LAUTECH Journal of Engineering and Technology 18 (1) 2024: 173-180

DEVELOPMENT OF AN INTELLIGENT SURVEILLANCE SYSTEM FOR GESTURE RECOGNITION USING MACHINE LEARNING

Omodunbi B. A., *Soladoye A. A., Adeyanju I. A. and Adeyo S. O.

Department of Computer Engineering, Federal University Oye-Ekiti. Nigeria.
Corresponding author: afeez.soladoye@fuoye.edu.ng

ABSTRACT

Over the years, video surveillance systems have been in use in various contexts such as traffic control, crowd control and protecting wildlife. In a dispensation characterized by security concerns and technological advancements, it has become necessary to develop intelligent systems for surveillance. Intelligent video monitoring has become a vital tool for boosting security and safety in public areas. These systems combine the use of computer vision, machine learning, and artificial intelligence techniques to analyse video data and alert security personnel to potential threats. However, traditional manual monitoring methods need a lot of manpower and are easily compromised. In addition, the cost of video surveillance goes up with mass data storage. This research proposes an Intelligent Surveillance System for Gesture Recognition using machine learning techniques. Video data showing normaly and anomaly gestures were captured using a phone camera connected to a personal computer (PC) using Iriun Webcam. The coordinates of this and landmarks obtained from these videos were gathered using media pipe showing different gesture classes. The hold-out evaluation method was employed with 70-30% split. The acquired videos were trained for gesture recognition using four different machine learning pipelines namely: linear regression, Ridge Classifier, Random Forest Classifier, and Gradient Boosting Classifier. Ridge classifier gave the best average accuracy, precision and F1 score of 99.8, 99.8 and 99.4% respectively when evaluated. This study shows that Ridge classifier provides a good classifier for gesture recognition using video coordinates and landmarks.

Keywords: *Gesture recognition, Surveillance, Machine Learning, Ridge classifier.*

INTRODUCTION

Intelligent video surveillance makes use of advanced technologies like artificial intelligence (AI), machine learning (ML), and computer vision to automatically monitor and analyse video feeds from security cameras. These technologies are used by intelligent video surveillance systems to detect, recognize, and track objects, people, and vehicles in real time. They can also automatically warn and notify security professionals in the event of any suspicious or unusual behaviour. Recognition of human activities is a recent field that is intended to provide techniques and methods allowing the detection and classification of human activities and extended now to recognize normal or abnormal activities (Neha *et al*., 2022). Video surveillance systems can now identify potential dangers and unusual activities in real-time courtesy of the development of high-resolution cameras and sophisticated video analytics. However, the efficiency of these systems is somewhat constrained by some issues. The first issue is that current systems frequently have high false alarm rates, which can cause security staff to become overworked and unable to react to real threats. Second, video surveillance systems must balance the need for security with each individual's right to privacy, generating ethical issues and creating legal challenges. In this research, we employed

MediaPipe for gesture recognition using machine learning.

MediaPipe is an open-source framework created by Google. It offers a comprehensive and adaptable selection of machine learning solutions for diverse multimedia processing applications, including computer vision and media understanding. It is used to attain estimates of 2D human joint coordinates in each image frame (Kim *et al*., 2023). It is used for creating pipelines that carry out inference operations over arbitrary sensory data. A perception pipeline can be created using MediaPipe, as a graph containing modular components, such as model inference, media processing algorithms, data conversions, etc. While sensory information like audio and video streams enter the graph, perceived descriptions like object localization and facial landmark streams leave the network (Lugaresi *et al.*, 2019). It makes it easier to design applications that analyse and work with audio, video, and image data. MediaPipe is known for its versatility, real-time performance, and ease of integration into various platforms. Due to its real-time processing capabilities and compatibility with both desktop and mobile platforms, MediaPipe can be used for a variety of practical applications. It consists of several core components that provide practical solutions for multimedia processing. They include face detection, hand tracking, pose estimation, object detection and tracking, face mesh, holistic, and audio processing.

Machine learning is a subfield of artificial intelligence (AI) that is centered on the development of algorithms and techniques that allow computers to learn from data and make predictions or decisions without being explicitly programmed. To put it in another way, as computers are exposed to more data, machine learning algorithms enable them to automatically perform better on a task. The basic concept underlying machine learning is to identify patterns and relationships within data, which can then be used to make predictions or decisions on new, unseen data. This is usually accomplished by training the machine learning model, and exposing it to a dataset containing examples of the inputs and their corresponding outputs. The model learns from this data by adjusting its internal parameters or structure to minimize errors or discrepancies between its predictions and the actual outputs (Omodunbi *et al.*, 2024).

Machine learning plays a significant role in the field of gesture recognition, a field within computer vision and human-computer interaction that focuses on understanding and interpreting human gestures captured through cameras or sensors. These systems use machine learning algorithms to analyze data captured from sensors such as cameras, depth sensors, or motion sensors, and classify gestures into predefined categories. Machine learning techniques have proved beneficial in extracting features and classifying human gestures/activities (Ambati and El-Gayar, 2021).

Gesture recognition is an area of computer vision and artificial intelligence that deals with the identification and analysis of human gestures and movements using computer algorithms. It involves the use of computer vision techniques to extract meaningful features from the video data captured by cameras or other sensors. These features are then analysed by machine learning algorithms to recognize patterns of movement that correspond to specific gestures. Gesture recognition can be accomplished through a variety of ways, including vision-based methods that utilize cameras to detect and track the user's body motions and sensor-based (Kim *et al*., 2023) methods that employ sensors to recognize and interpret the user's movements.

Violence not only threatens societal stability and growth but also puts people's lives and property in grave jeopardy. The surveillance system is essential for keeping an eye out for instances of violent behavior. With the rising usage of surveillance in daily life and the exponential growth of video data, it is becoming more and more unfeasible to manually identify violent behavior in surveillance due to the excessive manpower involved. As a result, developing an automated system to identify acts of violence is essential

Sahoo *et al*. (2022) proposed an end-to-end fine-tuning method of a pre-trained CNN model with a score-level fusion technique to recognize hand gestures in a dataset with a low number of gesture images. They evaluated the effectiveness of the proposed technique using leave-one-subject-out cross-validation (LOO CV) and regular CV tests on two benchmark datasets. Luo *et al*. (2022) proposed a human behavior recognition model that is based on an improved EfficientNet. The MBConv-building block of EfficientNet-module is streamlined, residual structure is added, and a better activation function is chosen to improve the model. To overcome the common issues of privacy, security, and robustness with traditional computer vision human behavior recognition, they employed Light Detection and Ranging (LiDAR) for human behavior recognition. This enhanced the EfficientNet model to have a better training effect on small datasets with fewer parameters and computational effort.

Human motions can be complicated and differ greatly from person to person, making gesture detection difficult. However, recent developments in artificial intelligence and machine learning have significantly advanced the field, and gesture detection systems are now more precise and dependable. This research aims to analyse the performance of different Machine Learning techniques. Four different machine-learning algorithms were applied to the dataset. They are linear regression, Ridge Classifier, Random Forest Classifier, and Gradient Boosting Classifier. The results of these algorithms are compared and presented in the form of tables and graphs and Ridge Classifier is identified to be the best algorithm based on the results obtained.

METHODOLOGY

The implementation of this system involves two stages: the development and evaluation stages. The development phase involves creating and refining the intelligent video surveillance system. It includes problem definition, data collection, algorithm and model selection, training and model development, system integration, and performance optimization. In the evaluation stage, the performance, effectiveness, and reliability of the developed system are assessed. It includes selecting evaluation metrics, using an evaluation dataset, evaluating system performance, conducting a comparative analysis, refining the system based on evaluation results, and validating through user feedback.

Figure 1 shows the workflow of the proposed intelligent video surveillance system, as this shows the stages involved in the recognition of the gestures starting from the acquisition of the video to tracking the video via media pipe and feature extraction from the video before it was sent for gesture recognition using the aforementioned machine learning algorithms and this would initiate notification or alert if any abnormal gesture or posture is detected in the video.

Figure 2 further shows the research methodology and approaches employed in this study to achieve the actions shown in Figure 1. This study started with the acquisition of videos of different postures as the input dataset, after which the necessary features were extracted from the videos which were

the landmark and coordinates of the postures and this data were further organized and classified using four machine learning algorithms, this trained model was

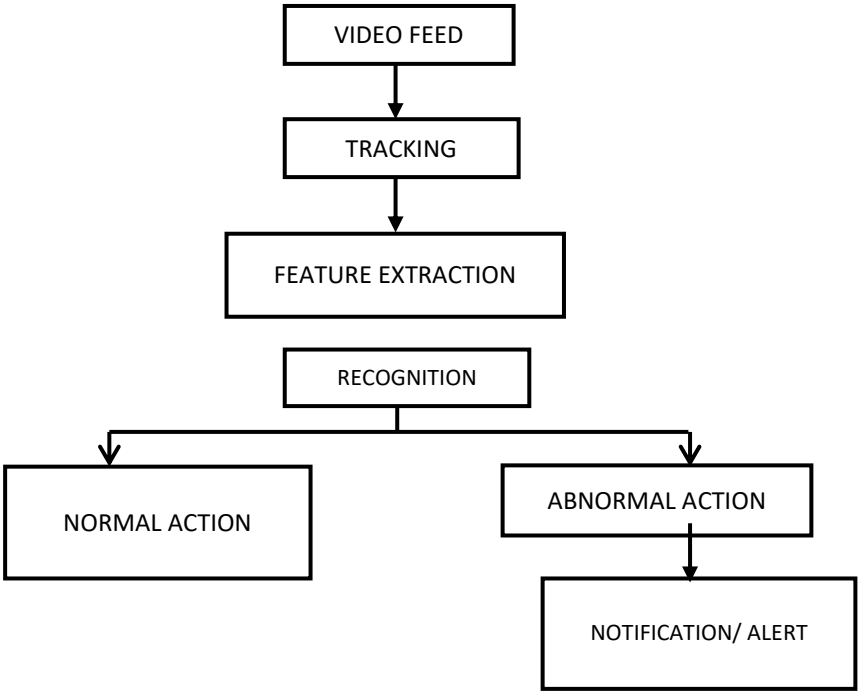

Figure 1: Block diagram of the Intelligent Video Surveillance System

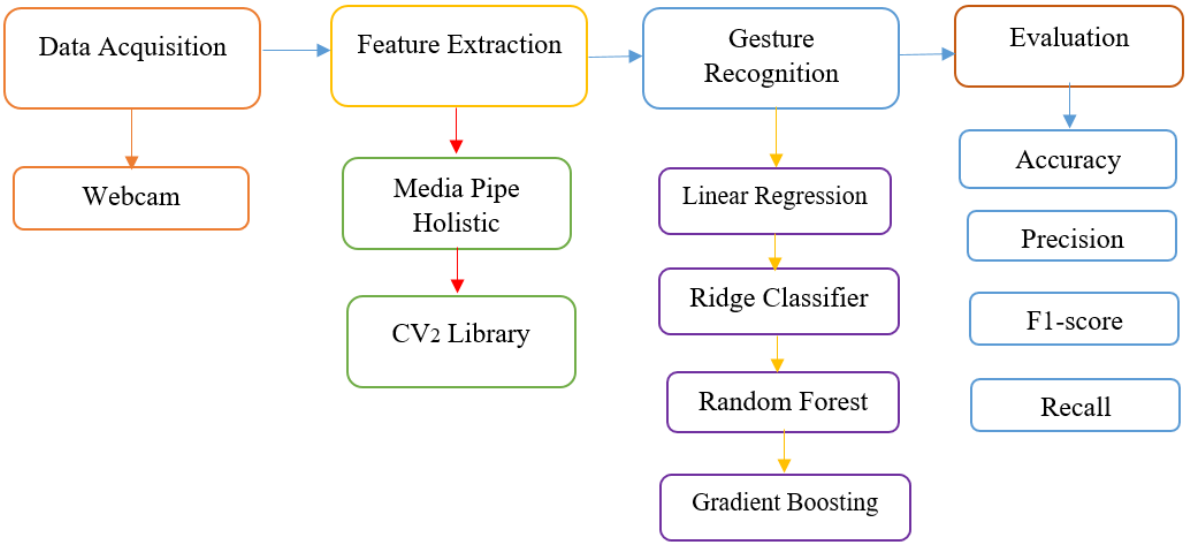

Figure 2: Overview of Research Methodology

afterward evaluated using various performance evaluation metrics. This representation is diagrammatically represented in Figure 2 for easier clarification.

Data Acquisition

The proposed system uses the Mediapipe framework to obtain data in real-time. The real-time data (video) was captured using a phone camera connected to a personal computer (PC) using software called “Iriun Webcam”, installed on both

the phone and PC and connected to a single network. This was done to get a better and broader angle, as the phone camera can be moved around and offers a fuller picture than the PC's webcam. A total of nine poses were done namely: Normal, fight cheek, fight kneel, punch, elbow fight among others. The mediapipe holistic framework was used in conjunction with the Python opencv module, to mark and record the coordinates and landmarks of the human body when obtaining real-time data. A block of code was written to capture real-time data from the webcam. These data include the various coordinates and landmarks gotten from an individual doing poses, using the mediapipe holistic solution, and cv2 library. The coordinates identified in every frame were read out to the csv file for the specific pose as we pose (class).

Feature Extraction

The acquired gesture videos are comprised of different postures like the normal human posture and four other threatening gestures like gestures of punch, fight, and elbow among others. The coordinates and landmarks of these postures were obtained using the medi pipeline holistic model. This helps in extracting the coordinates of these aforementioned postures and their landmark, therefore doing away with irrelevant features. These coordinates and landmarks were then exported as CSV files with the landmarks and coordinates as the attributes while they were classified based on their corresponding labels. This amounts to a total of 1342 coordinates and 2004 landmarks. These extracted features were further used as the input dataset to train the employed machine learning algorithms for gesture recognition.

Gesture Recognition

After the useful features have been extracted from the acquired video, the dataset was later processed for gesture recognition using four different machine learning algorithms namely: Linear Regression, Ridge Classifier, Random Forest and Gradient Boosting. These four classifiers were implemented on Google Colab. These classifiers made use of the Hold-out evaluation method with 70-30 splitting method. The tested model was eventually evaluated using accuracy, precision, f1score and recall as the evaluation metrics.

Results and Discussion

This section presents an overview of the results obtained from this research. The results were obtained from testing the trained model with webcam videos. The MediaPipe Holistic component integrates the pose, hand, and face tracking into a single, integrated solution. When including all three components, MediaPipe Holistic provides a unified topology for a ground-breaking 540+ key points (33 poses, 21 per hand and 468 facial landmarks).

Figures 3 (a-d) represent the confusion matrixes for the models: Gradient Boosting, Random Forest, Risge classifier and Linear Regression respectively, which show the number of positive and negative detections (predictions) made by the trained model. The horizontal axis (x–axis) represents the true values (ground truth), and the vertical axis (y – axis) represents the predicted values.

Experimentation Results

Table 1 presents the various results obtained after evaluating the four different machine-learning pipelines for gesture recognition. From the table, the Ridge Classifier has the best accuracy of 99.8%, the Random Forest Classifier comes in second, with 99% accuracy, and then the Gradient Boosting Classifier with 98.8% accuracy, and then the Linear Regression Model with an accuracy of 98.3%. Overall, the Ridge Classifier has the best result of the four pipelines.

A visualization of Table 1 is shown in Figure 4 above, with each pipeline represented with different colours: from left to right linear regression, ridge classifier, random forest classifier and gradient boosting classifier respectively, and each cluster representing each evaluation metric, showing their different results.

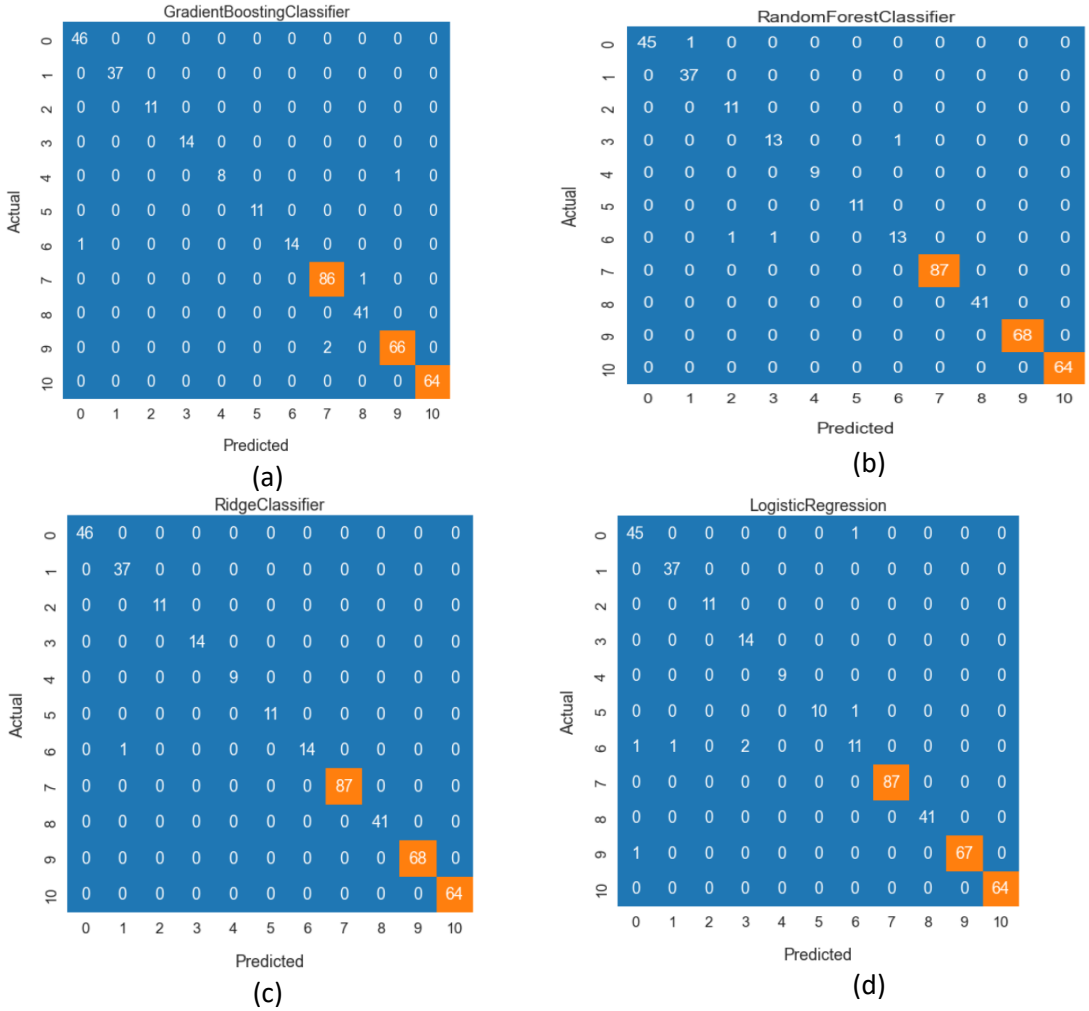

Figure 3: Confusion matrics obtained from the Experimentation

Table 1: Gesture recognition model training results

	Evaluation Metrics			
ML Pipeline	Accuracy	Precision	Recall	F1-score
Linear Regression	0.983	0.968	0.964	0.965
Ridge Classifier	0.998	0.998	0.994	0.996
Random Forest Classifier	0.990	0.977	0.979	0.978
Gradient Boosting Classifier	0.988	0.992	0.980	0.986

Real-Time system testing

Additional real-time testing was carried out after training the model, this was done to rate the performance of the system. The model was tested on live video from the webcam as shown in Figure 5. The results of the various tests are shown in the images below. The detection result of the model is depicted by the bounding boxes. The text on top of the boxes shows the type of gestures detected, and the score reflects the level of confidence of the detection.

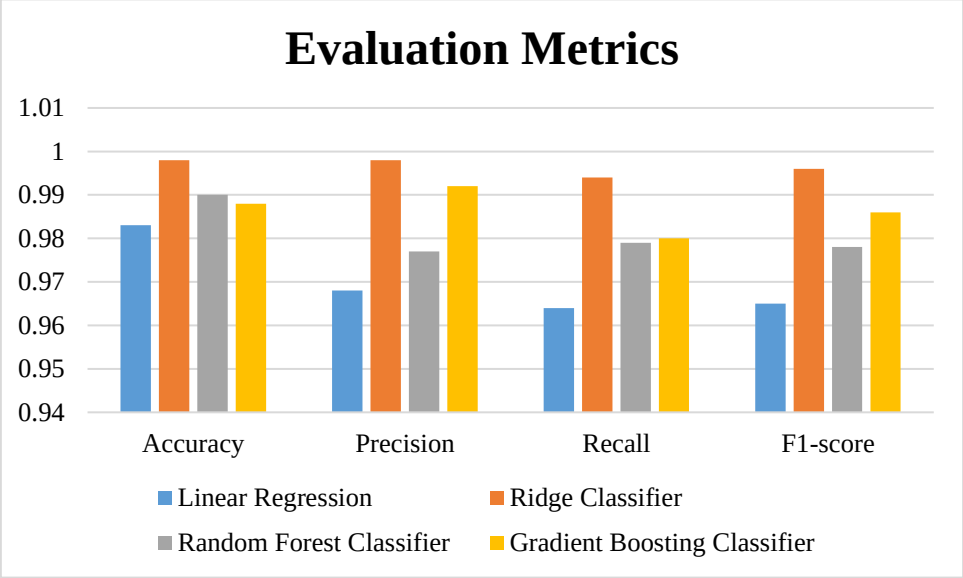

Figure 4: Chart showing the comparison of the different machine-learning pipelines for gesture recognition

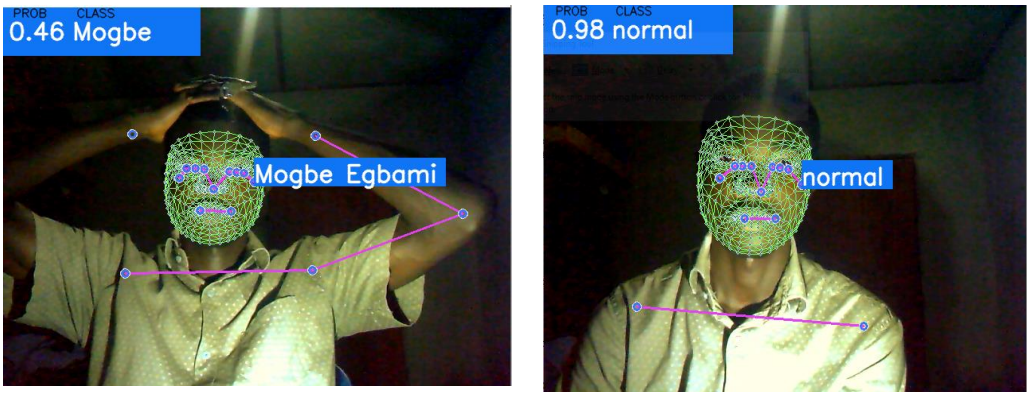

Figure 5: Images obtained from testing the model on real-time video via the webcam

CONCLUSION AND RECOMMENDATION

In conclusion, this research presents a solution in the form of an intelligent surveillance system for gesture recognition. The system's main objective is the recognition of odd and suspicious behaviours. This research leverages the capabilities of machine learning, and computer vision approaches to enhance the security and efficacy of video surveillance in various environments. The result of this project is a reliable and effective system that can detect unusual gestures in real-time by training the machine learning pipelines to recognize gestures of concern. The Media pipe holistic solution was used to obtain the data in real-time using a webcam. The machine learning pipelines: linear regression, Ridge Classifier, Random Forest Classifier, and Gradient Boosting Classifier, achieved an accuracy of 98.3%, 99.8%, 99%, 98.8%, a precision of 96.8%, 99.8%,

97,7%, 99.2%, a recall of 96.4%, 99.4%, 97.9%, 98%, and an F1-score of 96.5%, 99.6%, 97.8%, 98.6% respectively. Overall, the Ridge Classifier has the best result of the four pipelines, making it the best solution for the task.

This study highlights the significance of precise detection and proactive assessment. The system demonstrates adaptability in reducing security threats by using gesture recognition to spot deviations from expected behaviour. Security staff can take immediate action after receiving real-time alerts with contextual information, which increases the effectiveness of threat response. However, future works can be carried out to improve the gesture recognition capabilities in crowded environments, to properly analyse the crowd, and spot anomalies in crowd behaviour. Future work would focus on employing the acquired video gesture with the help of Computer Vision techniques rather than using vidoes' landmarks and coordinates for features extraction and later for gesture recognition.

Declaration of Competing Interest

No conflict of interest was declared by the authors.

REFERENCES

Ambati, L. S., & El-Gayar, O. (2021). "Human Activity Recognition: A Comparison of Machine Learning Approaches". Journal of the Midwest Association for Information Systems, 49-60.

Kim, J.-W., Choi, J.-Y., Ha, E.-J., & Choi, J.-H. (2023). Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model. Applied Sciences, Volume13(4).

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., . . . Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172v1.

Luo, C., Cheng, S., Xu, H. & Li, P. (2020). "Human Behavior Recognition model based on improved EfficientNet". Proceedings of the 8th International Conference on Information Technology and Quantitative Management. Procedia Computer Science, 199. pp. 369-376

Neha, K., Rutuja, D., & Vaibhav, K. (2022). Intelligent Video Surveillance System using Deep Learning. International Research Journal of Engineering and Technology (IRJET). Volume 9, Issue 5, 1616-1621.

Omodunbi, B. A., SOladoye, A. A., Okomba, N. S., Ayinla, M. O. and Odeyemi, C. S (2024) "Development of a medical condiction prediction model using natural language processing with K-nearest neighbour", FUOYE Journal of Engineering and Technology, 9(1), pp-25-32

Sahoo, J. P., Prakash, J. A., Pawel, P., & Samantray, S. (2022). Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network. Sensors, Volume 22, Issue 3.