Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 13, No.3 pp279–292July-September,2025

p-ISSN: 2663-628X
e-ISSN: 2663-6298

DEEP LEARNED FEATURE TECHNIQUE FOR HUMAN ACTION RECOGNITION IN THE MILITARY USING NEURAL NETWORK CLASSIFIER

Adeola O. Kolawole[1,*], Martins E. Irhebhude[1], and Philip O. Odion[1]

[1]Department of Computer Science, Faculty of Military Science and Interdisciplinary Studies, Nigerian Defence Academy, Kaduna Nigeria

*Corresponding author email: adeolakolawole@nda.edu.ng;

Received: 04 Mar. 2025 Accepted:28 May. 2025 Published:01 Jul. 2025 https://doi.org/10.25271/sjuoz.2025.13.3.1499

ABSTRACT:

Assessing military trainee in an obstacle crossing competition requires an instructor to go along with participants or be strategically placed. These assessments sometimes suffer from fatigue or biasedness on the part of instructors. There is the need to have a system that can easily recognize various human actions involved in obstacle crossing and also give a fair assessment of the whole process. In this paper, VGG16 model features with neural network classifier is used to recognize human actions in a military obstacle-crossing competition video sequence involving multiple participants performing different activities. The dataset used was captured locally during military trainees' obstacle-crossing exercises at a military training institution to achieve the objective. Images were segmented into background and foreground using a Grabcut-based segmentation algorithm. On the foreground masked images, features were extracted and used for classification with neural network. This method used the VGG16 model to automatically extract deep learned features at the max-pooling layer and the input presented to neural network classifier for classification into the various classes of human actions achieving 90% recognition accuracy which is at training time of 104.91secs. The accuracy obtained showed 3.6% performance improvement when compared to selected state-of-the-art model. The model also achieved 90.1% precision value and recall of 90.2%. Although many studies have focused on human recognition action recognition in several application areas, this study introduced a novel model for real time recognition of fifteen different classes of complex actions involving multiple participants during obstacle crossing competition in a military environment leveraging on the strength of deep learning and neural network classifier. This study will be of immense unbiased benefit to the military in the assessment of a trainee's performance during training exercises or competitions.

KEYWORDS: VGG16, Neural Network Classifier, Obstacle Crossing Exercise, Military Training, Deep Learning

1. INTRODUCTION

Human activity recognition is the process of recognizing human action that take place in a video sequence (Belmonte Fernandez, & Montoliu, 2020 ; Kolawole, Irhebhude, & Odion, 2025; Patel *et al*., 2020; Sansano). Techniques for human behaviour analysis can be through image based data (videos or images) or sensors which study motion data from wearables (Sansano, Montoliu, & Belmonte Fernandez, 2020). The video-based technologies involve extracting features from videos and images of human activity streams captured by cameras placed within an environment where humans are present, this approach gives an intuitive understanding of the complexities involved in the task (Shiraly, 2022). Human activities from input data are analyzed by machines for Human Action Recognition (HAR) and researchers have continued to improve HAR systems in different situations. HAR can be applied in; healthcare systems such as monitoring of patients, gesture recognition and also in the military operations (Luo, 2020).

HAR is a complex research field that involves recognition and prediction from video image samples consisting different human activities. The task of activity recognition is related to human actions present state, and action prediction envisages the future state. These two tasks are used in real-world applications; such as surveillance, human-computer interaction, autonomous vehicles, healthcare and video retrieval (Kong & Fu, 2022). Identifying specific human actions and movements from video or image data is useful for integration into technologies that use modern devices in real world application. Deep learning algorithms play an important role in processing videos or images for HAR; however, it is difficult to build one from the scratch as it comes with the difficulty in obtaining large amount of labelled data and the huge computational resources required. Transfer learning involving pre-trained model have been identified as solution to this challenge (Harahap *et al*., 2023).

The VGG16 CNN deep learning method is one of the most significant approaches for image recognition task (Latumakulita *et al*., 2022). An action, such as sitting or raising a hand, could be identified using a single frame (image) while interaction with one or more objects leads to greater complexity of actions such as jumping, running and swimming, as they require longer video frame sequence as such actions involve different physical body movements (Host & Ivasic-Kos, 2022).

* Corresponding author

Human actions can be categorized into four based on the involvement of body parts used; human gesture (movement of the body parts with specific meanings), action (it can be activity performed by an individual or combination of multiple gestures), interaction (involves performing activities by two actors i.e. human and objects) and group activity (complex combination of the three categories involving multiple participants and objects) (Manaf & Singh, 2021). These classes of actions are found in obstacle crossing exercises especially in military training institutions where several events involving humans and other objects take place with varying levels of complexity. Monitoring these activities manually is very difficult due to complexity of the exercise, hence the need to have an automated way of assessing these activities to ease the burden of identifying faults or errors during such exercise in the military environment.

This study recognizes different categories of human actions from locally gathered datasets using image classification and deep learning algorithms which can be beneficial in real time video analysis. The study by Kolawole, Irhebhude and Odion (2025) classified fifteen (15) different actions during an obstacle crossing competition in a military institution and achieved 86.5% recognition accuracy. An error margin of over 10% gives room for improvement. Therefore, this study builds on the effort made by using VGG16 extracted features to recognize human actions in military obstacle crossing competition with multi-layer perceptron classifier.

This study proposes using VGG16 CNN architecture for action recognition in cadet's obstacle crossing exercise. This research aims to contribute to learning in areas of developing a computer vision and deep learning techniques for human action recognition, especially in military activities.

The rest of the paper is structured as follows: the literature review and other concepts are discussed in section 2. The related works are discussed in section 3, the proposed methodology, dataset used, VGG16 feature extraction, and classification are discussed in section 4. Section 5 discusses the results of the implemented work. Performance comparison of the results obtained is discussed in section 6. Finally, the conclusions, recommendations and future work are discussed in section 7.

Literature Review:

This literature review explores existing research in human action recognition using traditional machine learning and deep learning models. It also discusses the conceptual view of area of study and review of related work.

Kong and Fu (2022) grouped action recognition into action representation and action classification. Representation converts series of videos to vectors, while classification will infer a labelled action from the vector. It plays an important role in recognition, but without a decent learning method a true representation result would not be achieved (Khowaja & Lee, 2020). Different real-world applications are made possible by action recognition, prediction systems, visual applications, entertainment, autonomous driving, video retrieval, and human computer interaction. Machine Learning or DL techniques can be used in training HAR models (Kolawole, Irhebhude, & Odion, 2025).

For handcrafted features, ML-based techniques can be utilized and the DL-based framework can be used for automatic feature extraction (Gupta *et al*., 2022). DL models have recorded great success in a variety of difficult research areas, including image recognition and natural language processing. The main advantage of DL is its ability to automatically learn representative features from massive amounts of data. As a result, this technology may be appropriate for human activity recognition. While some early attempts can be found in the literature, many difficult research problems such as detection accuracy, device heterogeneities, environmental changes, among others remain unsolved (Li *et al*., 2021).

Several deep learning algorithms such as CNN have been used for image classification; they consist of many layers including, convolutional, pooling and fully connected layers (Tripathi, 2021). Studies have shown that pre-trained networks such as; Visual Geometry Group (VGG16) (Simonyan & Zisserman, 2014), ResNet (Amos, & Irhebhude, Kolawole, 2023; Irhebhude, Kolawole, & Zubair, 2024) have been applied in recognition tasks and recorded promising results. According to some studies, the VGG16 pre-trained deep learning models have been employed for the extraction of deep learned features, it is considered to be one of the excellent vision model architectures (Thakur, 2024). The author further stated that the sixteen (16) in VGG16 refers to it having 16 layers that have weights with a large network of about 138 million parameters. This section gives a background knowledge on DL, VGG16, neural network classifier and related studies in the use of deep learning for human action recognition.

Deep Learning:

Deep learning model is fundamental to the proposed technique used in this study. The ability to learn high level features and capability to give high performance has made DL popular (Bento, 2021). DL uses artificial neural network (ANN) as their main structure, they differ from other algorithms because expert input is not required during feature design and engineering phase (Bento, 2021). DL algorithms take in dataset, learns its patterns and how to use features extracted on their own to represent the data. Subsequently, DL algorithm can combine different representations of each dataset with each one identifying a distinct pattern or characteristic into an abstract high level representation of the dataset (LeCun, Bengio, & Hinton, 2015). DNN, Hybrid Deep Learning (HDL), and transfer learning models are the three categories of DL methods used in HAR (Athavale *et al*., 2021; Deep & Zheng, 2019; Harahap *et al*., 2023; Li & Wang, 2022; Tufek *et al*., 2019). The transfer learning includes pre-trained DL models like VGG16, ResNet etc. and have been successfully applied in various classification studies such as activity recognition (Athavale *et al*., 2021; Harahap *et al*., 2023; Li & Wang, 2022), emotion recognition (Amos, & Irhebhude, Kolawole, 2023), pneumonia image classification (Jiang *et al*., 2021), etc. The development of DL can be separated into various categories in terms of algorithm and structure, and the CNN is the most utilized DL network type because of its ability to automatically detect important features without supervision (Alzubaidi *et al*., 2021). The VGG16 CNN architecture is briefly described in the next subsection.

Visual Geometry Group (Vgg16):

VGG16 also called VGGNet is a pre-trained deep learning model with sixteen layers developed by Simonyan and Zisserman (2014) at Oxford University, where it achieved an accuracy of 92.7% on the ImageNet dataset at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The VGG refers to the

Visual Geometry Group, while 16 refers to the 16 layers of the network including 13 convolutional layers and 3 fully connected layers. The uniqueness is in the use of 16 layers that have weights as opposed to relying on several hyper parameters, each layer works to process image information incrementally and improve the accuracy of its predictions (Thakur, 2024). VGG16 is an improvement on AlexNet with 3x3 convolutional kernels and 2x2 pooling layers, smaller convolution layers was added to deepen the network architecture and improve feature learning (Jiang *et al*., 2021). The ability of the VGG model to capture high-level features in complex tasks gives it a higher performance over other models, especially in object detection and recognition. The VGG design focused on increasing network depth, while using simple and uniform convolutional layers in image classification problems (Grigoryan, 2023).

A VGG16 comprises of different blocks; input, convolutional layers, pooling layers, fully connected layers and softmax as shown in Figure 1, the first two blocks contain two convolutional layers each with 64 and 128 filters respectively, followed by ReLU max pooling layers while the last three blocks have three convolutional layers each with 256, 512 and 512 filters respectively followed by ReLU and a max pooling layer. It follows the arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end, it has 3 fully connected layers (FC) followed by a SoftMax for output. The network is massive with about 138 million approximate parameters (Grigoryan, 2023).

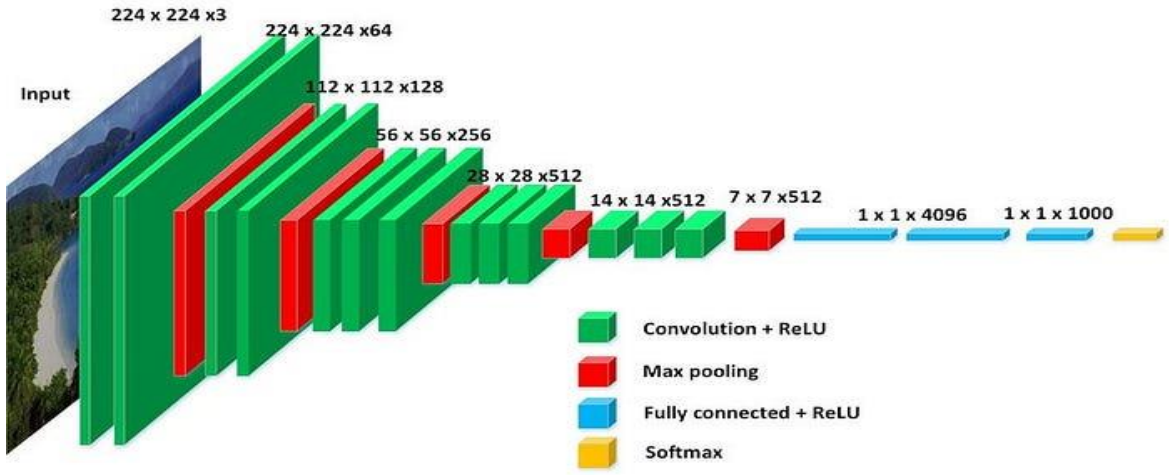

Figure 1:VGG16 Architecture (Arif, 2023)

(a) Input Layer – The input layer serves as the first layer of the VGG16 architecture. This layer has an input size of $224 \times 224 \times 3$ and feeds the input data to the network while the output is passed to the convolutional layer.

(b) Convolutional Layer – The convolutional layer captures a wide range of features at different levels for richer representation. It uses a filter with smaller 3x3 receptive field window size to detect local features and convolution stride of 1 pixel, rather than large fields in the first convolutional layer in a bid to improve accuracy (Qassim, Verma, & Feinzimer, 2018). The number of filters increases as it moves deeper into the network. There are five sets of convolutional layers and max pooling layer. This layer performs stride, padding and activation functions using ReLU the input image passes through the first convolutional layer with ReLU and the pooling layer having 64 filters and input dimension $224 \times 224 \times 64$. The input image further moves to the second convolution layer with a filter size of 128 and input dimension of $112 \times 112 \times 128$ for deeper feature extraction. More features are extracted at the third, fourth and fifth blocks with filter size $56 \times 56 \times 256$, $28 \times 28 \times 512$ and $14 \times 14 \times 512$ respectively.

In CNN, the stride value is a fundamental parameter that represents the number of pixel size by which the filter moves across the entire area of an input image during convolution operation. It controls how the convolutional filters interact with input image and affect the size of the output feature map. An adjustment in the stride can affect outputs size, computational efficiency, and models' performance. ReLU activation function introduces nonlinearity to the extraction process, the VGG16 network's hidden layers use ReLU nonlinearity instead of Local Response Normalization like AlexNet and aids in modelling complex interactions within data (Simonyan & Zisserman, 2014). The hidden layer contains neurons known as nodes; each node in each layer interacts with data and connects to each node in the following layer, passing on information learned about the data.

(c) Max-Pooling Layer – The Max-pooling layers are utilized for spatial down-sampling feature maps and reducing dimensions of the features while preserving important features, making the network more robust and invariant. The max-pooling layer consists of kernel and astride of 2 which helps to reduce overfitting (Athavale *et al*., 2021).

(d) Fully Connected Layer - Three fully connected layers are included in VGG16; they are used for classification tasks based on features extracted from the images. The first two layers have 4096 channels each while the third layer has 10000 channels, and the final layer is SoftMax (Simonyan & Zisserman, 2014)

(e) SoftMax Layer – This layer is the final layer of the architecture and it outputs the prediction of the classes.

Transfer learning techniques such as the VGG16 have been applied in many real-world applications. Pardede *et al.* (2021) proposed a transfer learning architecture for fruit ripeness detection using VGG16, replacing the upper layer of the architecture with an MLP block used as classifier and the SoftMax activation as output layer. The experimental result

demonstrates the strength of the VGG16 and MLP achieving good performance accuracy of 90%.

Multi-layer Perceptron (MLP):

Guha *et al*. (2021) proposed a feature selection model on multiple extracted features using MLP classifier and achieved best accuracy of 95.19% average classification. The MLP is known for its versatility and capability to distinguish between multiple classes of activities (Dutta, 2024). This characteristic makes it a wide choice for classification task.

MLP consists of input and output layers, with one or more hidden layers. Through training, MLP have the capability of learning, set of training data which is made up of several input vectors needed for training. The MLP is seen as feedforward algorithm as it is continually fed with training data, during training the network's weight are combined with input in a weighted sum and subjected to activation function until intended input-output mapping is achieved (Bento, 2021). The number of layers and neurons are referred to as the hyper parameters of a neural network which needs tuning, adjustment of weight is done through backpropagation. The notation in MLP is given as follows (Banoula, 2023):

Given $i = 0,1,, \ldots n$, where n is number of inputs, weight of the neuron are the quantities represented as w_i the input (x_i) correspond to the features and output (y) corresponds to their predictive class.

$$\sum_{i=0}^{n} w_i x_i = f(z) = y \qquad (1)$$

The weighting steps involve; multiplying each input feature by its weight $(x_i w_i)$,

in the second step they are added together as $(x_0 w_0 + x_1 w_1 + \cdots + x_n w_n)$. The next step involves applying an activation function $f(z)$ to the sum and produces an output (y) as shown in equation 1 (Taud & Mas, 2018).

Review of Related Works:

Traditional machine learning and deep learning techniques have been applied for human action recognition resulting in various degrees of recognition accuracies and performance (Dang *et al*., 2020). Ankita *et al*. (2021) developed a hybrid CNN-LSTM lightweight deep learning framework for HAR and achieved a prediction accuracy of 97.89% on the University of California Human Action Recognition (UCI-HAR) dataset. Further research on combining different deep learning models on various datasets was suggested by the authors to achieve a more accurate result.

Nafea *et al*. (2021) introduced a CNN model with varying kernel dimensions along with Bidirectional Long Short-Term Memory Network (BiLSTM) to extract spatial and temporal features from sensor data for HAR. Filters with different kernel sizes were utilized in each layer to extract useful information from the CNN. The output of the CNN and BiLSTM were combined to give the final output used for recognition, UCI dataset (consisting of six activities namely; sit, walk, walk up, walk down, stand, and laying down) and WISDM dataset (consisting of the following activities; standing, sitting, walking, upstairs, downstairs, and jogging) were used for experiment on the CNN-BiLSTM model and achieved accuracy values of 97.05% and 98.53% respectively. The researchers suggested further examination on contrasting features extracted by CNN-BiLSTM against handcrafted features.

Muhammad *et al*. (2021) proposed the BiLSTM mechanism with Dilated Convolutional Neural Network (DCNN) for a video based human action recognition system in order to address the issue of distinction between visual and temporal features in video frames. Discriminative features were extracted by the DCNN with residual blocks and then fed to the BiLSTM to learn long term dependencies. SoftMax was utilized to improve loss function and increase accuracy. Three benchmark datasets were used for evaluation; UCF101, UCF sports and J-HMDB achieving recognition rates of 98.3%, 99.1% and 80.2% respectively.

A multi-view action recognition system was proposed by Amin *et al*. (2021) using the VGG19 CNN model for feature extraction and conflux LSTMs network to learn multi view patterns. The actions were classified using SoftMax classifier. The Northwestern-UCLA multi view action 3D dataset (Wang *et al*., 2014) and Multi-camera action dataset (MCAD) (Li *et al*., 2017) were used in the experiment with the proposed conflux LSTM model, and accuracy values of 88.9% and 86.9% were reported respectively.

HAR plays a crucial role in pattern recognition with numerous applications, all HAR systems can be divided into two categories; The first uses sensor-based technology for activity recognition. This technology means wearable sensors are attached to the body so that the system can collect information in the form of velocity, frequency, acceleration, patterns etc. in movements made by individuals. The second approach is vision-based, which involves recognizing human activity from video or images. In this scenario, the system collects data using a camera to identify actions (Isakava, 2022). Vision-based approach to human action recognition uses images, videos or camera feeds. For video-based HAR three types of data can be extracted; visual data (such as contour, texture, colour and spatial features are extracted from individuals and objects in a frame), Temporal data (motion patterns of people, objects and cameras) and text data (Shiraly, 2022).

Liang, Lu and Yan (2022) used the CNN-LSTM model to extract spatiotemporal information based on the KTH (Caputo& Schuldt, Laptev, 2004) dataset for HAR. Six categories of actions were selected from the KHT video datasets (walking, jogging, running, clapping, boxing and waving), and the experimental result was compared to three other models. LSTM+CNN model had the best performance with 89.0% accuracy while the CNN, KNN and Spatio temporal interest point with KNN (STIP+KNN) models recorded 86.00%, 83.00%, and 84.0% respectively. This showed the ability of all the models to identify human actions but the need for improvement in achieving higher accuracy in multiple and continuous actions, especially in complicated environments. Different studies that used deep learning approach are reported here.

Hayat *et al*. (2022) developed a framework for monitoring activities. of elderly people in outdoor and indoor environments using data collected from gyroscope and accelerometer in smart phones. Activities such as; sitting, walking, going upstairs, going downstairs, standing, and lying were included in the dataset. The UCI dataset was used for the study and it was evaluated using different machine learning and deep learning techniques namely; RF, KNN, SVM, ANN and LSTM achieving overall recognition performances of 82.68%, 85.38%, 87.90%, 91.23% and 94.53%

respectively using 2-fold cross validation. With 10-fold cross validation, RF, KNN, SVM, ANN and LSTM gave accuracies of 85.63%, 87.18%, 89.08%, 92.25% and 95.05% respectively. The authors concluded that the results yielded significant performances as it is challenging to build HAR due to wide varieties of activities and their similarity in nature. Recommendation was made on improving the model by validating technique on larger datasets that contains more activities.

In a study by Patil *et al*. (2022), a deep learning model was developed to detect and classify human actions in restricted military areas for security purpose. Video images were captured with the use of webcam mounted to track human activities within the restricted environment. Normalization techniques, denoising, resizing and segmentation were done as preprocessing stage after which CNN was used for automatic feature extraction and LSTM for detection and the classification of human behaviour into legal and illegal activities. The authors did not report recognition accuracy achieved but identified the need to consider complex actions for further research and also improvement in the accuracy of the recognition system.

Keshinro, Seong and Yi (2022) employed two deep learning algorithms; ConvLSTM and Long-term Recurrent Convolutional Network (LRCN) for predicting human intention. The UTKinect-Action3D dataset (Xia, Chen, & Aggarwal, 2012) was used for the experiment. The dataset contained RGB images from 10 participants who performed the following actions; standing, picking, walking, throwing, pushing, waving, clapping hands and carry actions. The authors trained the proposed model by selecting four actions (pick, throw, carry and wave) from the dataset and achieved 74% accuracy on the ConvLSTM while the LRCN gave a poor accuracy of 25%.

Putra, Shima and Shimatani (2022) presented a Deep Neural Network (DNN) and shared weight techniques comprising of pre-trained CNN, attention layer, RNN and softmax layers for classifying human activity from multiple viewpoints. Multi-view images from IXMAS dataset (Weinland, Ronfard, & Boyer, 2006) and the i3DPost dataset (Kavi *et al*., 2016) were used in conducting experiments and recorded promising recognition accuracy of 97.2% and 96.87% respectively. The authors recommended further improvement in the model's performance by involving more subjects and considering evaluation with datasets consisting of more complex actions in different situations.

Qi *et al*. (2022) developed a smartphone based DCNN framework for HAR and automatic labelling of images. Depth vision data and IMU signals were collected using Microsoft Kinect camera and smartphone, the dataset comprised of 12 daily human activities. To improve accuracy, Kalman filter, noise removal and calibration were used as preprocessing steps. The proposed model achieved 93.89% accuracy.

Jain *et al*. (2022) proposed HAR hyper parameter deep learning model (HAR-HPTDL) which utilized a bidirectional LSTM model as feature extractor, a sparrow search algorithm to tune hyper parameters and a softmax layer for classification of human activities. To improve the overall performance of the HAR system, an entropy-based feature reduction was combined with Chi square-based feature selection. The experiment was conducted on the HMDB51, UCF101, UCF11, and IXMAS datasets with the proposed HAR-HPTDL model achieving recognition accuracy of 99.68%, 98.15%, 94.87%, 65.98% on the datasets respectively. The findings revealed the ability of the model to perform effectively on difficult tasks; however, it required significant amount of time and resources to achieve the success. Future improvement of the model in terms of inclusion of RGB and depth data for HAR was suggested.

Huang *et al*. (2022) noted CNN models have many invalid filters that affect the performance and contributes less to outputs. The authors proposed a filter activation model to reactivate useful filters that uses one network instead of multiple networks like in ensemble learning for HAR to boost accuracy. Gaussian noise activation was used to remove noise and entropy-based measure to select the useful filters. Experiments were performed on UCI-HAR, OPPO, UniMiB-SHAR, PAMAP2, WISDM, and USC-HAD datasets which gave accuracy values of 97.18%, 81.36%, 77.47%, 92.18%, 99.02% and 99.54% respectively.

Jaén-Vargas *et al*. (2022) proposed a DL approach for HAR with the introduction of sliding windows as a feature extraction technique in other to find the optimal size that will reduce processing cost and give the most accurate result for activity classification. In evaluating the performance of the model, experimental data comprised samples of three common daily movement activities (walking, sit-to-stand and squatting) from IMU and MOCAP dataset that were used on four different DL models; DNN, CNN, LSTM, and CNN-LSTM. Two sets of windows with varied lengths were used; short sliding window size of 5,10,15,20,25 frames and long sizes of 50,75, 100 and 200 frames to compare the performance in terms of accuracy and F1-score. The authors concluded the LSTM and CNN-LSTM performed better with F1-score above 80% and both achieving 99% accuracy on IMU data with 20 frames. Similarly, the LSTM and CNN-LSTM achieved high accuracy of 99% and 98.88% respectively on the MOCAP dataset utilizing windows from 20 frames. A further study that considers a greater variety of activities and complex HAR problems was suggested.

In addressing some existing challenges in vision based HAR, Poulose, Kim and Han (2022) proposed a Human Image Threshing (HIT) machine-based activity recognition using image dataset generated from smartphone camera, an IMU sensor and a stretch sensor. The dataset comprised of 9 activities; sitting, standing, laying, walking, push up, dancing, sit-up, running, and jumping. Masked RCNN was utilized for human body detection while facial image threshing (FIT) was employed to crop and resize images. The proposed HIT model considered four transfer learning models; VGG, Inception, ResNet and EfficientNet architecture. The ResNet achieved highest accuracy of 98.53% while VGG, Inception and EfficientNet recorded 96.38%, 93.18% and 89.94% respectively.

Dahou *et al*. (2022) presented a modified optimization algorithm called Binary Arithmetic Optimization Algorithm (BAOA) to improve the performance of HAR. The BAOA was used for feature selection CNN was applied to learn and extract features from the input data and SVM was adopted for classification based on the different activities. An experiment was conducted using the proposed combination of BAOA and CNN models on UCI-HAR, WISDM-HAR, and KU-HAR public datasets and achieved 95.23%, 99.5% and 96.8% accuracy respectively.

Azzag, Zeroual and Ladjailia (2022) presented CNN technique with the aim of achieving good result for human action recognition from real time videos. The model was based on static camera and frame by frame classification. The CNN model

architecture was used for training and classification using publicly available Weizmann dataset consisting of 10 classes of activities. The proposed method gave 84.44% accuracy value. The authors identified further research areas on having systems that can work with movement cameras, recognize and/or predict next movement of a person, and detect object and movement at the same time.

Ding, Abdel-Basset and Mohamed (2023) introduced HAR-DeepConvLG, a hybrid deep learning model for recognizing activities and solving HAR challenges in IoT applications. To improve classification accuracy while classifying human activities, the model included three convolutional layers and the addition of a squeezing and excitation block (SE) that learned and extracted spatial representation features from the input sensor data. A fusion of three parallel routes which were outputs from the convolution layer were fed into the LSTM, and GRU layer was used to extract temporal representation features. The performance of the proposed model was assessed through four commonly used HAR datasets and compared to other existing DL models. The experiment revealed that the HAR-DeepConvLG outperformed the other existing models with accuracy values of 97.52%, 98.48%, 97.85%, and 98.55% on the UCI-HAR, WISDM, PAMP2, and University of Southern California Human Activity Dataset (USC-HAD) datasets, respectively. Further study into the development of an online deep learning model and usage of large amount of multi-variant data was recommended.

Bi *et al*. (2023) discovered the issue of annotation, especially in real-world activity as a challenge to HAR tasks as it is time-consuming and requires specific knowledge and skills. Therefore, the authors attempted to address this issue by proposing a framework that integrates active learning and semi-supervised learning to extract the most useful information from samples for annotation and also take advantage of information from unlabeled instances. This was done in other to achieve a reduction in annotation cost and reduce issues related to insufficient labelled data. The proposed model was evaluated on three benchmark datasets; PAMAP2, USCHAD and UCIHAR and achieved F1 values of 0.76,0.45 and 0.91 respectively.

Vrskova *et al*. (2023) proposed a three-dimensional neural network (3DCNN) model with convolutional long short-term memory (ConvLSTM) layers to recognize human activity recognition from videos. The authors conducted experiment on three datasets; LoDVP Abnormal Activities dataset (Vrskova *et al*., 2022) comprising eleven classes, UCF50 dataset which contains 50 action categories and MOD20 dataset (Perera *et al*., 2020) which has 20 classes. Confusion matrix, recall, precision, accuracy and F1 score values were used to validate the performance of the model. The LoDVP, MOD20 and UCF20 attained precision rates of 89.12%, 83.89% and 87.76% respectively while recall values were 87.69%, 81.09% and 88.63% respectively and F1 scores were 89.32%, 81.57% and 87.84% respectively. The overall accuracy for UCF50 was 87.78% while LoDVP and MOD20 datasets achieved 93.41% and 78.21% respectively. The overall results demonstrate the success rate of utilizing the proposed neural network to classify abnormal human behaviour in public places from videos. However, the authors recommend enhancing the performance of the model by incorporating more sensor data and investigating other related tasks.

Raj and Kos (2023) classified human activities into three types based on the position of the human body; static, dynamic, and postural conditions. Standing, sleeping, and sitting were grouped as static activities because the human bodies remain steady during data collection, Running and walking were classed as dynamic activities, because human bodies do not remain steady during data acquisition, the postural transition activities are in between the static and dynamic. The authors demonstrated the ability of DL to improve activity recognition by presenting a 2D-CNN model for HAR using the WISDM dataset which achieved overall accuracy of 97.2%.

In a similar study, Garcia-Gonzalez *et al*. (2023) explored the CNN+LSTM model for HAR on real-life activities, gathering four categories of activities (inactive, active, walking, and driving) with the use of smartphone sensors and achieved an accuracy of 94.80%. The authors suggested further research to determine the best models for real-life datasets.

Dhiravidachelvi *et al*. (2023) designed an Intelligent Hyper parameter Tuned Deep Learning-based HAR (IHPTDL-HAR) which incorporates a deep learning based DBN model for recognizing human actions in the healthcare environment. The authors applied a DL-based DBN to recognize users' activities, Hierarchical Clustering (HC) for outlier detection and the Harris Hawks Optimization (HHO) algorithm was used for hyper parameter tuning. The experimental analysis was carried out on a localized dataset consisting of seven classes; FLD (Falling, lying down), LDSD (Lying down, sitting down), LOAF (Lying, on all fours), LS (Lying, sitting), SUFLSUFS (Standing up from lying, standing up from sitting), SUFSSUFSOG (Standing up from sitting, standing up from sitting on the ground) and FSD (Falling, sitting down). An accuracy value of 98.53% was reported when training and testing data split in a ratio 80:20. The authors suggested future study on implementing the proposed model in real-time smart hospital health care.

In order to recognize daily behaviour in the home environment, Li *et al*. (2023) proposed HAR algorithm based on wide time-domain convolutional neural network and multi-environment sensor (HAR_WCNN). Experiment was performed using the CASAS dataset by Cook *et al*. (2012) which contained information from daily activities from five families; Cairo, Milan, Kyoto7, Kyoto8 and Kyoto11. The study reported accuracies of 91.99%, 95.35%, 86.68%, 97.08% and 90.27% respectively demonstrating the model's excellent performance in identifying and classifying behaviour.

According to Harahap *et al*. (2023), using a pre-trained deep learning model for HAR eliminated the difficulties of developing a deep learning algorithm from scratch. Emphasis was laid on conducting more research to determine the advantages of the pre-trained deep learning model. The proposed pre-trained VGG16 model was used for the classification of two types of human activities; sitting and running, which were referred to as static and dynamic activities. A public dataset consisting of 1680 images and local dataset containing 100 images were used for the experiment. The focus was to achieve high accuracy in classifying human activities from data recorded by camera. The experiment yielded accuracy rates of 98.8% and 97% on public and local datasets respectively.

Surek *et al*. (2023) utilized a deep learning model consisting of ResNet and a hybrid 2D-Vision transformer architecture (ViT) with LSTM for HAR on RGB videos. The ViT splits the input images into patches which were fed into a transformer as sequence of patches. The proposed approach was applied on publicly available HMDB51 dataset using accuracy as

performance evaluation measure. The study result showed 96.7% accuracy on ResNet model while the ViT architecture reported low accuracy value of 41.0%. However, for improved performance the authors suggested the use of ensemble models on more complicated datasets.

Salunke *et al*. (2023) developed a system for recognizing human activities and detect threats in a military restricted area, a web-based google teachable machine platform was used to train the model and classify activities. Dataset used for the experiment was obtained from live web camera, the system was also designed to give an alert if suspicious activities were found within restricted areas. Authors did not clearly state the number of classes of activities considered to be legal or illegal. It was concluded that the proposed method achieved better results than other activity detection methods although performance evaluation metrics used and accuracy achieved from the model were not clearly stated and further improvement on accuracy value of activity recognition was recommended. Similarly, Kolawole, Irhebhude and Odion (2025) used Histogram of Oriented Gradient (HOG) and Region feature descriptors (HOGReG) to recognize 15 different actions in a military obstacle crossing competition and achieved 86.4% recognition accuracy. The authors further recommended use of a more robust feature extraction technique, and the introduction of a deep learning model to further improve the accuracy.

Studies reviewed showed that traditional and deep learning techniques were adopted for the prediction of human action in and outside military environments. Tradition methods used feature techniques like HOG, region among others, while deep learning techniques adopted methods like LSTM, VGG16 and others. Although, the studies by Salunke *et al*. (2023) and Kolawole, Irhebhude and Odion (2025) were conducted in a military environment, there is room for improvement in terms of accuracy and inclusion of multiple human actions in specific domain. This research builds on the efforts made in Kolawole, Irhebhude and Odion (2025), by introducing a hybrid model that uses VGG16 to extract features with MLP classifier for the recognition of human actions in an obstacle crossing exercises.

Proposed Methodology:

The developed model used VGG16 pre-trained model as feature extraction to handle classification tasks and predict output for the different human actions in obstacle crossing competition in a military environment. The proposed technique involves several steps to extract features from images of cadets performing different obstacle crossing actions. The video samples were framed into images and used as input for the experiment; For this research, the proposed methodology as shown in Figure 3 consists of different procedures to ensure an efficient description of human actions.

segmentation was done using image segmenter app in MATLAB. Several features were extracted from the images before classification and final recognition of various actions.

The pseudocode explaining the proposed model for human action recognition in obstacle crossing competition is described as follows;

Step 1: Capture videos of the military obstacle crossing competition.
Step 2: Extract frames from the video samples.
Step 3: Select frames containing action of interest.
Step 4: Segment the selected frame into foreground and background
Step 5: Select the masked images of the segmented foreground actions
Step 6: Modify the VGG16 transfer learning model and extract features from the maxpooling layer.
Step 7: Label the selected features
Step 8: Use the extracted features as input for MLP classifier.
Step 9: Evaluate the performance of the experiment.

Figure 2 shows the step-by-step flowchart of the experimental process, beginning with video input of obstacle-crossing activities dataset, followed by framing them into images, segmentation and feature extraction process for easier classification by the MLP classifier and finally recognition of different classes of actions.

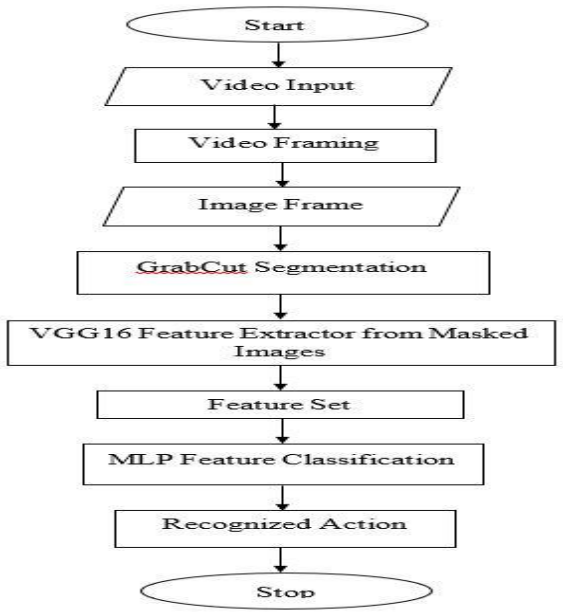

Figure 2: Flowchart of Experimental Pipelin

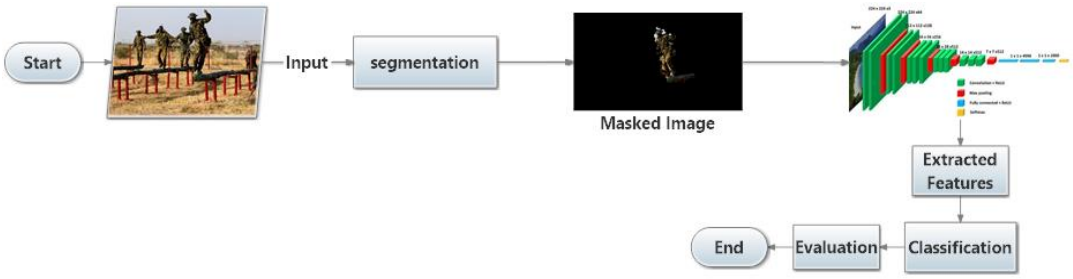

Figure 3: Proposed Methodology

Dataset:

The dataset utilized in this study is the one used by Kolawole, Irhebhude and Odion (2025) for action recognition in a military obstacle crossing competition. The captured videos were recorded with the use of Mavic 2 Enterprise drone, with 24GB memory storage onboard as reported. The dataset consists of 300 video files of cadets performing different activities that take place during obstacle crossing. The video samples are of dimension 1280×1720-pixel, frame rate of 30 frames per second (fps) and video length between 180 to 240 seconds. The video samples were resized from 1280×1720 to 448×252 due to the large video dimension while maintaining aspect ratio, this was achieved by framing the samples with video framing algorithm run on matrix laboratory software MATLAB R2023a. The resizing was necessary because of the limited computing resources and for easy framing. The collected dataset were 15 different activities namely; clear jump, barbed wire crawling, 6/ft wall climbing, scrabble net, hand or monkey bridge crossing, Tarzan rope, 9/7ft ditch, tunnel, Niger bridge, balancing, rough and tumbling, high wall with ladder, high wall tyre ladder minefield and horizontal and vertical wall. Each activity class contains 500 images totaling 7500 images used for the experiment.

Input Image/Segmentation:

Segmentation was carried out as reported in Kolawole, Irhebhude and Odion (2025). Some selected samples of input images from all the classes of activities with the actions performed are shown in Figure 4. The segmentation resulted in masked and black/white images. The masked images as shown in Figure 4 with their labelled classes of activities were used as input for further processing.

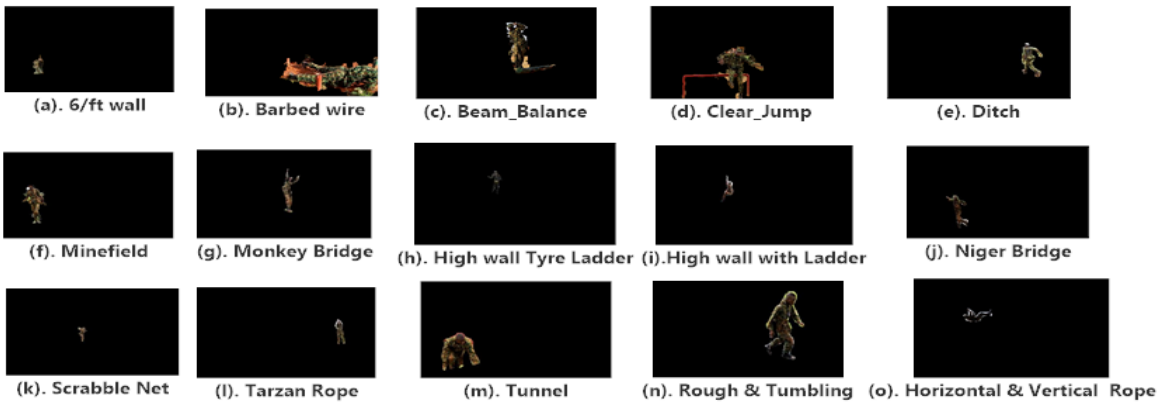

Figure 2: Sample Images from classes of Obstacle crossing activities

Figure 4 shows samples of segmented masked images of cadets performing the fifteen different classes of activities with their corresponding actions. The actions taken during the obstacle crossing activity for each class is described in Table 1 as reported by Kolawole, Irhebhude and Odion (2025).

Table 1: Obstacle Crossing Activities and Actions

Serial	Name of Obstacle	Actions Taken
	6/ft wall	Kick, hold and lie flat on the wall before dropping off
	Barbed wire	Crawling with toes and elbow under the barbed wire.
	Beam Balance	Mount log with two hands open horizontally
	Clear jump	Jumping over the bar without body contact
	Ditch	Jump over the ditch
	Mine field	Tip toe through empty tyres
	Monkey bridge	Jump and grab the bars
	High wall tyre ladder	Climb tyre ramp and jump down on both legs
	High wall with ladder	Using the vertical rope to get up the tower and rappel down
	Niger Bridge	Cross in chain like formation
	Scrabble net	Climbing up the scrabble net with hands on the vertical part and legs on horizontal part
	Tarzan rope	Holding rope firmly, swing across, jump and land
	Tunnel	Crawl inside the tunnel using skills applied in barbed wire
	Rough & tumbling	Firmly climb tyre ramp till the other side
	Horizontal and vertical rope	Mount the vertical rope with aid of legs, hold horizontal rope with both hands and legs moving to the end of horizontal rope, hold firm the vertical rope and drop down on both legs

Feature Extraction:

The CNN model is useful in extracting complex information from input images and effective in image classification (Jnagal, 2018). The proposed methodology employed the use of VGG16 model to extract deep-learned features automatically. The segmented foreground mask which served as the input image was presented to the VGG16 pre-trained model. The VGG16 features were extracted at the max pooling layer before the fully connected layer yields a feature length of 4096.

In this study, the processed input image of size 224×224 was labelled according to different actions giving a vector of 15 classes as shown in equation 2.

$$y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{14} \end{bmatrix} \qquad (2)$$

where Ys are the class labels.

Classification:

The MLP classifier was used for the classification. It maps the input data sets to a set of given output. The already preprocessed data from the previous blocks would be trained on the MLP network. MLP classifier achieved a good recognition accuracy of 95.19% in Guha *et al*. (2021) on multiple extracted features. According to Dutta (2024), the characteristics of the versatility and capability of the MLP classifier to distinguish between multiple classes of activities make it a wide choice for the classification task. Different parameters can be used like hidden layer size, activation function, number of epochs, and algorithm for weight optimization node.

In this study, the default hyper parameters values were used as shown in Table 2. These parameters achieved a satisfactory result without additional tuning.

Table 2: Selected Hyperparameter

Hyper parameter	Name
Preset	Wide Neural Network
Number of fully connected layer	1
First layer size	100
Activation	ReLU
Iteration Limit	100
Regularization strength (Lambda)	0
Standardized data	Yes

The performance of the algorithm was measured using accuracy, confusion matrix and Receiver Operating Characteristic (ROC) curve.

Performance Evaluation metrics:

For easy view and evaluation of the performance of the experiment conducted, the following metrics were used to study the training; accuracy, confusion matrix, receiver operating characteristics area curve (ROC), true positive rate, false positive rate, positive predictive rate and false discovery rate were used because of their evaluation efficiency (Abdullahi, & Irhebhude, Kolawole, 2021; Goma, & Irhebhude, Kolawole, 2021).

2. RESULTS AND DISCUSSION

The proposed methodology used VGG16 feature descriptors with 70% of the dataset was used for training while 30% was used for testing.

A total of 4096 features extracted gave an accuracy of 90% and a training time of 31.9755secs. The results are shown in the confusion matrix and ROC curve in Figures 5-8.

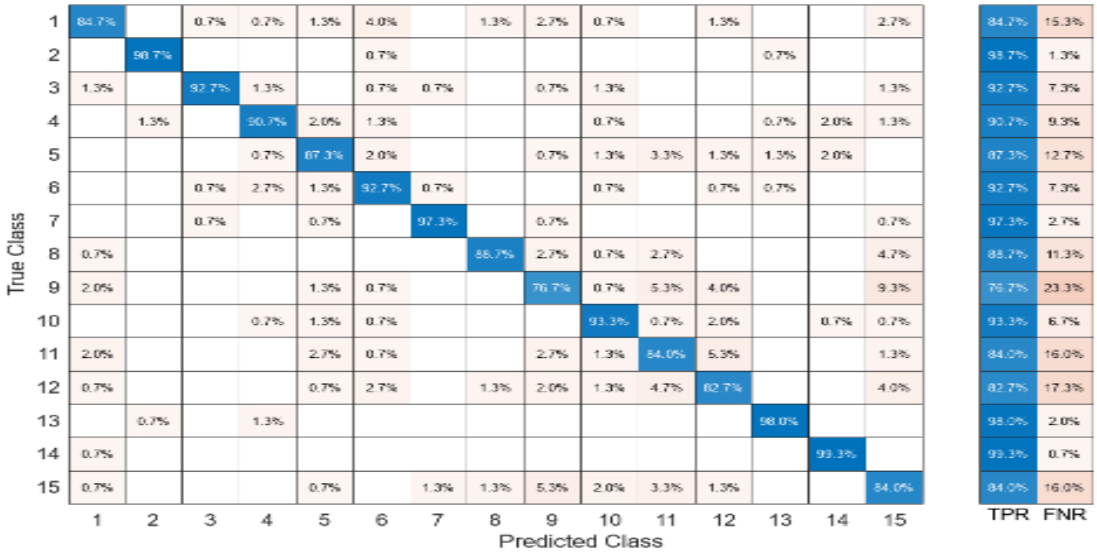

Figure 5: Confusion Matrix with Number of Observation using VGG16 Descriptor

The confusion matrix in Figure 5 shows the number of observations for each class of actions, class 14 recorded the highest number of correct prediction (149) followed by classes 2 and 13 which stood at 148 and 147 respectively. Other classes had varying number of correct predictions and some few wrongly predicted instances. Class 9 had the least number of correct observations with high number of actions placed in the wrong classes. The overall performance of the matrix shows how well the classifier understood the distinction between the different classes of actions and it is clear that the VGG16 has better performance in handling actions in minefield, barbed wire and high wall activities (these actions include; tiptoeing, crawling, climbing and jumping) compared to other classes as a result of the similar action movement in these classes. Further experiments on TPR and FNR are reported in Figure 6.

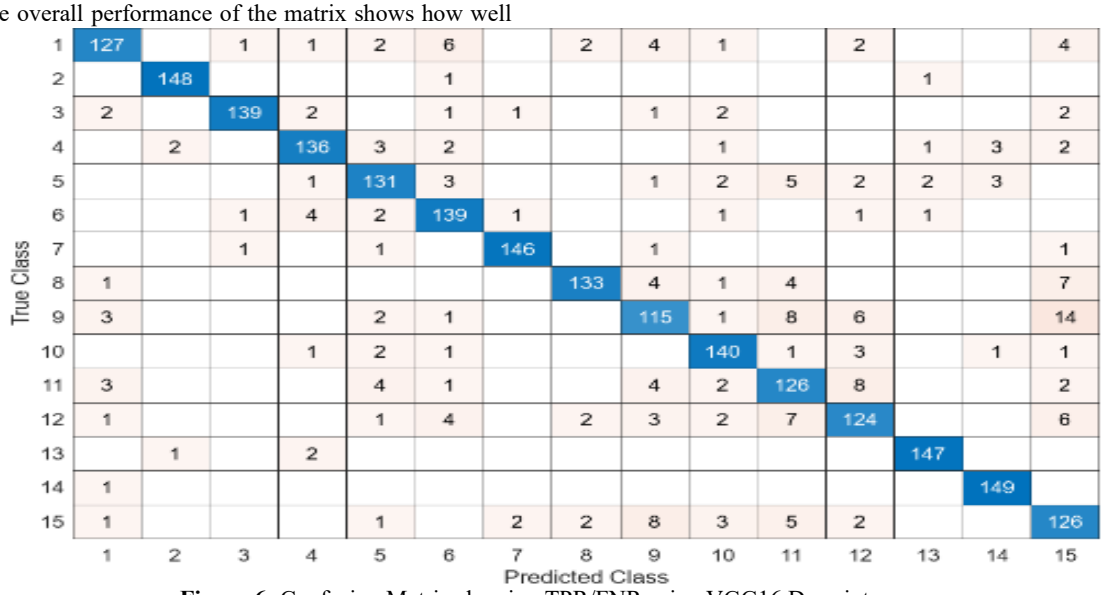

Figure 6: Confusion Matrix showing TPR/FNR using VGG16 Descriptor

Figure 6 shows the confusion matrix of TPR and FNR rates for different classes of human actions recognized after running the experiment. With a TPR value of 99.3% as seen in Figure 6, class 14 recorded the highest percentage of true positive classification which is in line with the number of observations reported from Figure 5. The highest percentage of incorrect prediction was recorded by class 9 with a FNR value of 23.3% which shows a high rate of misclassification spread across other actions. The PPV and FDR are also reported in Figure 7.

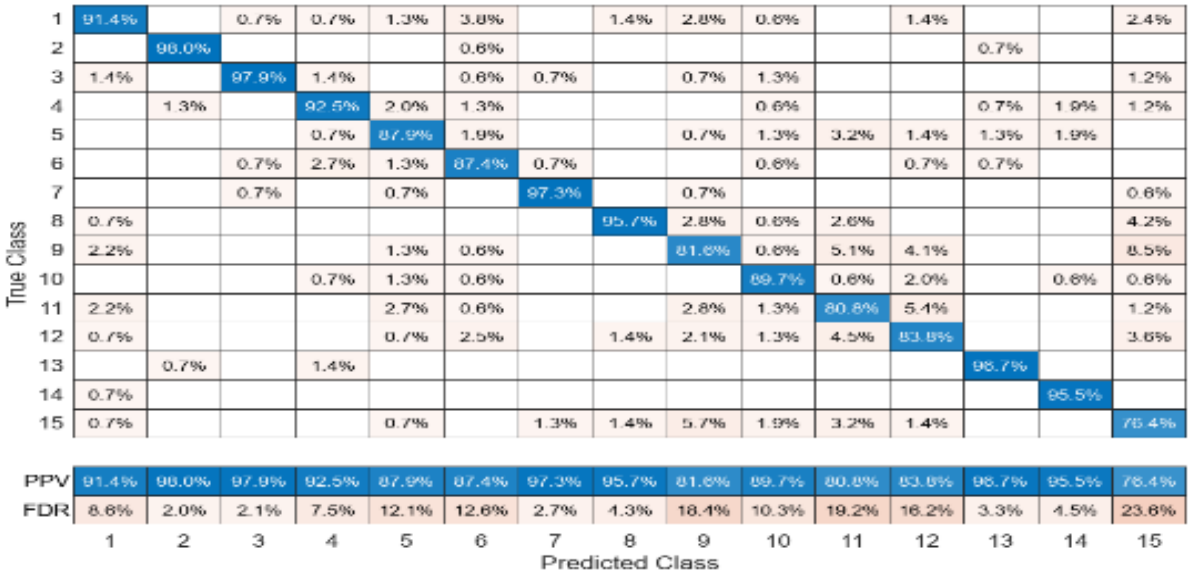

Figure 7: Confusion Matrix of PPV/FDR using VGG16 Descriptor

In Figure 7, class 2 had the highest true positive samples predicted with a 98.0% PPV while class 15 (horizontal & vertical rope activity) which involved holding on a rope and leg movement action. had the highest percentage of wrong prediction with an FDR value of 23.6% as shown in Figure 7 This shows that the pattern and shape movement while performing the action made it difficult for the classifier to identify these activities leading to the high misclassification in that class.

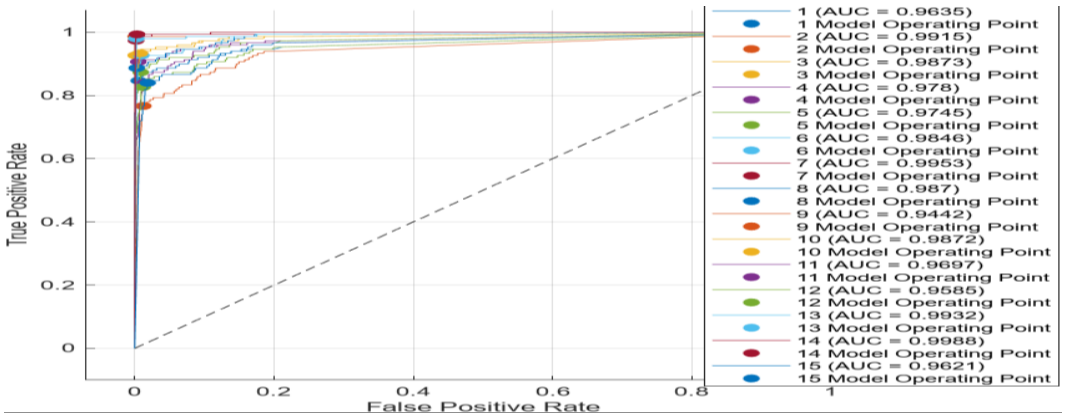

Figure 8: ROC Curve for VGG16 Descriptor

\

The ROC curve shown in Figure 8 gives further insight into the performance of the model. It is observed that most of the curves from the different classes are tilted closer to the top left corner which indicates a good-performing model that has the ability to distinguish between the different classes. Similar to other results reported in the confusion matrices, class 14 with AUC of 0.9988 shows a very high classification ability of the model to predict this class.

The obtained results show a similarity in recognition ability. The following classes of actions had good prediction using this model; class 2 (barbed wire activity) which involved crawling action and elbow movement on the ground, class 7 (9/7ft ditch activity) which involved jumping action, class 13 (high wall tyre activity) which involved climbing and jumping action and class 14 (minefield activity) which involved tiptoeing action.

The results also showed that the model had few misclassifications amongst the different classes of actions. The following classes reported various rates of incorrect predictions in a similar pattern across all the results presented; class 3 (6/ft & 10ft wall activity) which involved kicking and lying action against the wall, class 9 (Niger bridge activity) involving crossing over while holding hands in groups, class 11 (rough & tumble activity) involve jumping action, class 12 (high wall ladder activity) which involved climbing a rope and sliding down, and class 15 (horizontal & vertical rope activity) which involved holding on a rope and leg movement action. The misclassification rate is a result of shape and pattern movement in performing such actions and also similarities in patterns of actions performed across the different classes.

The study concludes that the VGG16 model performed well in activities involving multiple actions especially in crawling with elbow movement, jumping and tiptoeing. However, kicking, lying, jumping, climbing, holding and sliding were difficult to correctly identify as a result of similarity in inter-class group, especially with the jump action.

The evaluation results presented from the experiment conducted using VGG16 as the proposed approach model attained an overall accuracy of 90.1% with a training time of 104.91seconds with a good classification ability in recognizing actions in classes 13 and 14. This performance shows the high positive impact of employing deep learning model in extracting features for obstacle crossing activities recognition.

Performance Comparison Against a State-of-the-Art Model:

The performance of the proposed feature technique of VGG16 model and neural network classifier were compared with the approach used in Kolawole, Irhebhude and Odion (2025) using the same dataset.

Table 3: Performance Comparison Against State-of-the-art Model

Author	Method	Accuracy (%)	Training time (secs)	Precision	Recall
Kolawole, Irhebhude and Odion (2025)	HOGReG	86.4	31.975	80.1	86.6
Proposed model	VGG16	90	104.91	90.1	90.2

As shown in Table 3, despite having a high training time of 104.91secs, the proposed model still recorded better performance in accuracy with 3.6% increase. Similarly, the precision and recall values showed 10% and 3.6% increase respectively which shows that the proposed technique has a greater advantage in recognizing human actions in obstacle crossing.

CONCLUSION

In conclusion, this study used a VGG16 deep learned features with a neural network classifier to classify human action in a military obstacle-crossing competition. The VGG16 extracted features was computed at the max pooling layer of the pre-trained model. The neural network classifier helped in classifying the actions into 15 different classes of activities. The model achieved an improved overall accuracy of 90%, with training time of 104.91secs. This study will be relevant in real-time monitoring and analysis of activities in obstacle-crossing exercise in military environment, it will further eliminate slow process of assessing these activities.

For future work, a more robust hybrid technique is further recommended to improve the accuracy obtained and use of

dimensionality reduction algorithm to speed up the computation training time.

Acknowledgement:

The research team hereby acknowledges the Department of Computer Science, Faculty of Military Science and Interdisciplinary Studies, which were used during this study.

Ethical Statement:

The authors confirm that this study did not involve human participants, animal subjects, or sensitive personal data. Therefore, specific ethical approval was not necessary for this research.

Author Contributions:

A.O.K., contributed in the conceptualization, software, investigation, M.E.I., performed the formal analysis, data curation, review and editing of the manuscript, and visualization, and P.O.O., contributed the resources, and supervision.

Funding:

This research did not receive any specific funding from public, commercial, or non-profit organizations.

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, *8*(1), 1-74. https://doi.org/10.1186/s40537-021-00444-8

Amin, U., Muhammad, K., Hussain, T., & Baik, S. W. (2021). Conflux LSTMs Network: A Novel Approach for Multi-View Action Recognition. *Neurocomputing*, *435*, 321-329. https://doi.org/10.1016/j.neucom.2019.12.151

Ankita, Rani, S., Babbar, H., Coleman, S., Singh, A., & Aljahdali, H. M. (2021). An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones. *Sensors*, *21*(11), 1-17. https://doi.org/10.1109/JSEN.2023.3312478

Athavale, V. A., Gupta, S. C., Kumar, D., & Savita, S. (2021). Human action recognition using CNN-SVM model. *Advances in Science and Technology*, *105*, 282-290. https://doi.org/10.4028/www.scientific.net/AST.105.282

Azzag, H. E., Zeroual, I. E., & Ladjailia, A. (2022). Real-Time Human Action Recognition Using Deep Learning. *International Journal of Applied Evolutionary Computation (IJAEC)*, *13*(2), 1-10. https://doi.org/10.4018/IJAEC.315633

Banoula, M. (2023). *An Overview on Multilayer Perceptron (MLP)*. Retrieved 19 October 2023 from

Bento, C. (2021). Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis.

Bi, H., Perello-Nieto, M., Santos-Rodriguez, R., Flach, P., & Craddock, I. (2023). An active semi-supervised deep learning model for human activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, *14*(10), 13049-13065. https://doi.org/10.1007/s12652-022-03768-2

Cook, D. J., Crandall, A. S., Thomas, B. L., & Krishnan, N. C. (2012). CASAS: A smart home in a box. *Computer*, *46*(7), 62-69. https://doi.org/10.1109/MC.2012.328

Dahou, A., Al-qaness, M. A., Abd Elaziz, M., & Helmi, A. (2022). Human activity recognition in IoHT applications using arithmetic optimization algorithm and deep learning. *Measurement*, *199*, 1-11. https://doi.org/10.1016/j.measurement.2022.111445

Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, *108*, 1-23. https://doi.org/10.1016/j.patcog.2020.107561

Deep, S., & Zheng, X. (2019). Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition. 2019 29th International Telecommunication Networks and Applications Conference (ITNAC), Auckland, New Zealand (pp.1-4).IEEE. https://doi.org/10.1109/ITNAC46935.2019.9078016.

Dhiravidachelvi, E., Kumar, M. S., Anand, L. V., Pritima, D., Kadry, S., Kang, B.-G., & Nam, Y. (2023). Intelligent Deep Learning Enabled Human Activity Recognition for Improved Medical Services. *COMPUTER SYSTEMS SCIENCE AND ENGINEERING*, *44*(2), 89-91. https://doi.org/10.32604/csse.2023.024612

Ding, W., Abdel-Basset, M., & Mohamed, R. (2023). HAR-DeepConvLG: Hybrid deep learning-based model for human activity recognition in IoT applications. *Information Sciences*, *646*, 1-22. https://doi.org/10.1016/j.ins.2023.119394

Dutta, S. (2024). *Understanding Classification MLPs: An In-Depth Exploration*. Medium. Retrieved 1 FEB 2025 from.

Garcia-Gonzalez, D., Rivero, D., Fernandez-Blanco, E., & Luaces, M. R. (2023). Deep learning models for real-life human activity recognition from smartphone sensor data. *Internet of Things*, *24*, 1-22. https://doi.org/https://doi.org/10.1016/j.iot.2023.100925

Grigoryan, A. A. (2023). *Understanding VGG Neural Networks: Architecture and Implementation*. Retrieved 12 April 2024 from

Guha, R., Khan, A. H., Singh, P. K., Sarkar, R., & Bhattacharjee, D. (2021). CGA: a new feature selection model for visual human action recognition. *Neural Computing and Applications*, *33*(10), 5267-5286. https://doi.org/10.1007/s00521-020-05297-5

Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: a narrative review. *Artificial Intelligence Review*, *55*(6), 4755-4808. https://doi.org/10.1007/s10462-021-10116-x

Harahap, M., Damar, V., Yek, S., Michael, M., & Putra, M. R. (2023). Static and dynamic human activity recognition with VGG-16 pre-trained CNN model. *Journal Infotel*, *15*(2), 164-168. https://doi.org/10.20895/infotel.v15i2.916

Hayat, A., Morgado-Dias, F., Bhuyan, B. P., & Tomar, R. (2022). Human Activity Recognition for Elderly People Using Machine and Deep Learning Approaches. *Information*, *13*(6), 1-13. https://doi.org/10.3390/info13060275

Host, K., & Ivasic-Kos, M. (2022). An overview of Human Action Recognition in sports based on Computer Vision. *Heliyon*, *8*(6), 1-25. https://doi.org/10.1016/j.heliyon.2022.e09633

Huang, W., Zhang, L., Wang, S., Wu, H., & Song, A. (2022). Deep ensemble learning for human activity recognition using wearable sensors via filter activation. *ACM Transactions on Embedded Computing Systems*, *22*(1), 1-23. https://doi.org/10.1145/3551486

Irhebhude, M. E., Kolawole, A. O., & Abdullahi, F. (2021). Northern Nigeria Human Age Estimation From Facial Images Using Rotation Invariant Local Binary Pattern Features with Principal Component Analysis. *Egyptian Computer Science Journal*, *45*(1).

Irhebhude, M. E., Kolawole, A. O., & Amos, G. N. (2023). Perspective on Dark-Skinned Emotion Recognition using Deep-Learned and Handcrafted Feature Techniques. In A. H. Seyyed (Ed.), *Emotion Recognition - Recent Advances, New Perspectives and Applications* (pp. 1-23). https://doi.org/10.5772/intechopen.109739

Irhebhude, M. E., Kolawole, A. O., & Goma, H. K. (2021). A gender recognition system using facial images with high dimensional data. *Malaysian Journal of Applied Sciences*, *6*(1), 27-45. https://doi.org/10.37231/myjas.2021.6.1.275

Irhebhude, M. E., Kolawole, A. O., & Zubair, W. M. (2024). Sign Language Recognition Using Residual Network Architectures for Alphabet And Diagraph Classification. *Journal of Computing and Social Informatics*, *4*(1), 11-25. https://doi.org/10.33736/jcsi.7986.2025

Isakava, T. (2022). *A gentle introduction to human activity recognition*. Retrieved 21 February 2024 from ndatalabs.com/blog/human-activity-recognition

Jaén-Vargas, M., Leiva, K. M. R., Fernandes, F., Gonçalves, S. B., Silva, M. T., Lopes, D. S., & Olmedo, J. J. S. (2022). Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models. *PeerJ Computer Science*, *8*, 1-22. https://doi.org/10.7717/peerj-cs.1052

Jiang, Z.-P., Liu, Y.-Y., Shao, Z.-E., & Huang, K.-W. (2021). An improved VGG16 model for pneumonia image classification. *Applied Sciences*, *11*(23), 1-19. https://doi.org/10.3390/app112311185

Jnagal, A. (2018). *Image processing with deep learning- A quick start guide*.

Kavi, R., Kulathumani, V., Rohit, F., & Kecojevic, V. (2016). Multiview fusion for activity recognition using deep neural networks. *Journal of Electronic Imaging*, *25*(4), 043010-043010. https://doi.org/10.1117/1.JEI.25.4.043010

Keshinro, B., Seong, Y., & Yi, S. (2022). Deep Learning-based human activity recognition using RGB images in Human-robot collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1548-1553. https://doi.org/10.1177/1071181322661186 Khowaja, S. A., & Lee, S.-L. (2020). Semantic Image Networks for Human Action Recognition. *International Journal of Computer Vision*, *128*(2), 393-419. https://doi.org/10.1007/s11263-019-01248-3

Kolawole, A. O., Irhebhude, M. E., & Odion, P. O. (2025). Human Action Recognition in Military Obstacle Crossing Using HOG and Region-Based Descriptors. *Journal of Computing Theories and Applications*, *2*(3), 410-426. https://doi.org/10.62411/jcta.12195

Kong, Y., & Fu, Y. (2022). Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, *130*(5), 1366-1401. https://doi.org/10.1007/s11263-022-01594-9

Latumakulita, L. A., Lumintang, S. L., Salakia, D. T., Sentinuwo, S. R., Sambul, A. M., & Islam, N. (2022). Human Facial Expressions Identification using Convolutional Neural Network with VGG16 Architecture. *Knowl. Eng. Data Sci.*, *5*(1), 78-86.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Li, W., Wong, Y., Liu, A.-A., Li, Y., Su, Y.-T., & Kankanhalli, M. (2017). Multi-Camera Action Dataset for Cross-Camera Action Recognition Benchmarking. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA (pp.187-196). IEEE. https://doi.org/10.1109/WACV.2017.28.

Li, X., Zhao, P., Wu, M., Chen, Z., & Zhang, L. (2021). Deep learning for human activity recognition. *Neurocomputing*, *444*, 214-216. https://doi.org/10.1016/j.neucom.2020.11.020

Li, Y., & Wang, L. (2022). Human activity recognition based on residual network and BiLSTM. *Sensors*, *22*(2), 1-18. https://doi.org/10.3390/s22020635

Li, Y., Yang, G., Su, Z., Li, S., & Wang, Y. (2023). Human Activity Recognition Based on Multienvironment Sensor Data. *Information Fusion*, *91*, 47-63. https://doi.org/10.1016/j.inffus.2022.10.015

Liang, C., Lu, J., & Yan, W. Q. (2022). Human action recognition from digital videos based on deep learning. Proceedings of the 5th International Conference on Control and Computer Vision, Xiamen, China (pp. 150-155). Association for Computing Machinery. https://doi.org/10.1145/3561613.3561637.

Luo, F. (2020). *Human activity classification using micro-Doppler signatures and ranging techniques* [Doctoral Dissertation, Queen Mary University of London].

Manaf, A., & Singh, S. (2021). Computer Vision-Based Survey on Human Activity Recognition System, Challenges And Applications. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, Coimbatore, India (pp.110-114). IEEE. https://doi.org/10.1109/ICSPC51351.2021.9451736.

Muhammad, K., Ullah, A., Imran, A. S., Sajjad, M., Kiran, M. S., Sannino, G., & de Albuquerque, V. H. C. (2021). Human action recognition using attention based LSTM network with dilated CNN features. *Future Generation Computer Systems*, *125*, 820-830. https://doi.org/10.1016/j.future.2021.06.045

Nafea, O., Abdul, W., Muhammad, G., & Alsulaiman, M. (2021). Sensor-Based Human Activity Recognition with Spatio-Temporal Deep Learning. *Sensors*, *21*(6), 1-20.

Pardede, J., Sitohang, B., Akbar, S., & Khodra, M. L. (2021). Implementation of Transfer Learning Using VGG16 on Fruit Ripeness Detection. *I.J. Intelligent Systems and Applications*, *2*, 52-61. https://doi.org/10.5815/ijisa.2021.02.04

Patel, C. I., Labana, D., Pandya, S., Modi, K., Ghayvat, H., & Awais, M. (2020). Histogram of Oriented Gradient-Based Fusion of Features for Human Action Recognition in Action Video Sequences. *Sensors*, *20*(24), 1-32. https://doi.org/doi.org/10.3390/s20247299

Patil, S., Shelke, S., Joldapke, S., Jumle, V., & Chikhale, S. (2022). Review on human activity recognition for military restricted areas. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* *10*(12), 603-606. https://doi.org/10.22214/ijraset.2022.47926

Perera, A. G., Law, Y. W., Ogunwa, T. T., & Chahl, J. (2020). A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems*, *50*(5), 405-413. https://doi.org/10.1109/THMS.2020.2971958

Poulose, A., Kim, J. H., & Han, D. S. (2022). HIT HAR: Human Image Threshing Machine for Human Activity Recognition Using Deep Learning Models. *Computational Intelligence and Neuroscience*, *2022*, 1-21. https://doi.org/10.1155/2022/1808990

Putra, P. U., Shima, K., & Shimatani, K. (2022). A deep neural network model for multi-view human activity recognition. *PloS one*, *17*(1), 1-20. https://doi.org/10.1371/journal.pone.0262181

Qassim, H., Verma, A., & Feinzimer, D. (2018). Compressed residual-VGG16 CNN model for big data places image recognition. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, USA (pp. 169-175).IEEE. https://doi.org/10.1109/CCWC.2018.8301729.

Qi, W., Wang, N., Su, H., & Aliverti, A. (2022). DCNN based human activity recognition framework with depth vision guiding. *Neurocomputing*, *486*, 261-271. https://doi.org/10.1016/j.neucom.2021.11.044

Raj, R., & Kos, A. (2023). An improved human activity recognition technique based on convolutional neural network. *Scientific Reports*, *13*(1), 1-19. https://doi.org/10.1038/s41598-023-49739-1

Salunke, U., Shelke, S., Joldapke, S., Chikhale, S., & Jumle, V. (2023). Implementation Paper on Human Activity Recognition for Military Restricted Areas. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *11*(v), 1-7. https://doi.org/10.22214/ijraset.2023.52658

Sansano, E., Montoliu, R., & Belmonte Fernandez, O. (2020). A study of deep neural networks for human activity recognition. *Computational Intelligence*, *36*(3), 1113-1139. https://doi.org/10.1111/coin.12318

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Cambridge, UK (pp. 32-36). IEEE. https://doi.org/10.1109/ICPR.2004.1334462.

Shiraly, K. (2022). *Latest Advances in Video-Based Human Activity Recognition Modern HAR in Video & Images*. Retrieved 17 February 2024 from

Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)10.1109/ACPR.2015.7486599*, Kuala Lumpur, Malaysia (pp.730-734). IEEE. https://doi.org/10.48550/arXiv.1409.1556.

Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., & Coelho, L. d. S. (2023). Video-based human activity recognition using deep learning approaches. *Sensors*, *23*(14), 1-15. https://doi.org/10.3390/s23146384

Taud, H., & Mas, J. F. (2018). Multilayer Perceptron (MLP). In M. T. Camacho Olmedo, M. Paegelow, J.-F. Mas, & F. Escobar (Eds.), *Geomatic approaches for modeling land change scenarios* (pp. 451-455). Springer International Publishing. https://doi.org/10.1007/978-3-319-60801-3_27

Thakur, R. (2024). *Beginner's Guide to VGG16 Implementation in Keras*. Retrieved 13 APril 2024 from

Tripathi, M. (2021, 2022). *Image Processing using CNN: A beginners guide*. Retrieved 31 December 2022 from

Tufek, N., Yalcin, M., Altintas, M., Kalaoglu, F., Li, Y., & Bahadir, S. K. (2019). Human action recognition using deep learning methods on limited sensory data. *IEEE Sensors Journal*, *20*(6), 3101-3112. https://doi.org/10.1109/JSEN.2019.2956901

Vrskova, R., Hudec, R., Kamencay, P., & Sykora, P. (2022). A new approach for abnormal human activities recognition based on ConvLSTM architecture. *Sensors*, *22*(8), 2946. Vrskova, R., Kamencay, P., Hudec, R., & Sykora, P. (2023). A New Deep-Learning Method for Human Activity Recognition. *Sensors*, *23*(5), 1-17. https://doi.org/10.3390/s23052816

Wang, J., Nie, X., Xia, Y., Wu, Y., & Zhu, S.-C. (2014). Cross-view action modeling, learning and recognition. *Proceedings of the IEEE conference on computer vision and pattern recognitio*n, Columbus, United States (pp. 2649-2656). IEEE. https://doi.irg/10.1109/CVPR.2014.339.

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, *104*(2-3), 249-257. https://doi.org/10.1016/j.cviu.2006.07.013

Xia, L., Chen, C.-C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. 2012 IEEE computer society conference on computer vision and pattern recognition workshops, Providence, RI, USA.