Human Activity Recognition using LSTM & CNN

Veeri Dheeraj
School of Computer Science and Engineering
Lovely Professional University
Phagwara, India
veeridheeraj15@gmail.com

Imran Khureshi
School of Computer Science and Engineering
Lovely Professional University
Phagwara, India
iamimran117@gmail.com

Vamshidhar Reddy
School of Computer Science and Engineering
Lovely Professional University
Phagwara, India
vamshireddybanthi@gmail.com

Krishna Anvith V
School of Computer Science and Engineering
Lovely Professional University
Phagwara, India
vattikutianvith@gmail.com

Ajay Sharma
upGrad Campus
upGrad Education Private Limited
Banglore,Karnataka,India
ajaysharmadharmani@gmail.com
ORCID: 0000-0001-6620-4805

Shamneesh Sharma
upGrad Campus
upGrad Education Private Limited
Banglore,Karnataka,India
shamneesh.sharma@gmail.com
ORCID: 0000 -0003-3102-0808

***Abstract*— The issue of human activity detection stands as the new crucial research area, embracing healthcare, surveillance, robotics, and human-computer interaction among its wide variety of applications. The current article highlights the current strategies, difficulties, and perspectives in the research domain of human activity identification. This research work takes up sensor technologies evolution and the impact on data collection, especially wearable sensors, sensors in the ambient, and ubiquitous computing devices. The current work represents various human feature extraction methods that are time-domain features, frequency-domain features, and spatial-temporal features. [] The future perspective of the following abstract also considers the use of machine learning algorithms in activity recognition, which encompasses classical classifiers such as Support Vector Machines (SVM) and Random Forests.**

Keywords*— Human Activity Detection, Long- Short term memory, Computer Vision, Machine Learning*

I. INTRODUCTION

A. Background

Human Activity is a wide range of actions done by humans like eating, sleeping, walking to more complex behaviours. The present work is one of the important tasks which involves human activity detection, in the field of computer vision and artificial intelligence, and classify human actions automatically from video and sensor data. The pixel level identification in image processing is the field which generates interest due to its wide applications in other domains like security, healthcare, sports analysis, human- computer interaction, and robotics. The capability to recognize. and award human activities without a human supervisor is a fact in which such activity can be proved to be invaluable. Surveillance, lifts automation of monitoring the public spaces, improving means in the space of security and safety. In the healthcare industry, it provides a platform where patient's activities like movement and rehabilitation are traced. It is a valuable tool for the analysis in sports that produces invaluable strategies for athletes and training.

B. Motivation

Recognizing human activities through technological means having a very critical area of research, with thoughtful implications across various applications such as healthcare, surveillance, and smart environments. The feature of the system to detect and segment human actions with the help of sensors' data enables health improvement for people and taking the productivity of many applications to a new level. The present research work follows the purpose and full fill the scope to go on with improvement of human activity detection using deep learning algorithms.

Fig. 1. Image showing different activitie in video frames from the dataset

C. Research Contribution

In this paper the author has work and presents some insightful contributions in human activity recognition using deep learning methods and especially Long- Short- Term Memory (LSTM) networks. The primary activity of our group concerns the researching and establishing together a new neural network model designed personally for activity recognition. Such a special-temporal contextual enhancement, which is desirable for accurate identification of diverse activities, is essentially brought about by this architecture that effectively incorporates temporal dependencies that exist in human activities. This provides an extended benchmark dataset including the real-life activity data captured from the different environments. This is used to work out a standardized evaluation process and allows comparison of different approaches. The incorporation of LSTM networks within our architecture does not only allow the extraction of long-term temporal dependencies but also makes the modelling of chaos simpler, therefore, accurate classification is done.

D. Preliminaries

Datasets of videos act as info depositories for human activity detection due to their variety of rich information, which reveals dynamics in time and space, interactions of objects and contextual stimulators. They empower people do with a more complete real behaviour model that would not be represented sufficiently by a normal photo or sensor data

because the remaining moves are complex. This is done by the broad spectrum of activities in different settings - that is, video sets are a good basis for the applicability of developed models. The temporal framework provided by the video sequel is the basis for time course analysis that looks at the activity progression from one point in time to another, revealing the flowing nature of human behaviours in time. These datasets work like standards often accepted for evaluating detection algorithms and create equal, consistent conditions for comparing different approaches, which triggers constant development in the field. In addition to video-based activity detection that has got practical application for surveillance, healthcare, and smart environment which makes it possible instantaneous monitoring and reaction for abnormal activities. Study and pattern mining become faster and easier when human activities are recorded on videos. It simplifies and quickens the discovery of behaviour, social interactions, and routines. As a result, it makes a whole field of studies like psychology and sociology.

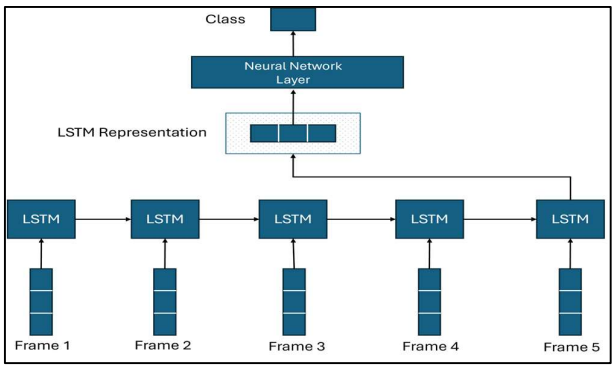


Fig. 2. The figure showning the Frames into LSTM and CNN

E. Proposed Framework

A learned LSTM (Long Short-Term Memory) paradigm was used for this project to develop the model. The core feature of LSTMs here is the ability to encode these long-standing dependencies in the sequences. To begin with, traditional RNNs are faced with the issue of the vanishing gradient problem, where all the gradients are increased exponentially as they flow backwards through time, rendering it almost impossible to learn long-term dependencies. Therefore, LSTMs solve this problem altogether by introducing a mechanism of gating which allows it to remember or declutter data as a function of time. LSTMs are specifically designed to deal with the mentioned problem through introducing the memory cell that functions as a 'warehouse' where information can be saved for relatively longer. LSTMs can be as well employed as CNNs network architectures for instance – video and images analysis. The memory cell is controlled by three gates: the input gate, the output gate, and the forget gate. The gates here ponder the flow of information because they decide if to input, delete or output the data from the cells' memory. A control mechanism called the input gate determines the amount of information fed into memory cell. The forget gate uses gate-controlled input so only specific if each information is deleted from the memory cell, on the other side memory cells output the information through what is controlled by the output gate. Backpropagation enables LSTM networks to select relevant information and ignore the ones which pass through the network by implicitly learning. This technique helps the model to learn long term dependencies. Memory manipulations are done by three gates –

1) Forgot Gate: Through the mechanism of a memory gate (Forget gate), the information that is no longer necessary is eliminated. Two aspects gate xt (input at the exact time period) and ht-1 (previous cell output) are multiplied by the weight matrices then bias is added, carries out the function of taking the input as input and delivers a binary output. For the particular states i.e. cell state, the output of 0 is information disappears and for output 1 is a next are recall for further use. The equation for the forget gate is:

$$\Phi\tau = (\omega\phi [\eta\tau-1, \xi\tau] + \beta\phi) \quad (1)$$

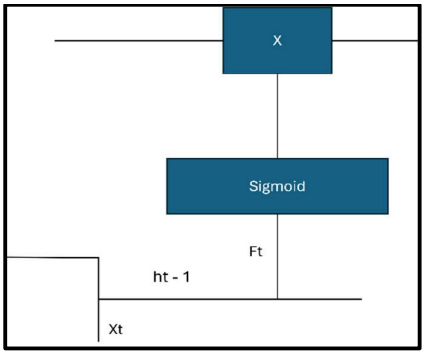


Fig. 3. Forgot Gate for the elimination of of unnecessary information in the CNN and LSTM network

2) Input Gate: The formation of cell memory occurs through the information provided by the input gate to the cell state. It powerfully governs the information and the values of the information related with ht-1 and xt using sigmoid function and the controlling similar to the forget gate filter. Subsequently, a vector is chopped out of the tanh function that has every output from ht-1 to x.At last, the The multiplication of vectors of the magnitude space and the regulated quantities gives the required details. The equation for the input gate is: The equation for the input gate is:

$$\iota\tau = (\Omega\iota \ [\eta\tau-1, \xi\tau] + \beta\iota) \quad (2)$$

$$X\tau = \tau\alpha\nu\eta(\Omega\chi [\eta\tau-1, \xi\tau] + \beta\chi) \quad (3)$$

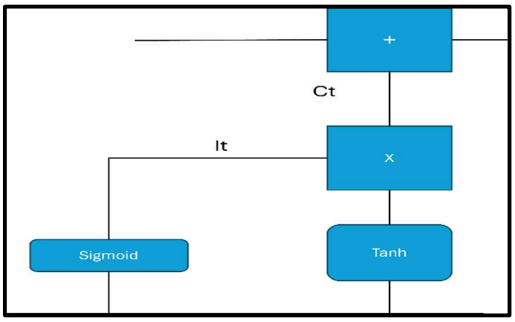


Fig. 4. Input Gate video frames processing in for LSTM network

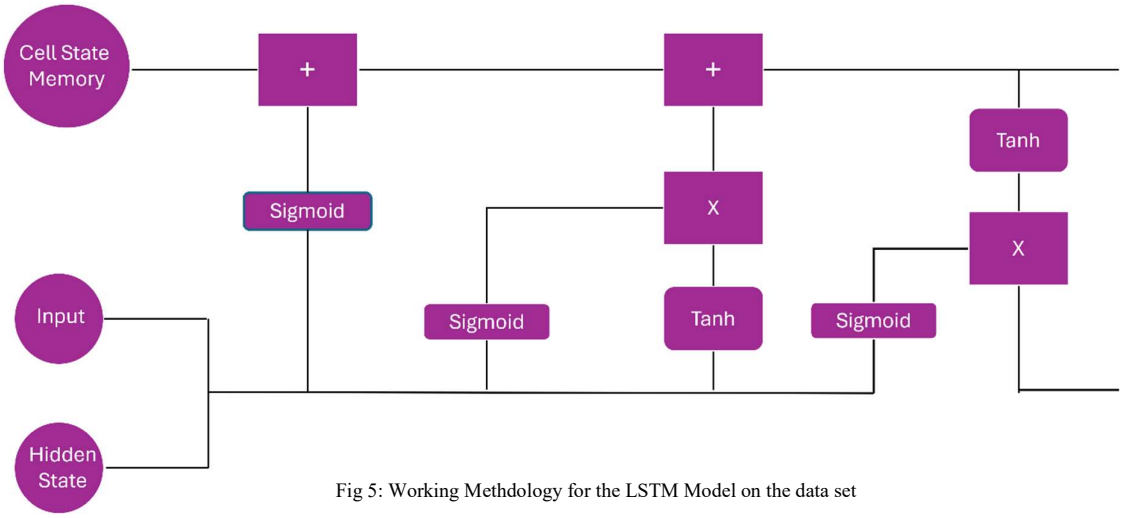


Fig 5: Working Methdology for the LSTM Model on the data set

The We multiply time to reflect, it is the our time and by it, we can realise the information that we have been ignoring. Next, we include ιτ ∗ Xτ. This represents a new candidate value, recentered for the sum we have used to update each state value.

$X\tau = \phi\tau\chi\tau{-}1 + \iota\tau\chi\tau$ where

⊙ denotes elementwise multiplication tanh is tanh activation function.

It consists of two sub-components: the input gate itself, which decides which values to update, and the tanh layer, which generates new candidate values to add to the cell state.

3) Output Gate: In the process of that, the critical data related to the current cell state is extracted, which will be presented in the output. Initially, vector is created when a cell is multiplied by the cell itself. The data is processed through the sigmoid function and then it is made sure that only required value is retained based on inputs ht-1 and xt. In the end, inside the cell multiplication of the vector and the regulated values is done which represents the output and trickled down to the next one. The equation for the output gate is: The equation for the output gate is:

$$O\tau = (\Omega o [\eta\tau{-}1, \xi\tau] + \beta o) \quad (4)$$

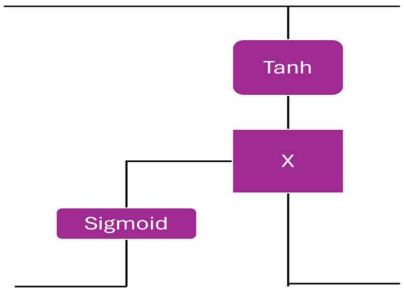


Fig. 6. Output gate used to retain the values from LSTM and CNN based network

II. METHODOLOGY

Our methodology encompasses several stages, including data preprocessing, model design, training, and evaluation. We describe the preprocessing steps involved in preparing the UCF50 dataset for training, including video normalization, frame extraction, and annotation. Subsequently, we introduce the architecture of our deep learning models, which combine CNNs and LSTM to capture spatial and temporal features from video sequences effectively. Details regarding model hyperparameters, optimization techniques, and training strategies are provided.

A. Data Collection

The UCF dataset is a remarkable base of research in connection with the examination of human activity based on video data. This part gives the description of the data collection approach including the sources of video materials, their annotation process, and metadata creation. The content of the video in the UCF50 dataset was collected from many sources, including different video repositories and online platforms, plus some of the videos were recorded and supplied as datasets. The general principle was the selection of the representatives that would include a variety of activities humans to do under specific conditions and circumstances. Videos were selected to fit into 50 selected activities, which were initially given as the dataset. The duration of each activity takes approximately 140 videos.

B. Data Preprocessing

Our dataset for training purpose deals with classifying dry, wet, mushy, and burned. Horserace, Biking, Swing, Pizza tossing are points activities that are very attractive and involve people actively. Resized frame dimensions: 64 x64 pixels Sequence length. In 20 shots per sequence. Dataset directory: "UCF50" Classes (activities) in the dataset: "Horserace" (horse-riding), "Biking" (bicycling), "Swing" (swimming), and "Pizza Tossing" (pizza tossing). Once all frames contributed during the holding period and processed, the video reader is then liberated for other uses. Lastly, the function returns the list of detected frames (video in a discontinuous state).

Steps involved in this :

- Let the frame tackle the issue.
- Presently, get past the current line.
- Use frame parameters given to indicate desired size.
- Apply process of standardization onto the resized image

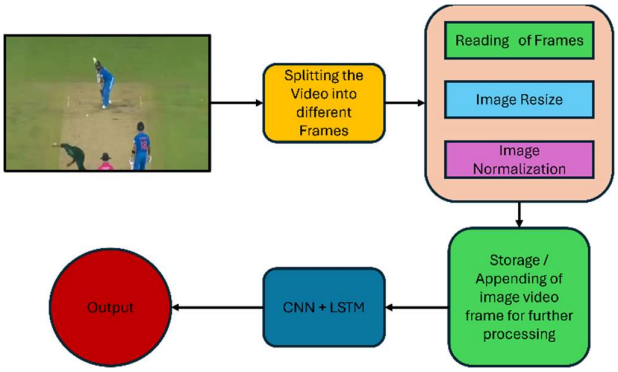
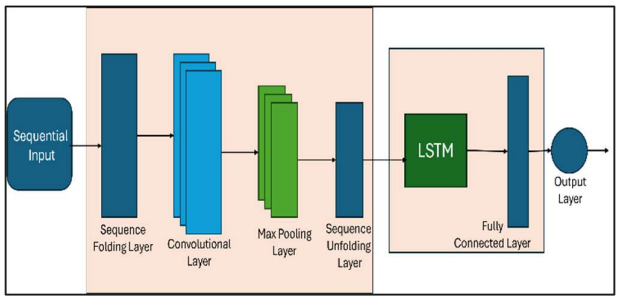


Fig. 7. Preprocessing of Splitting Frames from the dataset

C. Convolutional LSTM

In the proposed ConvLSTM-based model architecture, we have designed a deep neural network comprising multiple layers to effectively capture spatiotemporal patterns inherent in video data for the task of human activity recognition. The model architecture is structured sequentially, with each layer serving a specific function in feature extraction and classification. Beginning with the ConvLSTM2D layers, which utilize convolutional LSTM cells, the model extracts spatial and temporal features simultaneously from input video sequences. The choice of parameters, such as filter sizes and activation functions, is meticulously made to enhance the model's ability to discern meaningful patterns across frames. Additionally, recurrent dropout is incorporated within the ConvLSTM layers to regularize the network and prevent overfitting.

Fig. 8. CNN + LSTM working methodology model used for the traning and testing on the video frame image dataset



Following each ConvLSTM2D layer, max- pooling layers are applied to reduce the spatial dimensions of the feature maps, aiding in capturing higher-level features while reducing computational complexity. Furthermore, dropout layers are inserted using the Time Distributed wrapper after each max-pooling layer to further mitigate overfitting by randomly dropping out units during training. A depth wise convolution enabling the final dense layer processing is one of the architectural methods of the last flat convolutional layer for harnessing multidimensional output tensors to one dimension vector. A dense layer, on the contrary, serves as an input layer, and then a tangent hyperbolic activation function with a probability distribution yield for each activity class is loaded.

D. Training the Model

Our code can be started by creating an Early Stopping callback as Choice 1. Coming back on the validation loss during training has a high significance for tracking down the network. It begins the training stopping process in the event that validation loss failure to be lowered for a specified number of epochs known as the coefficient of attenuation, the patience parameter. This is the way of the mechanism which is stopping the training when the overfitting of the model begins to decrease the generalization ability of it. In addition to that, the model weights (the best ones in terms of validation loss) are going to be those that are restored in the model for the final evaluation. Finally, the callback that has been created comes to being used. During this process, the model is compiled with the categorical cross-entropy loss function which is a commonly used loss function for multi-class classification tasks like human activity recognition. The Adam optimizer is utilized for gradient-based optimization solving some challenges of vanilla SGD and demonstrating advantageous features like adaptive learning rates and momentum.

III. RESULT ANDANALYSIS

A. Result on the working model

The model's performance was evaluated using various metrics including a multilabel confusion matrix, precision, recall, F1 score, and accuracy. The multiclass confusion matrix offers insights into the model not only predictive performance but also in multiple classes' aspects. Each element of the batting matrix is a cell for the true negative (TN), true positive (TP), false negative (FN) and false positive (FP) predictions of each label.

TABLE.1 SHOWING THE THE TRUE POSITIVE RATE AND FALSE POSITIVE RATE

S. No	True Positive Rates	False Positive Rates
1	0.8919	0.1078
2	0.9375	0.0
3	0.8929	0.0
4	0.8077	0.0354

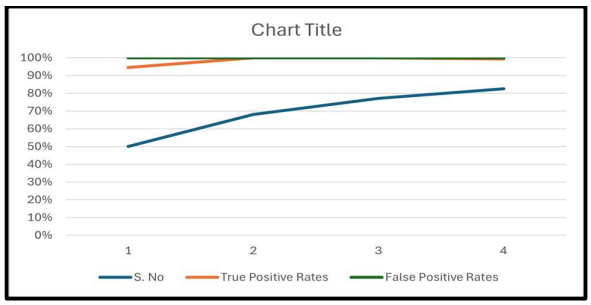


Fig. 9. Image showing the ROC values for the video image data set.

For instance, in the first class, there were 91 true negatives, 11 false positives, 4 false negatives, and 33 true positives. Precision, recall, and F1 score were computed using the micro-averaging method, which considers each instance

equally and calculates metrics globally across all classes. The precision micro-average was found to be 0.8, recall micro-average was 0.89, and F1 score micro-average was 0.892, indicating overall effectiveness in predicting the correct labels across all classes. Additionally, the accuracy of the model was computed to be 0.8.

B. Loss and Accuracy

In this study, we utilized a convolutional LSTM (ConvLSTM) model for our classification task and monitored its training and validation loss metrics. The plotted graph illustrates the changes in both training and validation loss over epochs.

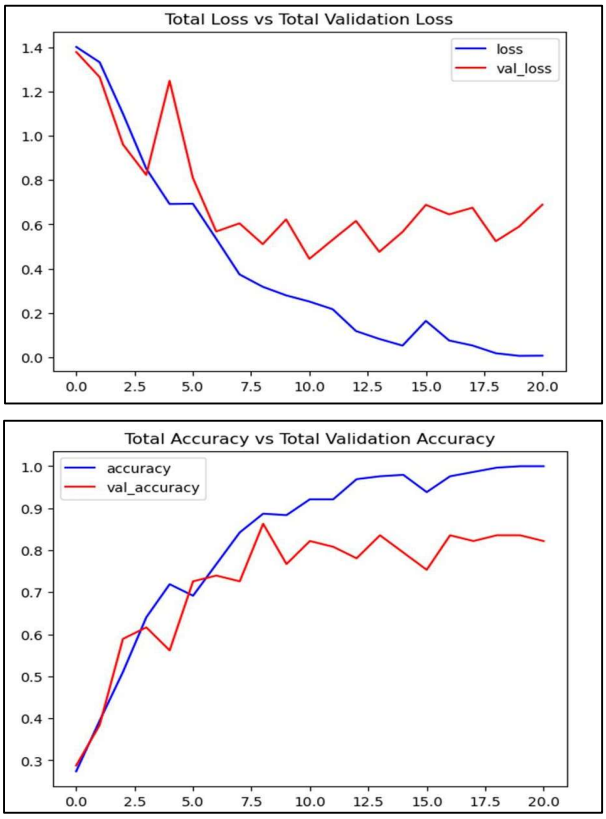


Fig. 10. Image showing the Graph validation for Loss and Accuracy

C. Predictions

In our case, the predictions may be the presence of certain desirable patterns in the input sequence of data. The performance evaluation will be based on comparison of these predictions with ground truth label through using standard metrics such as accuracy, precision, recall, and F1 score. Visualizations including confusion matrices should not be excluded from machine learning as they provide full detail about the model's forecasting capabilities and, thus help to determine if the model classifies cases correctly across the different classes. A review of predictions is where we gain an insight about the advantages and disadvantages of the ConvLSTM model which is of utmost importance in making any decision regarding the model changes and future applications of the same that can achieve better accuracy and generalization.

Fig. 11. Predicted Output on the working model

IV. CONCLUSION

A. The conclusion includest

This study has proved that ConvLSTM model is quite smart when capturing different human activities, even though the range of activities is diverse, and the environmental variation is strong, and the data is noisy. ConvLSTM models achieves this well by pooling the convolutional layers with LSTM cells that results in features that are hierarchical and the network thus, gets to retain the contextual information over time thus, more precise activity classification can be achieved even in the challenging environments like the dynamic ones, noisy ones or both. After a careful research and evaluation, in our work we have managed to indicate the outstanding advantages of ConvLSTM model against the traditional machine learning algorithms.

V. ACKNOWLEDGEMENTS AND FUTURE SCOPE

Mr. Ajay Sharma's guidance has been crucial in shaping this research. Through their in-depth knowledge and support, they have cultivated a solid ground that has finally enabled the completion of the project.

This research concentrated on identifying a small set of human activities in a controlled scenario. Future research may dive deeper into various activities and analyse the workability of the system in the most complex situations. On the other hand, the combination of deep learning algorithms or sensor fusion with video data could potentially lead to the improvement of the activity recognition system in terms of the accuracy and robustness. Furthermore, this study will be the basis for designing human activity recognition apps that can be used for different purposes such as assisted living for the elderly or personalized fitness coaching.

VI. REFERENCES

[1] Gadekallu, T. R. *et al.* Hand gesture recognition based on a Harris Hawks optimized Convolution Neural Network.

[2] Luwe, Y. J., Lee, C. P. & Lim, K. M. Wearable sensor-based human activity recognition with hybrid deep learning model.

[3] Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M. & Schmidhuber, J. Flexible, High-Performance Convolutional Neural Networks for Image Classification.

[4] Rajpurkar O.M., Kamble S.S., Nandagiri J.P., Nimkar A.V. Alert Generation on Detection of Suspicious Activity Using Transfer Learning.

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. Nothing else matters: Model-agnostic explanations by identifying prediction invariance (2016).

[6] Aghdam, H. H., Heravi, E. J. & Puig, D. Explaining adversarial examples by local properties of convolutional neural networks.

[7] Ajay Sharma, Ankit Gupta, Varun Jaiswal. Machine Learning for Intelligent Multimedia Analytics: Techniques and Applications.

[8] Palumbo, F., Gallicchio, C., Pucci, R. & Micheli, A. Human activity recognition using multisensory data fusion based on reservoir computing. *J. Ambient Intell. Smart Environ.*

[9] Yin, J., Yang, Q. & Pan, J. J. Sensor-based abnormal human-activity detection. *IEEE Trans. Knowl. Data Eng.* **20**, 1082–1090 (2008).

[10] Li, Y., Zhang, H., Xue, X., Jiang, Y. & Shen, Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, e1264.

[11] Ajay Sharma, Tarun Pal, Varun Jaiswal Decision support algorithms for data analysis.

[12] Alom, M. Z. *et al.* A state-of-the-art survey on deep learning theory and architectures. *Electronics* **8**, 292. (2016)

[13] Fan, L., Wang, Z. & Wang, H. Human Activity Recognition Model Based on Decision Tree. *2013 International Conference on Advanced Cloud and Big Data*.

[14] Shi, X., Li, Y., Zhou, F. & Liu, L. Human Activity Recognition Based on Deep Learning Method. *2018*

[15] Paul Viola and Michael Jones. Robust real-time object detection. In International Journal of Computer Vision, 2001.

[16] Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey. Machine Vision and Applications, 19(5-6):345–357,2008.

[17] Xia L., Li Z. A new method of abnormal behavior detection using LSTM network with temporal attention mechanism. J. Supercomput. 2021;77:3223–3241. doi: 10.1007/s11227-020-03391-y. [CrossRef] [Google Scholar]

[18] Lee J., Shin S. A Study of Video-Based Abnormal Behavior Recognition Model Using Deep Learning. Int. J. Adv. Smart Converg. 2020;9:115–119. [Google Scholar]

[19] Zhang J., Wu C., Wang Y., Wang P. Detection of abnormal behavior in narrow scene with perspective distortion. Mach. Vis. Appl. 2018;30:987–998. doi: 10.1007/s00138-018- 0970-7. [CrossRef] [Google Scholar]

[20] Vallathan G., John A., Thirumalai C., Mohan S., Srivastava G., Lin J.C.W. Suspicious activity detection using deep learning in secure assisted living IoT environments. J. Supercomput. 2021;77:3242–3260. doi: 10.1007/s11227-020-03387-8. [CrossRef] [Google Scholar]

[21] Ullah W., Ullah A., Hussain T., Muhammad K., Heidari A.A., Del Ser J., Baik S.W., de Albuquerque V.H.C. Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data. Futur. Gener. Comput. Syst. 2022;129:286–297. doi: 10.1016/j.future.2021.10.033. [CrossRef] [Google Scholar]

Updated_Final.docx

ORIGINALITY REPORT

7%	5%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	medium.com Internet Source	1%
2	journal.universitasbumigora.ac.id Internet Source	1%
3	Submitted to University of Greenwich Student Paper	1%
4	"Computer Vision – ECCV 2016", Springer Science and Business Media LLC, 2016 Publication	<1%
5	dokumen.pub Internet Source	<1%
6	ebooks.au.dk Internet Source	<1%
7	www.researchgate.net Internet Source	<1%
8	www2.mdpi.com Internet Source	<1%
9	www.researchsquare.com Internet Source	<1%

10	repository.tudelft.nl Internet Source	<1%
11	Jongju Kim, Heungseok Lee, June Ho Park. "Classification of Power System Stability Using Deep Learning", 2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2023 Publication	<1%
12	P. S. Nandhini, S. Kuppuswami, S. Malliga. "Chapter 65 Classification of Intrusions in RPL-Based IoT Networks: A Comparison", Springer Science and Business Media LLC, 2022 Publication	<1%
13	ebin.pub Internet Source	<1%
14	www.semanticscholar.org Internet Source	<1%
15	IFoA Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		