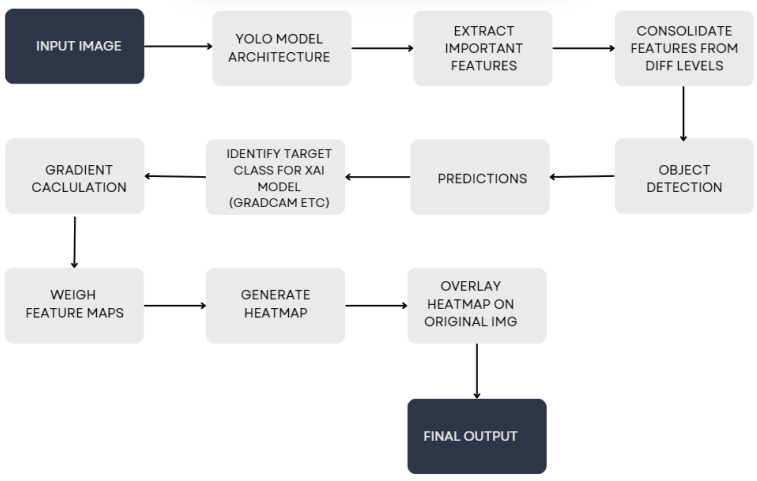

Figure 3. The architecture or flow of data for the Yolo and Grad-CAM input to output pipeline

Figure 3. presents the general format of the YOLO model integrated with the GradCAM explanation technique and the data exchange between them. The process first involves feeding the input image to the YOLO model, and while doing so, utilizing its backbone (for instance, CSPDarknet in the case of YOLOv5 or the E-ELAN architecture in the case of YOLOv7) to extract features of the images. Below the neck, multi-scale feature maps which are used for object detection, can be further processed by PANet or FPN. Usually important features extraction from the image is done using CSPDarknet.Feature aggregation across the different levels (by FPN or PAN) is for improving the features for detection.sed through the YOLO model, which utilizes its backbone (such as CSPDarknet for YOLOv5 or the E-ELAN structure in YOLOv7) to extract features from the image. These features are further processed by the neck (using PANet or FPN), which consolidates multi-scale feature maps for robust object detection. Extraction of important features from the image commonly uses CSP Darknet. Consolidation of features from different levels (often using FPN – Feature Pyramid Network or PAN – Path Aggregation Network) is to enhance feature representations for detection. The last stage of the object detection is therefore performed by predicting the class, the bounding boxes and the confidence scores for each object.Prediction data is then taken to the next component in the full architecture.Subsequently, these final outputs produced by these three XAI models, namely GradCAM, GradCAM++, and ScoreCAM, are compared. which utilizes its backbone (such as CSPDarknet for YOLOv5 or the E-ELAN structure in YOLOv7) to extract features from the image. These features are further processed by the neck (using PANet or FPN), which consolidates multi-scale feature maps for

robust object detection. Extraction of important features from the image commonly uses CSPDarknet. Consolidation of features from different levels (often using FPN – Feature Pyramid Network or PAN – Path Aggregation Network) is to enhance feature representations for detection. The final object detection is then done by predicting the class, bounding boxes, and confidence scores for each object. Prediction data is then moved on to the next part of the full architecture.Then these final outputs generated are compared between all three XAI models implemented, namely GradCAM, GradCAM++, and ScoreCAM.

5. Experimentation and Result Analysis

This study aimed at investigating human activity recognition in the HAR dataset with different YOLO models like YOLOv5, YOLOv7 & YOLOv8. Both YOLOv5 and YOLOv7 could be integrated with XAI approaches such as GradCAM, GradCAM++ and ScoreCAM to explain the predictions on the HAR dataset. On the other hand, YOLOv8 was an active candidate for object detection and not fully compatible with XAI models so we have not taken YOLOv8 into consideration for comparison with YOLOv5 and YOLOv7. To compare the results obtained by YOLOv5 and YOLOv7 of the test images, detections were explained using explanation heatmaps while other parameters such as fidelity, ambiguity and interoperability were also measured. On the performance metrics obtained, YOLOv7 was slightly better than YOLOv5 in most of the test scenarios. Thus, regarding the detection accuracy, XAI methods, and performance YOLOv7 can be considered as the most suitable at the moment, while further development of YOLOv8 may be more appropriate for the HAR recognition once the updated YOLOv8 with optimized settings for this kind of task is available.

Figure 4. A test of the bounding boxes generation using Yolo v5 and Yolo v7 models.

Figure 4. shows how YOLO models create bounding boxes to classify objects in a test image. In this example, the model identifies an action labeled as "eating" with a confidence level of 0.33. The study compares different YOLO versions, like YOLOv5 and YOLOv7, to show why YOLOv7 performs better than the others during this particular timeframe.

Table 1. Performance metrics of the Yolo v5 and Yolo v7 models

Metric	YOLOv5	YOLOv7
Accuracy	0.55	0.75
Precision	0.375	0.692
Recall	0.429	0.643
F1-Score	0.399	0.667

The Table 1. gives us a clear picture of how YOLOv5 and YOLOv7 perform in recognizing actions in our test scenario. When using the HAR dataset, YOLOv7 consistently outperformed YOLOv5. YOLOv5 had an accuracy of 55%, while YOLOv7 pushed that up to 75%, making it a stronger choice for real-time detection. Precision also saw a big jump, going from 37.5% in YOLOv5 to 69.2% in YOLOv7, meaning it made fewer false predictions. Recall, which measures how well the model captures relevant instances, improved from 42.9% with YOLOv5 to 64.3% with YOLOv7. Lastly, the F1-Score, which balances precision and recall, rose from 0.399 in YOLOv5 to 0.667 in YOLOv7. This data clearly shows that YOLOv7 is not only quicker but also more accurate and reliable than YOLOv5 for identifying actions in real-time.

Figure 5. The original test image used for testing all explainable AI models.

Figure 5. shows the original image used to test the performance of all three Grad-CAM models with both YOLOv5 and YOLOv7. This image served as a benchmark to see how each model highlighted important areas, helping us compare the strengths and differences between YOLOv5 and YOLOv7 in action.

Figure 6. The comparative analysis of YOLOv5 and YOLOv7 with integration to GradCAM on a test image

The Figure 6. gives us a closer look at how YOLOv5 and YOLOv7 performed when identifying a “Running” activity using the HAR dataset. With the same test image, YOLOv5 predicted “Running” with 38% confidence, while YOLOv7 was more certain, giving a confidence score of 44%. To see why, we used GradCAM to create heatmaps showing what each model focused on. YOLOv5’s heatmap was a bit scattered, highlighting some relevant features but also including a lot of unrelated areas. In contrast, YOLOv7’s heatmap was much more precise, homing in on key details like the movement of the legs and the runner’s posture. This clarity helped YOLOv7 make a stronger, more accurate prediction, showing it’s better at zeroing in on the right features for understanding specific activities.

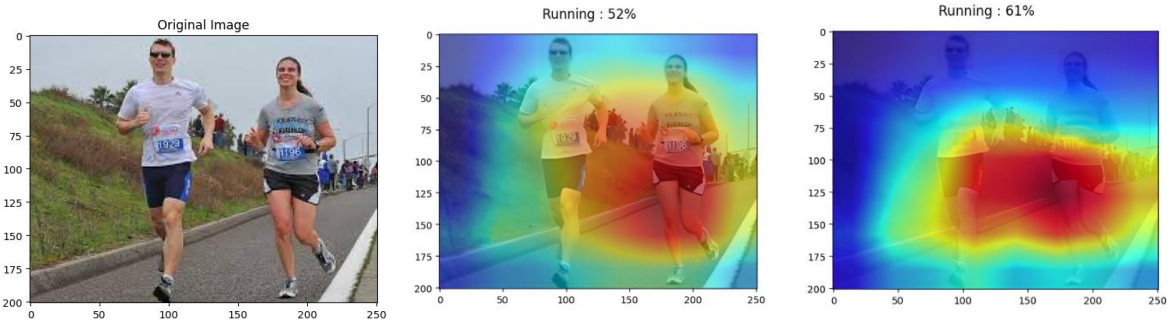

Figure 7. The comparative analysis of YOLOv5 and YOLOv7 with integration to GradCAM++ on a test image.

In Figure 7., we used the same image of a person running to compare how well YOLOv5 and YOLOv7 perform with GradCAM++. YOLOv5 recognized the action but only gave a confidence score of 52%. On the other hand, YOLOv7 was a bit more certain, scoring 61%. To understand the differences in their predictions, we created GradCAM++ heatmaps to see what each model was focusing on. YOLOv5's heatmap was pretty scattered, showing some relevant areas but also highlighting a lot of unnecessary details. In contrast, YOLOv7 had a much sharper focus and larger area focusing highly on the legs , zeroing in on important features like the runner's legs and posture. This suggests that YOLOv7 does a better job of identifying the key elements that matter when recognizing activities.

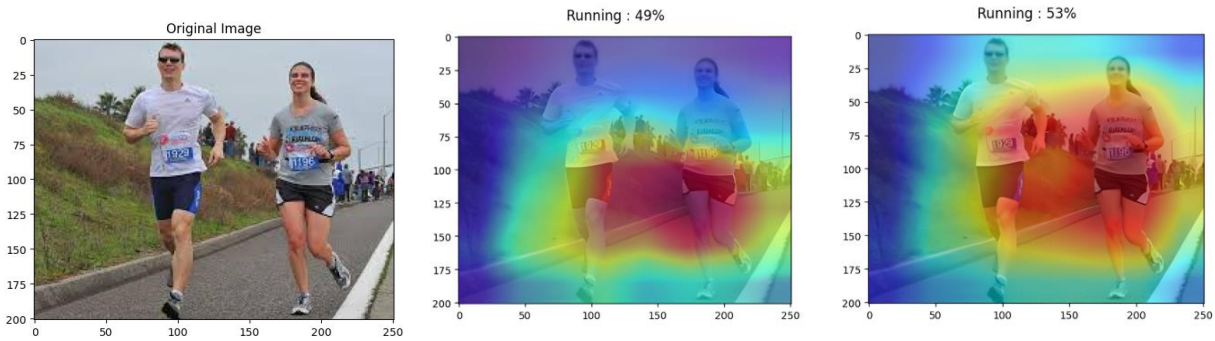

Figure 8. The comparative analysis of YOLOv5 and YOLOv7 with integration to Score-CAM on a test image

In Figure 8., we compared how well YOLOv5 and YOLOv7 performed using ScoreCAM on the Human Activity Recognition (HAR) dataset, specifically testing the same image of someone running. YOLOv5 predicted "Running" with a confidence score of 49%, while YOLOv7 edged ahead with a score of 53%. The ScoreCAM heatmaps for both models helped us see which parts of the image influenced their predictions.

YOLOv5's heatmap highlighted some important areas related to running but also included a few irrelevant spots, indicating that it had a broader focus. In contrast, YOLOv7's heatmap was much sharper and more accurate, concentrating on the runner's posture. This focus led to a clearer interpretation of the activity and gave a higher confidence in its prediction. Overall, this comparison shows that YOLOv7, with its more advanced architecture and precise ScoreCAM visualization, offers a better understanding and classification of activities compared to YOLOv5 but also has a tough competition against YOLOv7 with GradCAM++.

6. Comparative Analysis of XAI Models

Other than visualizations so far concerned, **c**omparing GradCAM, GradCAM++, Score-CAM on YOLOv5 and YOLOv7 on the basis of the performance metrics obtained, a comparative view of fidelity, ambiguity, and interoperability highlights the differences that determine model efficiency.

Table 2. Performance metrics table of the Deep Learning Models v5 and v7 with integration to the XAI models.

Metric	YOLOv5 with GradCAM	YOLOv5 with GradCAM++	YOLOv5 with Score-CAM	YOLOv7 with GradCAM	YOLOv7 with GradCAM++	YOLOv7 with Score-CAM
Fidelity	0.685	0.675	0.755	0.732	0.720	0.801
Ambiguity	0.143	0.047	0.068	0.027	0.032	0.061
Interoperability	0.74	0.69	0.71	0.79	0.78	0.7

The above Table 2. shows the results for YOLOv5 where GradCAM attains a fidelity score of 0.685 thus implying that that the explanations given by GradCAM are partially, superior in fidelity with the model prediction We also obtain an ambiguity score of 0.143 thus pointing out that GradCAM has relatively high degree of uncertainty. Interoperability for GradCAM is 0.74 which indicates the aspect that quality of the explanation is well supported by simplification. A variation of this is called GradCAM++ with the close fidelities to 0.675 and ambiguities to 0.047., although the introduced concept of Object stands at 0.69 it seems the authors aimed at finding middle ground between perfect explanations and perfectly ambiguous results.The score CAM show the highest fidelity of 0.755 meaning the most

relation of the given explanations to the decision making of the model but a rather high level of the samples ’ambiguity is 0.068 and a moderate but rather specific level of interoperability is 0.71 what can be characterized as complicated but relatively clear.vely high uncertainty. Interoperability in GradCAM is at 0.74 which shows that the explanation quality is well complemented with simplicity. A modification of GradCAM is called GradCAM++ and, though its fidelity scores are 0.675 and ambiguity scores 0.047 surpass those of GradCAM, the object of 0.69 points to a compromise between high precision of explanations and low interpretability of results. Score CAM gives the highest fidelity of 0.755 implying the most relation of the explanations given to the decision making of the model but a comparatively high ambiguity score of 0.068 and relatively moderate degree of interoperability score of 0.71 suggesting a convoluted yet somewhat understandable interpretation.

The performance of the XAI models is a little better when changing to YOLOv7. It further rates GradCAM on a scale of fidelity and ambiguity; whereas fidelity is 0.732 and the ambiguity is 0.027, the authors explain that GradCAM is more ‘aligned ’with YOLOv7 decision outputs. An interoperability score of 0.79 also means that it is very efficient in providing human interpretable features to it. The fidelity of GradCAM++ is slightly lower than that of GradCAM with a mean of 0.720; although the results of ambiguity and decreases in interoperability are quite insignificant with 0.032 of ambiguity and 0.78 interoperability. Lastly there is the Score-CAM which ranks the best per toss on the fidelity score of (0.801) though it records a relatively high ambiguity of (0.061) together with an interoperability of 0.7 claiming moderate fidelity between interpretation precision and visibility. This comparative work aims at comparing and contrasting the results of the analysis to show that ScoreCAM performs better than the other XAI models for both variants of YOLO; at the same time, it advances the argument that the choice of the XAI model to be used must include fidelity loss, ambiguity, and interoperability.

7. Conclusion and Future Work

The work shows how integration of YOLO models with Explainable AI (XAI particularly with Grad-CAM is possible and efficient for real-time recognition of human daily activities. This strategy contributes to addressing one of the topics of the rather hot discussions regarding the blurriness of neuronal nets and provides more insights into the subsequent model choices. Where it concerns matters of life and death or issues of security as in health my service provision or identifying terrorists, then it becomes quite pertinent to have an understanding on the mechanisms in which an AI model comes up with a decision. The presented procedure investigates the problem of defeating the trade-off trade-off and offers evidence that is feasible to create rather accurate interpretative models that could be used straightforwardly by naive audiences.meet one of the topics of frequent discussions about the opacity of neuronal nets and provides additional understanding of further model decisions. When it comes to the AI system affecting life and death, or security concerns, as is the case with healthcare or terrorists ’identification, then the mechanisms by which an AI model makes its decision must be well understood to establish trust. The presented procedure addresses the challenge of defeating the trade-off trade-off and provides proof that it is possible to develop effective models which are easy to understand by simple audiences.

It became more efficient with YOLO models wed with Grad-CAM’s explanation of the model’s decision-making scheme makes these models more suitable for the future usage of AI in primary real-world application. In as much as dealing guess work out, the system makes stakeholders have trust and faith in the result by showing which part of the input affected the decision made by the model, thus is inline with some of the major objectives in developing AI and thus, one for developing systems functionalities that are not only efficient but also usable. This work complements an attempt to solve the ‘black-box ’issue of AI models and extends the assumption of the rising importance of interpretable models in industries that demand unambiguous clarity on decision-making processes.Further research possibilities are highlighted as follows: One important direction is the addition of some other more detailed XAI techniques such as the Randomized Input Sampling for Explanation (RISE), which can create pixel level saliency maps that will enable the model to describe what part of the images is most consequential in terms of prediction. While Grad-CAM already provides visually interpretability, RISE might be able to advance that idea and come up with a more effective way of achieving importance maps for object localization. This could in turn result in developing a better appreciation of the decisions making especially in the YOLO models that are beneficial in undertaking object detection tasks.ary real-world usage. By allowing stakeholders to see what parts of the input influenced the model’s decision, the system becomes more trustworthy, aligning with key goals in AI development and so, one for constructing systems functionalities that are not only effective, but are also accessible. This approach makes a contribution to address the ‘black-box ’problem of AI models and strengthens the premise of increasing significance of explainable models in industries that require clear understanding of decision-making processes.There are several possible directions for additional investigation based on this study. One important direction is the application of other more complex XAI techniques such as the Randomized Input Sampling for Explanation (RISE), which can generate pixel level saliency map, which can provide a more richer and more detailed description of what part of images is most influential in making the model prediction. Although Grad-CAM already provides interpretable visualizations, RISE may be able to improve on that concept and provide a better method to create importance maps for

object localization. This could lead to a refined perspective of the process of making and executing decisions especially on YOLO models, which are efficient in performing object detection tasks.

Moreover, future studies could explore the use of XAI approaches to other successful state-of-art object-detection model like the faster region-based convolutional neural network (Faster R-CNN), single shot detection (SSD), or efficient detection (EfficientDet) that have also shown high accuracy-efficiency balance. A Comparison of these models alongside with more extensive and varied HAR Datasets may cast further light on the desirable approaches to Explainability for diverse models especially for fine-Grained localization and recognition of Objects. They may also be useful to contextualise how one XAI approach impacts, for example, precision or recall rates and inference time, metrics of explainability that are arguably more objective than journal readability. Presumably, as the work on the development of new versions of YOLO model proceeds, for instance, connected with the YOLOv8, there is a need for the further works that will cover the changes in the implications of the new models of YOLO. These optimizations particularly when enhancing the accuracy of recognition, time consumed or degree of interpretation shall likely have a synergy effect with enhanced applicability and growth in employs for the model in question.Understanding how these architectural changes affect explainability will help catalyze the design and creation of better understandable models for applications like self-driving cars, health care and security systems. (EfficientDet) which have also exhibited high accuracy-efficiency trade-off. A comparative analysis of these models in combination with more extensive and varied HAR datasets could shed more light on the desirable approaches to explainability for different models in particular the fine-grained localization and recognition of objects. Such comparisons may also help explain how a particular XAI technique affects, for example, precision or recall rates and inference time, more objective indicators of model explainability. Presumably, as the development of new versions of the YOLO model continues, for example, connected with the YOLOv8 model, it is necessary to consider further works exploring the changes in the architecture of the new models. These optimizations, especially when increasing the recognition accuracy, speed, or interpretability, shall likely go hand in hand with improvements in practical usefulness and expansion of the applications of the model itself. Realizing how these architectural changes impact explainability will advance the design and development of more understandable models to be used in related environments such as self-driving cars, healthcare applications, and security monitoring systems.

Second, more complex and detailed descriptions should be introduced in the further studies concerning XAI methods because the types of explanations, which are presented by the extensions of the LIME method, are also more complex and detailed. In general, the present visualizations are satisfactory although they can be improved and polished further to make it easy for such users who want to be presented with clear and obvious information. For instance, increasing the quality of heatmaps or saliency maps will un-complicate the two and assist the decision makers to make right decisions in the right time because those wrong decisions in relation to such areas are likely to cost lots of lives, resources, and time. The last but not the least, the enhancement of YOLO models and other object detection architectures for XAI will reveal new areas of AI application in sectors involving explanation. They can be thus made more transparent and accurate by applying new techniques for XAI like RISE, Grad-CAM, etc. This will make it possible for confidence in or utilization of AI systems in areas such as finance, the health care sector and the legal field that require model interpretability.This, then, makes this paper make the discussability of a mission-critical, complex and high performing AI system more achievable subsequent studies regarding XAI methods. The present visualizations as a whole are quite useful but require enhancement and polishing in order to make them easy for understanding for such users who intend to get clear and obvious information. For example, enhancing heatmaps or saliency maps can un-complicate those and help decision-makers make more accurate decisions quickly because the wrong decisions when it comes to these areas may result in massive loss of lives, resources, and time. Last but not least, improving YOLO models and other object detection architectures for XAI will uncover new applications of AI in sectors requiring explanation. Applying new XAI techniques such as RISE, Grad-CAM, etc., these models can be made more transparent and accurate. This will let AI systems to be trusted and used in various areas like finance, health care and law that necessary model interpretability. Therefore, this study brings the discussability of a mission-critical, complex and high performing AI system closer to reality.

References

1. Manaf, Abdul, and Sukhwinder Singh. "Computer vision-based survey on human activity recognition system, challenges and applications." *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. IEEE, 2021.

2. Hu, Yuncong. "Research on Human Activity Recognition Algorithm Based on Metric Learning." *2024 3rd International Conference for Innovation in Technology (INOCON)*. IEEE, 2024.

3. Maurya, Raman, et al. "Complex human activities recognition based on high performance 1D CNN model." *2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*. IEEE, 2022.

4. Twinkle, B. Kaur and P. Goel, "A Framework for Human Activity Recognition using Deep Learning Techniques," 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2023, pp. 1-6, DOI: 10.1109

5. Fukace, Vinicius K., Yandre MG Costa, and Igor Da P. Natal. "Integrating multiple public datasets for human activity recognition using machine learning." *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2023.

6. Lalwani, Pooja, and Ganeshan Ramasamy. "Human activity recognition using a multi-branched CNN-BiLSTM-BiGRU model." *Applied Soft Computing* 154 (2024): 111344.

7. Sinha, Kumari Priyanka, and Prabhat Kumar. "Human activity recognition from UAV videos using a novel DMLC-CNN model." *Image and Vision Computing* 134 (2023): 104674.

8. Xiao, Deqin, et al. "DHSW-YOLO: A duck flock daily behavior recognition model adaptable to bright and dark conditions." *Computers and Electronics in Agriculture* 225 (2024): 109281.

9. Shinde, Shubham, Ashwin Kothari, and Vikram Gupta. "YOLO based human action recognition and localization." *Procedia computer science* 133 (2018): 831-838.

10. Ankalaki, Shilpa. "Simple to Complex, Single to Concurrent Sensor based Human Activity Recognition: Perception and Open Challenges." *IEEE Access* (2024).

11. Quach, Luyl-Da, et al. "Tomato Health Monitoring System: Tomato Classification, Detection, and Counting System Based on YOLOv8 Model With Explainable MobileNet Models Using Grad-CAM++." *IEEE Access* (2024).

12. Kuroki, Michihiro, and Toshihiko Yamasaki. "Fast Explanation Using Shapley Value for Object Detection." *IEEE Access* 12 (2024): 31047-31054.

13. G Garcia-Garcia, Sagrario, and Raúl Pinto-Elías. "Human Activity Recognition implenting the Yolo models." *2022 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*. IEEE, 2022.

14. Dalabehera, Aditya Ranjan, et al. "Real-time Criticality Evaluation Through Vision-based Human-centric Emotion, Activity and Object Interactions." *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*. IEEE, 2024.

15. Hsiao, Tzu-Shan, et al. "Integrating YOLO and DG-STGCN Systems for Enhanced Human Action Recognition." *2024 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*. IEEE, 2024.