International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No. 1 (Dec-23)

DOI:http://dx.doi.org/10.12785/ijcds/1401121

A Robust Human Activity Recognition System Using 3D CNN

Anagha Deshpande[1] **and Krishna Warhade**[2]

[1,2] *School of Electronics and Communication Engineering, Vishwanath Karad MIT World Peace University, Pune-411038, Maharashtra, India*

Received 22 Nov. 2022, Revised 31 Oct. 2023, Accepted 30 Nov. 2023, Published 28 Dec. 2023

Abstract: Video analytics has become a critical area of study in the domain of computer vision due to the availability of abundant video data. Automating human activity recognition from video footage is becoming increasingly popular due to its use in the fields of video surveillance, healthcare, and industry. Neural network models are currently being used in a varied range of scientific, academic, and commercial applications to solve image processing problems. One of the key benefits of 3D convolution neural networks (3D CNN) is their capability to learn hierarchical representations of spatiotemporal features. In this presented work, we proposed a novel 3D CNN model for detecting human activity from video sequences. A key contribution of our research is the development of a pre-processing technique including key frame selection, background segmentation, and the modeling and training of an efficient 3D Convolutional Neural Network for classifying human activities. The proposed model is tested on benchmark datasets like KTH, Weizmann, and UT-I. The performance of the model in handling challenges in datasets is also evaluated. Our proposed technique demonstrates superior recognition accuracy and training speed compared to reference methods. The proposed method has promising applications in surveillance, healthcare, sports analysis, and human-computer interaction, where accurate activity recognition is vital.

Keywords: computer vision, data preprocessing, feature extraction, neural networks, video surveillance

1. INTRODUCTION

Human beings can perform numerous activities with different intentions. Human activities can be classified into two classes regular activities and aberrant activities depending on the situational context. Automating the recognition of human activities is a crucial and prominent research area in video analytics. Human activity detection has gained popularity in many practical applications such as healthcare like elderly monitoring, rehabilitation [1], [2], entertainment, simulation [3], and surveillance systems [4]. Human eyes can detect and identify activities from the video easily, but automating this process is challenging. The variability in visual features such as appearance, color, shape, and context within video sequences poses a challenge for researchers developing human activity recognition systems. However, by addressing these challenges and leveraging the variability in visual appearance, it may be possible to improve the effectiveness of human activity recognition systems [5]. Human Activity Recognition (HAR) systems are broadly categorized into sensor and vision-based systems. Vision-based activity recognition systems do not require the use of multiple, large, and inconvenient gadgets that must be worn on different parts of the body, unlike sensor-based activity recognition systems[6]. Vision-based systems use images or videos to identify and classify actions, potentially allowing users to interact with the system in a more natural way. By utilizing visual information, vision-based systems can potentially offer advantages in terms of user acceptance and trust compared to other approaches. Human Activity Recognition depends on contextual information mainly classified into usual and unusual human activity recognition. The usual or normal human activities like standing, running, clapping, sitting, etc. [7]. Unusual or abnormal activities like falling, vandalism, fighting, loitering, etc. [8]. Identifying the aberrant human behavior with the most accuracy will help promptly to control the situations. Machine learning and deep learning methods are rapidly being employed by researchers to automate the detection of human behavior. Convolutional Neural Network (CNN) is a fundamental and flexible deep learning framework. It is a specific form of a multi-layer neural network built with the purpose of fast and effectively identifying visual patterns from pixel images. A 3D CNN is a type of artificial neural network that is explicitly designed to process 3-dimensional data, such as video sequences or volumetric images. To abstract features from spatial and temporal dimensions, we proposed the use of a 3-dimensional CNN for the recognition and classification of various human activities. This paper focuses on human activity recognition of heterogeneous activities using various benchmark data collections like KTH, Weizmann,

E-mail address:anagha.deshpande@mitwpu.edu.in,krishna.warhade@mitwpu.edu.in

and UT-I. The objective of the research presented in the paper is to design and implement an end-to-end 3D CNN model for efficient human activity recognition. The study also highlights the impact of the preprocessing of the input data on the accuracy of activity recognition and the training time required for the proposed CNN architecture. The segments of the paper are described as given below: Section 2 comprises a study of the associated literature. The elaborate discussion on the proposed system architecture in section 3 and briefing of the considered data sets. Section 4 exemplifies the testing and experimental outcomes, and Section 5 briefs the conclusion and discussion.

2. LITERATURE REVIEW

There has been a significant amount of study dedicated to the identification of human activity, and this research continues to be an active area of study. There are two main categories of strategies that have been developed for the capture and classification of human activity: (a) traditional, manually crafted approaches for feature extraction [[9]], and (b) automated feature extraction techniques using deep learning [10]. The literature review in this work concentrates on the deep learning practices in the implementation of Human Activity Recognition (HAR). The fast development in deep CNN models and improved performance enabled computer vision researchers to employ the CNN to HAR application [11], [12], [13]. These types of neural networks need minimal preparation as CNNs are designed to uncover hidden shapes in the specified data and RNNs employ time series data, which is essential to obtain temporal information. Convolutional Neural Networks (CNNs) have primarily been used for 2D images as a class of deep learning for feature building. In the approach described in the paper [14] using 2D CNNs, convolutions are applied to still images extracted from input videos. Spatial features alone may not suffice for identifying all kind of activities, and the lack of temporal consideration potentially impact accuracy, particularly for activities like sitting. A 3D Convolutional Neural Network (3DCNN) can effectively gather feature representation from images without being significantly affected by image processing. In the paper, [15] describes the 3DCNN using the concept of motion cuboids using HOG (Histogram of Gradient) and HOF (Histogram of optical flow) local features. The accuracy of models tested on real-time, high-resolution videos is hindered by the use of handcrafted feature techniques. In the approach described by the author in [16], a 3D CNN is used for feature abstraction, and long short-term memory is applied for classification. To predict future actions, the proposed approach uses parallel binary classifiers. Tran et al. [17] learned C3D features using three-dimensional convolutional kernels with a filter size of 3x3x3 and linear classifiers to classify human activities. The method represents compact features with 52.8% accuracy for UCF 101 dataset. In [18] Ashok Sarabu improved the two-stream convolutional network approach in HAR to deal with the damaged spatiotemporal features using the convolution long-short-term memory. In [19] implemented an innovative approach using two-way Long Short-Term Memory and a residual connection approach for feature extraction and effectiveness verified on UCF101 and HMDB51 datasets. As per the literature survey, it can be inferred that convolutional neural networks (CNNs) tend to perform exceptionally well on tasks related to recognizing objects in visual data. Convolutions neural networks (CNNs) are resilient to changes in factors such as viewpoint, lighting, and the presence of surrounding distractions, based on previous research. The structure of the layers in a convolutional neural network (CNN) significantly impacts the network's ability to extract relevant features from the input feed and classify or predict accurately. In Human activity recognition problems, it is essential to acquire the temporal information embedded in successive images. To effectually integrate motion evidence in video analysis, we introduced a framework that utilizes a 3D CNN to get the discriminative features from data with both spatial and temporal dimensions. However, the video inputs 3D processing needs extremely high computing costs and consumes significant run time, making it sometimes impractical for real-time applications.

3. MATERIAL AND METHODS

Recently in video analytics applications, 3DCNN-based techniques using 3D convolution layers have gained a lot of popularity [20]. Fig. 1 shows the conceptual framework for the proposed Human Activity Recognition System. The summary of the proposed activity recognition system is presented below: The videos for the various human activities from benchmark datasets are loaded as input to the system. The video datasets are divided into 3 sets: training, validation, and testing with the split of 70:20:10. The initial S seconds of each video are considered, with N frames perceived and equal interlacing between them. Then each frame is transformed to a grayscale and resized. One set of video sequences is given directly as input to the 3DCNN model. One set of video sequences is applied as input to the 3DCNN model after applying the preprocessing techniques. 3DCNN model is trained and tested separately on the data with and without preprocessing. The results in terms of test and validation accuracy, loss, and confusion matrix, ROC curves are computed following training, validation, and testing. The image frames from videos were vertically flipped to perform data augmentation to address the problem of over-fitting in neural networks.

A. *Preprocessing*

The preprocessing step involves reading a video file, applying background subtraction, resizing the frames, converting the frames from RGB to grayscale, and finally returning the processed frames as a NumPy array. This sequence of operations prepares the video frames for input to the 3DCNN model. The process of removing foreground features from backgrounds in a video frame sequence is described as background subtraction. The frame difference approach, generally known as background image subtraction, is the fundamental idea of identifying the object's movement by comparing the current frame and a reference

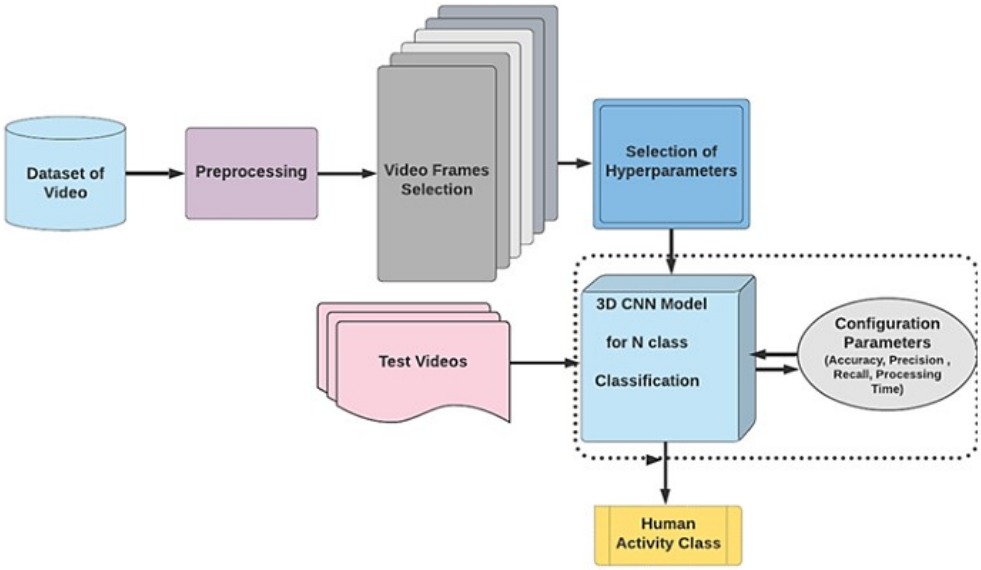

Figure 1. Conceptual Flow Diagram of Proposed Method

Figure 2. Conceptual Flow Diagram of Proposed Method

frame [21].

$$\Delta I(i, j) = ICurrent(i, j) – IPrevious(i, j) \quad (1)$$

In this work, Gaussian Mixture-based segmentation algorithm is used. The MoG (Mixture of Gaussians) technique, unlike the Kalman filter, monitors the evolution of multiple Gaussian distributions at the same time. MoG can handle multimodal background distributions and retains a density function for each pixel. The threshold parameter is selected as 16, and the learning rate is chosen automatically. Fig. 2 shows the sample video frames with background subtraction. A 3x3 elliptical structuring element for performing the opening operation, which involves the sequence of erosion followed by dilation. This operation is effective in reducing small noise and achieving smoother foreground regions, while maintaining the original shape and size of the objects within the image.

B. Proposed 3D Model

The use of 3D convolutional layers for feature extraction from video inputs has become increasingly popular in the field of video analytics in the past few years [22]. Human Activity Recognition requires the analysis of data from multiple timesteps in order to accurately identify the activity being performed. A 3DCNN model that operates on a stacked frame can learn hidden spatial and temporal patterns from the videos. The proposed Convolutional Neural Network model for this work is presented in Fig. 3. The proposed architecture comprises three convolution layers, three max-pooling layers, proceeded by one global average pooling layer, and three entirely interconnected layers. After experimenting with different filter dimensions (3x3, 5x5, and 7x7) in all layers of our 3D Convolutional Neural Networks (CNNs), we found that a 5x5 filter dimension performed better in the first convolutional layer for extracting high-level features. In the consecutive convolutional layers, a 3x3 filter dimension proved effective in capturing more detailed information while maintaining computational efficiency. The feed to the neural network is the stack of images of size 40×40×20×1 (Image height, Image width, No. frames, No. channels). The initial 3D convolution layer has a filter size of and a kernel of 5×5×5 with stride value one. Next to the initial 3D convolution layer, placed the MaxPooling3D layer of size 2×2×2 with a stride value of for temporal and spatial pooling. The max pooling layer will compress each feature map by a factor of two. The second 3D convolution layer comprises 64 filters with a kernel size of 3x3x3. Next to the second 3D convolution layer, a MaxPooling 3D with the size of 2x2x2 was added. Follows the third 3D convolution layer having 256 filters of size 3x3x3. After the third layer of 3D convolution, a MaxPooling layer of the same size is added. Next to that is the 3D Global max pooling layer. The global pooling layer down-samples the complete feature map into a unique, fixed-length vector, which can then be used as input to a fully connected layer for the classification of human activities. Dropout layers are added for preventing overfitting and improving the generalization performance. Fig. 5 indicates the model summary for the proposed model. Table I displays the hyperparameters that were fixed for the proposed model

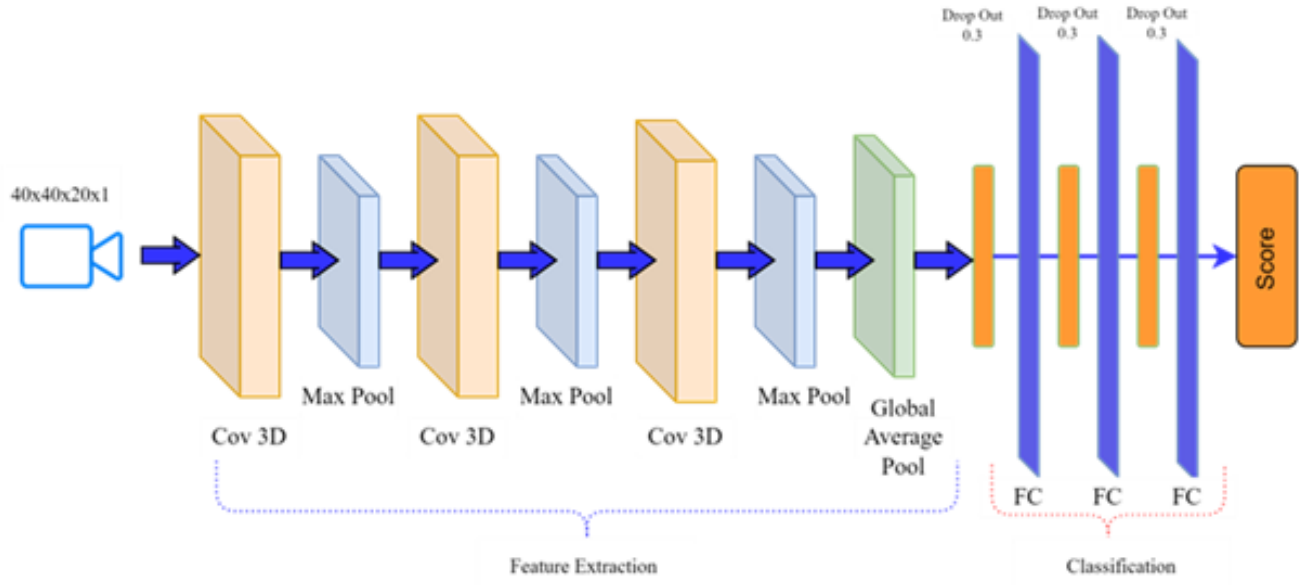

Figure 3. Proposed 3D CNN Model.

following a process of hyperparameter tuning. Dropout is a batch normalization strategy for deep neural networks that minimize overfitting and improve generalization ability equation (2) shows the loss function for categorial cross-entropy.

$$L_{CE} = -\sum_{i=1}^{N} t_i \log(P_i) \tag{2}$$

$$g_t = \nabla_{\theta_t} J(\theta_t) \tag{3}$$

The developed model was trained using Adam and Nadam [23] optimizers. The model performance in terms of accuracy was found superior for Nadam (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-07). The Nadam (Nesterov-accelerated Adaptive Moment Estimation) optimizer uses both the first moment (mean) and the second moment (uncentered variance) of the gradients to adaptively adjust the learning rate for each parameter. The first-moment estimate, like Adam, accounts for the gradient's direction, while the second-moment estimate adapts to the magnitude of the gradient. Nadam optimizer is used to enhance the training speed and to improve the model's performance.

$$m_t = \gamma m_{t-1} + n g_t \tag{4}$$

$$\theta_{t-1} = \theta_t - (\gamma m_{t-1} + n g_t) \tag{5}$$

It is computationally inexpensive and remarkably effective. The dropout value of 0.3 is decided empirically. The activation function used is ReLU to ensure the non-negative and integer value from the output from each layer. Experimentation was carried out using two different batch sizes to find the optimal batch size for the human activity task. The selection of batch sizes considered the available computational resources and dataset size. The training progress and performance were monitored, and on the basis results shown in Fig. 4, a batch size of 16 was selected as the hyperparameter value for further experimentation.

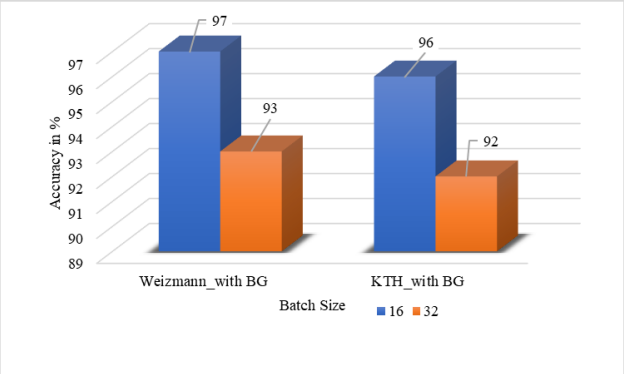

Figure 4. Batch size verses accuracy graph

C. Datasets

The KTH dataset [24] is substantially larger than the other dataset used in this work. Specifications of Dataset are as follows: Type of actions: Simple; Actions: Hand Clapping, Walking, Jogging, Boxing, Hand Waving, Running; No. Classes: 6; No. of Videos: 600; No. of actors: 25; Environment: Controlled; Frame rate: 25 fps; Resolution: 160X120; Complexity: Marginally Complex. Challenges in the datasets are Low-Resolution Scale change, Different Clothing, and Different scenarios. The Weizmann dataset [25] is relatively uncomplicated than a few other activity recognition datasets, but it can still provide valuable insights in comparing the model performance. Specifications of Dataset are as follows: Type of actions: Simple; Actions:

TABLE I. HYPERPARAMETERS USED IN THE PROPOSED 3DCNN MODEL

Model	3DCNN
Layers	3
Loss Function	Cross Entropy
Optimizer	Nadam
Learning Rate	0.001
Batch Size	16

Layer (Type)	Output Shape	Params#
Conv3D	(None, 40, 40,20,16)	2016
MaxPooling3D	(None, 20, 20,10,16)	0
Conv3D	(None, 20, 20,10,64)	128064
MaxPooling3D	(None, 10, 10, 5, 64)	0
Conv3D	(None, 10, 10, 5, 256)	442624
MaxPooling3D	(None, 5, 5, 3, 256)	0
GlobalAverage3D Pooing	(None,256)	0
Dense	(None,64)	16448
Dropout	(None,64)	0
Dense	(None,32)	2080
Dropout	(None,32)	0
Dense	(None,9)	297
Total Parameters:591,529 Trainable Parameters:591,529 Non-Trainable Parameters:0		

Figure 5. Summary of the proposed 3DCNN model

Jump, run, side, hand waving, walk, jumping jack, jump in, bend; No. Classes: 10; No. of Videos: 90; No. of actors: 9; Environment: Controlled; Frame rate: 25 fps; Resolution: 180*144; Complexity: Simple. Challenges in the datasets are low resolution, camera view change, and occlusions. The UT-I interaction [26] dataset was recorded in a parking lot environment with a relatively stable background and minimal camera movement. There were typically only two individuals present in each frame of the dataset, and the background remained largely unchanged throughout the recordings. Specifications of Dataset are as follows: Type of actions: Interaction; Actions: Handshaking, Kicking, Hugging, Pushing, Pointing, Punching; No. Classes: 6; No. of Videos: 120; No. of actors: 25; Environment: Realistic, Windy Background; Frame rate: 30 fps; Resolution: 720*480; Complexity: Complex. Challenges in the datasets are moving background, camera jitters, pedestrians in the scene, different clothing condition The UCF 101 data [27] collection is the most challenging and popular dataset used in the research community. The UCF 101 dataset consists of 133220 clips that fit into 101 classes. The time span of each video is from 3 to 10 seconds. Challenges in the datasets are variation in camera motion, scale, view angle, jumbled backdrop, lighting environments, object appearance and position

Figure 6. Challenge in the Datasets

4. RESULTS AND DISCUSSION

This section describes the experimental outcomes. The efficacy of the hypothesized 3DCNN architecture is validated using benchmark datasets like KTH, Weizmann, and UT-Interaction. The datasets were partitioned as training, testing, and validation with a ratio of 70:20:10. Table II shows the data split details for each dataset. In the training phase for the KTH dataset, 35 frames were selected from each video sample with a temporal gap of seven seconds between them. The selected frames were then pre-processed and used as input to train the model. For the Weizmann dataset, 20 frames were selected with a temporal gap of 2 seconds between them, while 21 frames were selected for the KTH dataset with a temporal gap of 3 seconds between them. This frame selection strategy was implemented to lessening the computational cost and improve the training speed of the proposed model. The model was trained using the hyperparameter values stated in Table I. In this study, a regularization technique called dropout was applied to the model with a value of 0.3. Batch size is determined through experimentation to manage a balance between efficiency and generalization. Early stopping is utilized to automatically

TABLE II. DATA SETS SPLIT DETAILS

Data Set	Train Samples	Validation Samples	Test Samples	Seconds S	Frames F
KTH	300	122	100	7	35
Wiezmann	214	68	30	2	20
UT-I	65	22	20	3	21

find the optimal number of epochs based on validation performance. The batch size for training was set to 16 instances, and the count of epochs was set to 60/80. The model was evaluated by calculating both the validation loss and the training loss, both with and without a preprocessing step involving background subtraction (BG) to evaluate the performance. Confusion matrices are useful for assessing the performance of a classification model in terms of correctly classified and misclassified classes. The confusion matrices for the KTH and Weizmann action datasets are illustrated in Fig. 7 and 8, respectively. These matrices provide insights into the performance of the proposed method. Additionally, Fig. 9 and 10 present the training and validation loss plots for the KTH and Weizmann action database, respectively, as generated by the proposed models. These plots help visualize the convergence and generalization capabilities of the models during training. Accuracy is a parameter that measures the model's ability to correctly classify or predict the target class or condition. It is defined as the proportion of true predictions by the model, relative to the total number of predictions. The evaluation metrics recognition accuracy with and without preprocessing step evaluated for the proposed 3DCNN model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. The Receiver Operating Characteristic (ROC) curve is a visual representation of the trade-off between the true positive rate and the false positive rate. The AUC (Area Under the ROC Curve) is a metric that quantifies the overall performance of the classifier. Fig. 11 and 12 represents the ROC curve for KTH and Weizmann dataset. The AUC-ROC values closer to 1.0 indicate a stronger discriminatory power of the model. Fig. 14 shows the overall assessment of recognition accuracy by the model with and without pre-processing. The comparison depicts that the accuracy improved after the implementation of the background subtraction step.

A. Robustness Test

The proposed model also tested for the subset from the most challenging UCF101 dataset. The model tested on a random subset of eight out of 101 activities from the data set. The model tested for randomly selected eight activities out of 101 from the data set. The data set comprises 1103 total video clips. The data was split in an 80:20:10 ratio with 772 videos in training data, 111 videos in validation data, and 220 videos in test data. The recognition accuracy

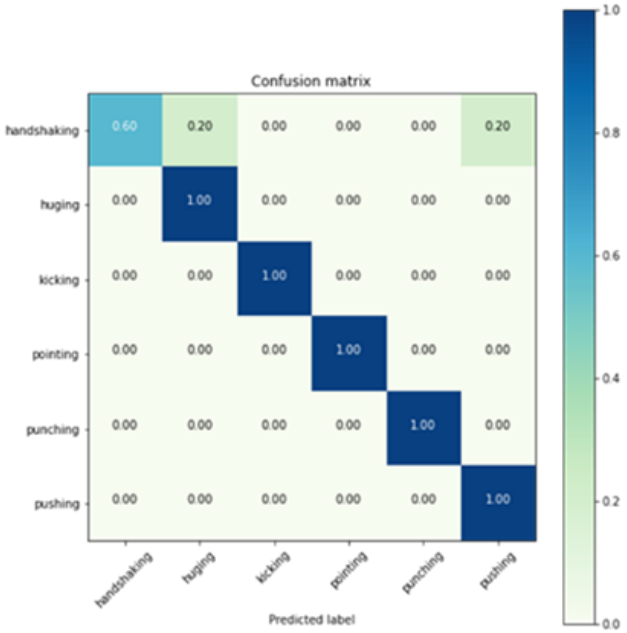

Figure 7. UT-I Dataset Confusion Matrix

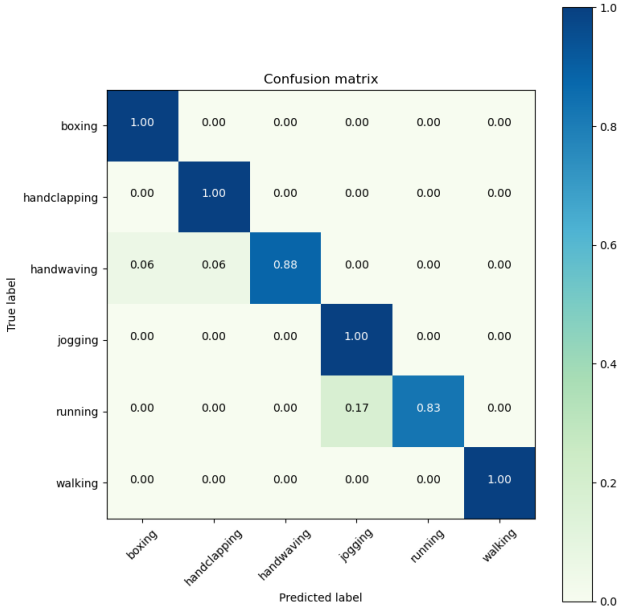

Figure 8. KTH Dataset Confusion Matrixl

TABLE III. TESTING ON THE WEIZMANN ROBUSTNESS DATASET

Scenario No.	Details of the specimens used	Recognition Accuracy
1	112 specimens of 8 classes (other than usual walk) + 10 specimens of unusual walk +10 specimens of walk with view angle change	93.33%
2	175 specimens of 9 classes including usual unusual view angle change walk	93.33%

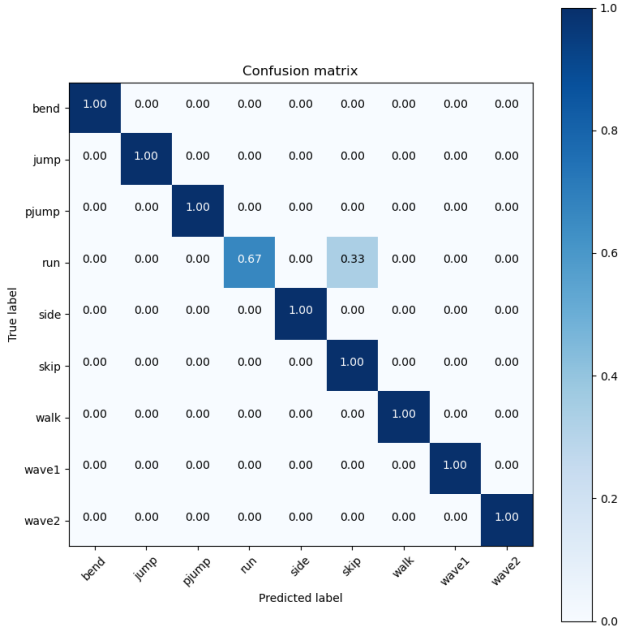

Figure 9. Weizmann Dataset Confusion Matrix

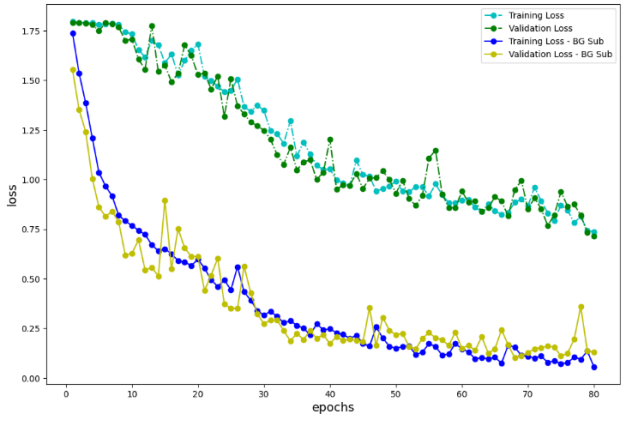

Figure 10. Validation loss and Training loss for KTH

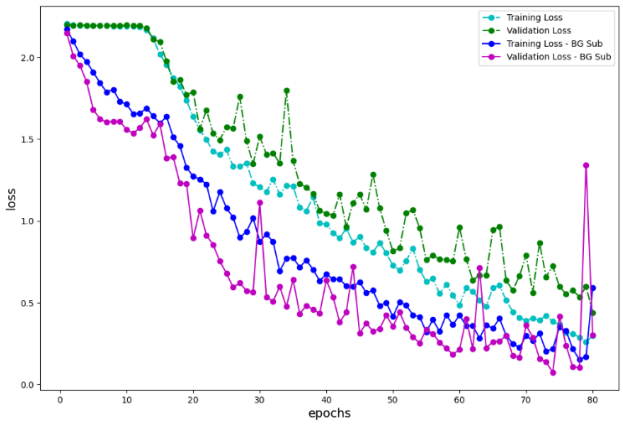

Figure 11. Validation loss and Training loss for Weizmann Dataset

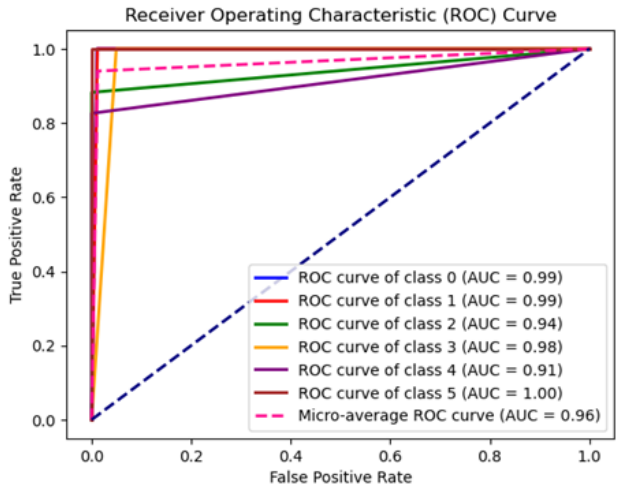

Figure 12. ROC Curve for KTH Datasetl

with the preprocessing step estimated was 85%. Fig. 15 shows the confusion metrics for the subset of the UCF data set. Weizmann's robustness data set was utilized to evaluate the hypothesized 3D CNN Model's endurance to challenging situations such as occlusion, irregular action execution, varied context, and different view angles. The developed architecture was trained and tested on the chal-

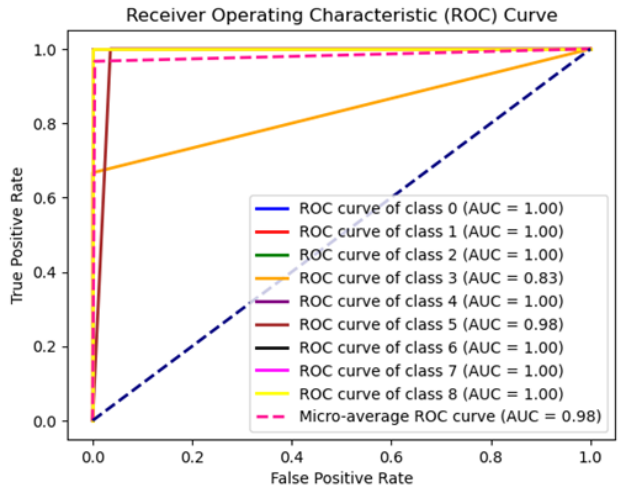

Figure 13. ROC Curve for Weizmann Datasetl

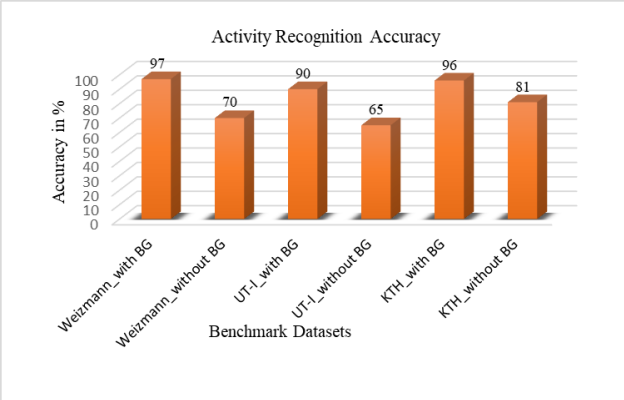

Figure 14. Accuracy graph with and without pre-processing

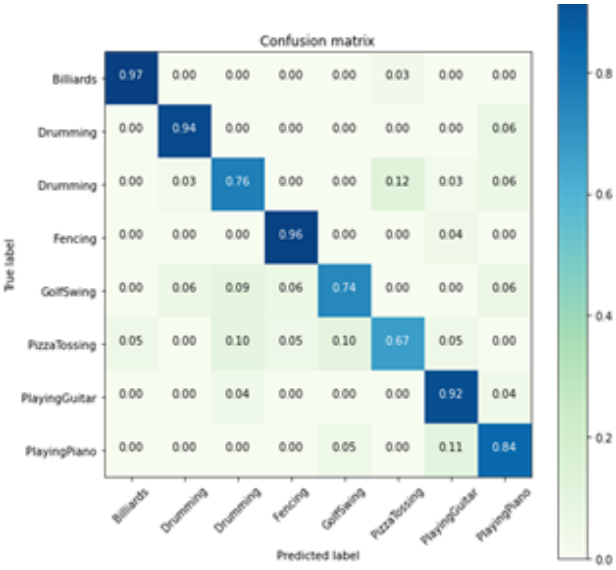

Figure 15. Modified UCF Dataset Confusion Matrix

Figure 16. Challenges in the Weizmann Datasets

lenge. Weizmann data-set. Fig. 16 shows sample images for real-time challenges like occlusion, unusual scenarios, a person with objects, and viewpoint variation in the data set for walking activities. The Weizmann camera view rotation change data set included recordings of walking behavior at 10 diverse camera view angles, oscillating from 0°to 90°. The experimentation for robustness testing was performed using two different scenarios as mentioned in Table IV and an average recognition accuracy of 93% was attained. Fig. 17 displays the confusion matrix for challenging Weizmann data set 2. The misclassification score is 0.0667, which shows that the model performs well in challenging scenarios. Fig.18 shows the graph for the model training time as one of the evaluation metrics examined. The model takes less training time without the background subtraction preprocessing step compared to background subtraction.

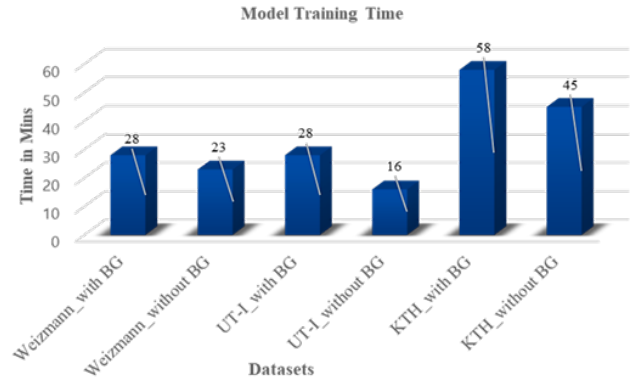

Figure 17. Proposed Model Evaluation Time

The evaluation parameters of the competing approach are compared against the previously published studies in human activity recognition with different approaches like History Trace Template for feature abstraction and support vector machine for classification [28]. The study in [29], which employs a Gaussian mixture random matrix for dimensionality reduction (CS-GMRM), primarily concentrates on reducing the dimensionality of the data. It

TABLE IV. COMPARISON TABLE FOR RECOGNITION ACCURACY IN % WITH EXISTING APPROACHES

References	Year	Method	KTH	Weizmann	UT-I
28	2013	HTT+SVM	90.22	93.41	-
29	2015	CS-GMRM	92.2	94.4	-
30	2017	SRC	96.66	97.78	-
31	2018	NMC	90.58	95.56	-
32	2018	3DCNN+3DMotion Cuboid	94.9	97.2	-
33	2019	SCM	-	-	87.50
34	2018	DT+MIL	-	-	85.80
35	2020	DBN	94.83	-	-
36	2021	Distance Transform + Entropy Features + ANN	91.4	-	-
37	2021	Transfer Learning	96.5	-	-
38	2022	ResInc-ConvLSTM	94.08	79	-
Proposed Model	2023	BG+3DCNN	96	97	90

recognizes that some actions in the KTH data set may exhibit mutual confusion due to substantial overlap in their visual and motion characteristics. The silhouettes method in [30] is susceptible to dissimilarities in lighting situations, camera viewpoints, and occlusions. However, the use of a sparse representation-based classifier is a novel aspect of the work. In the method described in [31], a separate view-invariant feature extraction and a nearest neighbor classifier are necessary, whereas the proposed approach integrates these two components into a single, automatically functioning model. The proposed model incorporates a higher temporal dimension with a larger and adaptive frame count as compared to the motion cuboid approach where 11 successive frames are utilized as input [32]. The inclusion of a greater temporal dimension with an increased number of frames enhances the representation of intricate motions and the handling of occlusions, leading to a more robust model. The model results were also compared with various methods like the novel sub-volume co-occurrence matrix descriptor [33], the approach based on sparse presentation of feature covariance matrices [34], deep belief networks [35], distance transform and entropy features of human silhouettes and ANN for classification [36]. Time-sliced averaged gradient boundary magnitude employed for feature extraction and classifier employed using transfer learning approach [37] and the residual inception convolutional recurrent layer integrated into ConvLSTM [38]. Table IV presents comparison insights for the relative performance of the model compared to the existing approaches. The comparison confirms that the proposed model is either superior or comparable to current methods in many cases.

5. CONCLUSION

In this study, we developed a framework using a 3D neural network that can classify human activities from video feeds. We used the KTH, Weizmann simple human activity data set, and UT-I human interaction databases for training and testing. The videos in these data sets were first converted to images of size 40X40. The available size of GPU and memory constraints forced to resize the images to a smaller size, which may impact the overall accuracy of the classification results. The design of the 3DCNN architecture significantly affects its performance. In the presented study, we designed and evaluated the proposed 3DCNN architecture and compared its accuracy to state-of-the-art methods. The results showed that this novel 3DCNN architecture had the highest accuracy and effectively addressed overfitting problems. The analysis of the confusion matrices as depicted in Fig 7, 8, and 9, revealed that the proposed architecture performed well in classifying human activities into groups, with minimal misclassifications. Additionally, the proposed model demonstrated robustness in classifying activities in challenging scenarios such as occlusion, variations in view angle, illumination, and scale. In our future work, we plan to address training and evaluation

challenges by working with a large data set. Additionally, we will consider extending the approach by incorporating an attention mechanism to emphasize the most relevant features in activity classification.

References

[1] F. L. A. S. I. Bisio, A. Delfino, "Enabling iot for in-home rehabilitation: Accelerometer signals classification methods for activity and movement recognition," *IEEE Internet Things Journal*, vol. 4, pp. 135–1146, 2017.

[2] A. S. M. Z. Uddin, "Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning," *Scientific Report*, vol. 11, 2021.

[3] F. P. S. Herath, M. Harandi, "Going deeper into action recognition: A survey", image and vision computing," *Image and Vision Computing*, vol. 60, pp. 4–21, 2017.

[4] R. Xu, Y. Guan, and Y. Huang, "Multiple human detection and tracking based on head detection for real-time video surveillance," *Multimedia Tools Appl.*, vol. 74, no. 3, p. 729–742, feb 2015. [Online]. Available: https://doi.org/10.1007/s11042-014-2177-x

[5] R. Kolkar and V. Geetha, "Issues and challenges in various sensor-based modalities in human activity recognition system," in *Applications of Advanced Computing in Systems*, R. Kumar, R. K. Dohare, H. Dubey, and V. P. Singh, Eds. Singapore: Springer Singapore, 2021, pp. 171–179.

[6] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Comput. Surv.*, vol. 54, no. 4, may 2021. [Online]. Available: https://doi.org/10.1145/3447744

[7] M. Ibrahim, J. Kainat, S. S. Ullah, and S. Al-Hadhrami, "An effective approach for human activity classification using feature fusion and machine learning methods," *Applied Bionics and Biomechanics*, vol. 2022, pp. 1–14, 02 2022.

[8] D. J. Samuel, R. E. Fenil, G. Manogaran, G. N. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. Ahilan, "Real-time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," *Computer Network.*, vol. 151, 2019.

[9] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, pp. 976–990, 2010.

[10] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International Workshop on Human Behavior Unterstanding*, 2011.

[11] R. Alfaifi and A. Artoli, "Human action prediction with 3d-cnn," *SN Computer Science*, vol. 1, 08 2020.

[12] S.-U. Park, J. Park, M. Al-masni, M. A. Al-antari Aisslab, M. Z. Uddin, and T.-S. Kim, "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services," vol. 100, 10 2016.

[13] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human activity recognition using binary motion image and deep learning," *Procedia Computer Science*, vol. 58, pp. 178–185, 2015, second International Symposium on Computer Vision and the Internet (VisionNet'15). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915021614

[14] K. K. Verma, B. M. Singh, H. L. Mandoria, and P. Chauhan, "Two-stage human activity recognition using 2d-convnet," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, pp. 125–136, June 2020.

[15] P. S. Sahoo and S. Ari, "On an algorithm for human action recognition," *Expert Systems with Applications*, vol. 115, pp. 524–534, 2019.

[16] R. Mutegeki and D. S. Han, "A cnn-lstm approach to human activity recognition," in *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 362–366.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[18] A. Sarabu and A. Santra, "Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks," *Emerging Science Journal*, vol. 5, pp. 25–33, 2021.

[19] W. Xiao and J. Qingge, "Tbrnet: Two-stream bilstm residual network for video action recognition," *Algorithms*, vol. 13, no. 7, p. 169, 2020.

[20] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, "Human activity classification using the 3dcnn architecture," *Applied Sciences*, vol. 12, no. 2, p. 931, 2022.

[21] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *2006 International Conference Pattern Recognition*, Hong Kong, China, 2006, pp. 441–444.

[22] S. Yu, Y. Cheng, L. Xie, and S. Z. Li, "Fully convolutional networks for action recognition," *IET Computer Vision*, vol. 11, no. 8, pp. 744–749, 2017.

[23] T. Dozat, "Incorporating nesterov momentum into adam," in *Proceedings of the 4th International Conference on Learning Representations*, 2016, pp. 1–4.

[24] I. Laptev, "Local spatio-temporal image features for motion interpretation," Ph.D. dissertation, Department of Numerical Analysis and Computer Science, Royal Institute of Technology (KTH), Sweden, 2004.

[25] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1395–1402.

[26] M. S. Ryoo and J. K. Aggarwal, "Ut-interaction dataset, icpr contest on semantic description of human activities (sdha)," cvrc.ece.utexas.edu/SDHA2010/HumanInteraction.html, 2010, dOI: Not applicable.

[27] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human action classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[28] G. Goudelis, K. Karpouzis, and S. Kollias, "Exploring trace transform for robust human action recognition," *Pattern Recognition*, vol. 46, no. 12, pp. 3238–3248, 2013.

[29] J. Jiang, X. He, M. Gao, X. Wang, and X. Wu, "Human action recognition via compressive-sensing-based dimensionality reduction," *Optik*, vol. 126, no. 9-10, pp. 882–887, 2015.

[30] H. Qian, J. Zhou, and Y. Mao, "Recognizing human actions from silhouettes described with weighted distance metric and kinematics," *Multimedia Tools and Applications*, vol. 76, pp. 21 889–21 910, 2017.

[31] K. P. Chou, M. Prasad, and D. Wu, "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, p. 1, 2018.

[32] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, "Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos," *Procedia Computer Science*, vol. 133, pp. 471–477, 2018.

[33] D. G. Lee and S. W. Lee, "Prediction of partially observed human activity based on pre-trained deep representation," *Pattern Recognition*, vol. 85, pp. 198–206, 2019.

[34] J. Wang, S. C. Zhou, and L. M. Xia, "Human interaction recognition based on sparse representation of feature covariance matrices," *Journal of Central South University*, vol. 25, no. 2, pp. 304–314, 2018.

[35] M. Abdellaoui and A. Douik, "Human action recognition in video sequences using deep belief networks," *Traitement du Signal*, vol. 37, pp. 37–44, 2020.

[36] P. Ramya and R. Rajeswari, "Human action recognition using distance transform and entropy-based features," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8147–8173, 2021.

[37] S. Zebhi, S. M. T. AlModarresi, and V. Abootalebi, "Transfer learning based method for human activity recognition," in *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, Tehran, Iran, Islamic Republic of, 2021, pp. 761–765.

[38] S. Khater, M. Hadhoud, and M. B. Fayek, "A novel human activity recognition architecture: using residual inception convlstm layer," *Journal Engineering Applied Science*, vol. 69, no. 45, 2022.

Anagha Deshpande Anagha Deshpande is a holder of a master's degree in Digital Systems from Pune University. Presently, she serves as an assistant professor at MIT World Peace University in Electronics and Communication Engineering, bringing with her 17 years of teaching experience. she serves as a research scholar in the School of Electronics and Communication Engineering, focusing her research primarily on signal and video processing.

Krishna Warhade Dr. Krishna K. Warhade holds a Ph.D. from IIT Bombay, where he specializes in communication and signal processing. He has over three decades of teaching experience. His research pursuits encompass diverse areas such as video processing, wavelets, biomedical signal processing, and agriculture. Notably, he has shared his insights and findings through publications in esteemed journals and conferences. Dr. Krishna K. Warhade is a Professor in the Electronics and Communication Engineering Department, Presently serving as the Director of the Doctoral Program at MIT World Peace University.