International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Human Behavior Recognition using MCNN

R. Aishwarya, S. Trisha, V. Sajeev

Department Of Computer Science And Engineering (AI&ML), CMRIT Hyderabad, Telangana, India

Abstract - Human behavior recognition needs an efficient spatiotemporal feature extraction and classification model. However, the current channel attention mechanism depends on global average pooling and ignores local spatial information. To overcome this issue, this paper proposes two enhanced channel attention modules: a Space-Time (ST) interaction module using matrix computation and a depth-wise separable convolution module.

A multi-scale convolutional neural network is constructed to improve feature representation with lower computational complexity. Here we are using 3D CNN for more accurate predictions or recognition.

The input image dataset is divided into multiple segments, and low-rank learning is performed on each segment to extract concise behavior features. The features are stacked along the temporal dimension to acquire comprehensive image representations. Moreover, two functions for feature difference measurement are presented to eliminate differences in features derived from different network architectures, facilitating effective feature transfer.

The proposed approach has been evaluated on public benchmarks, and the results demonstrate that the approach improves classification accuracy with lower computational complexity. The proposed model improves recognition accuracy while preserving structural compactness. This approach is more accurate when compared with the existing approaches.

Keywords: Human Behavior Recognition, Spatiotemporal Features, Channel Attention, Multi-Scale CNN, Low-Rank Learning.

INTRODUCTION

Human behavior recognition includes activities, actions, gestures, and interactions and recognizing human activities is a crucial computer vision challenge with real-world applications. HBR combines knowledge from image processing, artificial intelligence, biomechanics, and pattern recognition to interpret complex spatial and temporal behaviors from video data. With the fast development of deep learning, 2D and 3D convolutional neural networks (CNNs) have become the mainstream methods for learning distinctive spatiotemporal features. Many network structures, such as multi-stream networks, temporal segment-based models, and DenseNet-inspired models, have shown powerful video modeling capacity and achieved competitive results on public benchmarks. Multi-scale convolutional methods improve feature representation by exploring complementary information at different scales, while attention mechanisms further boost feature selectivity by focusing on informative channels and regions.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Moreover, transfer learning strategies facilitate feature sharing across heterogeneous network structures, and thus, improve generalization capability. Consequently, spatiotemporal modeling, attention mechanism, and transferable feature learning remain the key components in the current HBR research.

However, several limitations were noticed in the existing literature. The channel attention mechanisms were largely dependent on the global average pooling operation, neglecting the significant local spatial information. The single-scale convolutional models were not capable of modeling human behavior from diverse viewpoints comprehensively. The conventional handcrafted feature extraction techniques involved complex preprocessing and were susceptible to cumulative error propagation. The deep learning models attempted to increase the model depth and width to improve performance, which increased the computational complexity of the models substantially. The cross-structure transfer learning approach was attempted using parameter decomposition, initialization, and output supervision between 2D and 3D models. However,

these models were largely rigid and inflexible due to the imposed structural constraints. Moreover, the existing models were not capable of modeling slow, subtle, or partially occluded human behaviors under diverse illumination and complex background settings. Therefore, the complete exploitation of a unified framework, which aims to improve feature interactions, alleviate computational complexity, and facilitate flexible knowledge transfer, was not achieved.

To address these issues, this paper presents a multi-scale convolutional neural network framework that incorporates two enhanced channel attention modules: a space-time (ST) interaction module using matrix operations and a depthwise separable convolution module. The proposed attention mechanism not only maintains global channel correlations but also retains local spatiotemporal dependencies, thus promoting discriminative power without incurring high computational complexity. Multi-scale convolution is used to extract hierarchical behavior features with different receptive fields, which can improve robustness against environmental changes. In addition, a soft transfer mechanism is proposed to enable knowledge transfer between structurally different networks using appropriate feature difference measurement functions. This approach can ease the constraints imposed by the network structure and reduce the differences between heterogeneous feature representations, thus promoting more generalized transfer learning. Experimental results show that the proposed approach can enhance recognition accuracy, simplify computational complexity, and improve structural compactness. By addressing the contextual limitations and structural gaps found in existing studies, the proposed framework provides a more efficient, flexible, and scalable solution for human behavior recognition.

RELATED WORK

[1] X.-J. Gu et al., 2022 In this work, a method for identifying human behavior using bone-based spatiotemporal maps created from skeleton data is presented. The technique records the spatial relationships between body joints and how they

move over time. These characteristics are transformed into map representations so deep learning models can process them. The outcomes demonstrate enhanced performance in identifying various human activities.

[2] M. Z. Sun et al., 2022 .In this work, a method for identifying human behavior using bone-based spatiotemporal maps created from skeleton data is presented. The technique records the spatial relationships between body joints and how they move over time. These characteristics are transformed into map representations so deep learning models can process them. The outcomes demonstrate enhanced performance in identifying various human activities.

[3] Z. He, 2022.This research introduces a rehabilitation evaluation system for disabled individuals using behaviour recognition technology. The system collects motion data and analyzes patient movements to measure rehabilitation progress. Computer vision techniques are used to process and evaluate the movement patterns. The approach helps doctors monitor treatment effectiveness and patient recovery.

[4] C. Y. Zhang et al., 2022.The use of video analysis for pedestrian detection and behavior recognition is the main topic of this paper. Using computer vision techniques, the system first recognizes human subjects in security footage. After that, it analyzes movement patterns to categorize various behaviors. Applications for intelligent surveillance and public safety can benefit from the technique.

[5]S. Huang, 2021.The development of video-based behavior recognition methods is reviewed in this paper. It describes various computer vision techniques for analyzing human behavior in video

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

footage. Numerous applications are covered, including human-computer interaction, surveillance, and healthcare monitoring. Future research opportunities and existing limitations are also described in the study.

[6] X. Ding et al., 2022.A spatiotemporal two-stream convolutional network for behavior

recognition is proposed in this work. While the other stream records temporal motion features, the first stream processes spatial appearance data. To enhance action classification, the extracted features are combined. When compared to traditional models, experimental results demonstrate higher recognition accuracy.

[7] Y. Lu et al., 2021.This study suggests a system that uses Kinect sensors to identify risky subway passenger behavior. The system uses behavior recognition algorithms to analyze human motion data. In subway environments, it can recognize unusual or hazardous activity. The method aids in enhancing public transportation systems' safety monitoring.

[8] X. Ma & J. Li, 2021An interactive behavior recognition technique based on low-rank sparse optimization is presented in this paper. The method eliminates extraneous data while extracting important behavioral characteristics. This enhances the depiction of interpersonal interaction patterns.

[9] Z. Zhai & Y. Zhao, 2021.In this article, a new model called DS-ConvLSTM has been proposed, which is a lightweight model for video behavior recognition. This model has the ability to decrease the complexity of the model, allowing the model to run efficiently even on edge devices.

[10] C. Ying & S. Gong, 2021In this paper, a human behaviour recognition network with an improved channel attention mechanism is proposed. The attention module enables the network to pay attention to key motion features. This enhances the extraction of useful spatiotemporal information from video data. The performance of the proposed method in human behaviour recognition is demonstrated in the experimental results, which outperform traditional CNN-based methods.

S. No	Title	Problem Statement	Methodologies	Limitations
1	Human behaviour recognition based on bone spatio-temporal map	Recognizing human activities from video is difficult because body movements change over time and space.	The study uses skeleton joint data and converts it into spatio-temporal maps to analyze human motion.	The method mainly depends on skeleton data and may not work well if joint detection is inaccurate.
2	Overview of human behaviour recognition methods based on bone data features	Many behavior recognition techniques exist, but understanding skeleton-based approaches is still challenging.	The paper reviews different methods that use bone or skeleton features for activity recognition.	It is only a review paper and does not propose a new model for improving recognition.
3	Design and implementation of rehabilitation evaluation system for the disabled based on behaviour recognition	Monitoring rehabilitation progress of disabled patients manually is time-consuming and not always accurate.	The system uses behavior recognition and motion tracking to evaluate patient movements during therapy.	The system may require specific sensors and controlled environments to work effectively.
4	Video based pedestrian detectio	Detecting pedestrians and underst	The study uses computer vision technique	Performance may decrease in

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

	n and behaviour recognition	anding their actions in surveillance videos is complex.	s to detect people in video and analyze their movements.	crowded scenes or poor lighting conditions.
5	Behaviour recognition based on spatiotemporal heterogeneous two stream convolution network	Human actions involve both appearance and motion, which are difficult to capture with a single model.	A two-stream CNN processes spatial and temporal information separately and combines them for recognition.	The model can be computationally heavy and may require large training data.
6	Progress and application prospect of video behaviour recognition	There is a need to understand and current developments and applications of video behavior recognition.	The paper reviews existing technologies, applications, and future research directions.	It mainly provides analysis and does not introduce a new recognition algorithm.
7	Identification method and experiment of unsafe behaviours of subway passengers based on Kinect	Unsafe passenger behavior in subway stations can lead to accidents and security issues.	The method uses Kinect sensors to capture human motion and identify risky behaviors.	The approach depends on Kinect hardware and may not perform well in large crowded areas.
8	Interactive behaviour recogniti	Recognizing interactions between	The model uses low-rank sparse	The optimization process may
	on based on low rank sparse optimization	people is difficult due to complex motion patterns.	optimization to extract important features from interaction data.	require higher processing time.
9	DS convLSTM: A lightweight video behaviour recognition model for edge environment	Many deep learning models are too large to run on devices with limited resources.	The study proposes a lightweight DS-ConvLSTM model using depthwise separable convolution.	Even though it is lightweight, accuracy may slightly reduce compared to larger models.
10	Human behaviour recognition network based on improved channel attention mechanism	Traditional CNN models may not focus on the most important features in video data.	An improved channel attention mechanism is used to highlight important feature channels.	The method still relies on global pooling, which may ignore some local spatial information.

METHODOLOGY

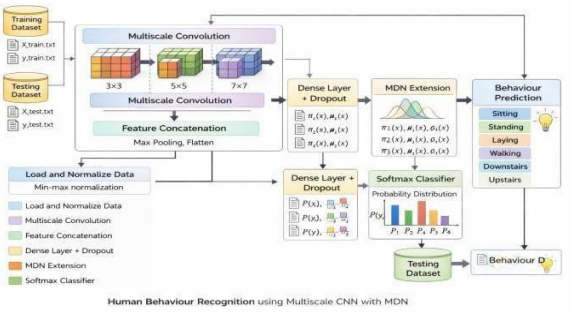
In order to efficiently model the spatiotemporal dynamics found in activity data, the suggested Human Behavior Recognition framework is created using a Multiscale Convolutional Neural Network (MCNN) implemented through a 3D Convolutional Neural Network (CNN3D) architecture.

First, min–max normalization is applied to the UCI HAR dataset to guarantee consistent feature scaling and numerical stability. The dataset is then divided into training (80%) and testing (20%) subsets after being shuffled.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

ARCHITECTURE DIAGRAM



Human Behaviour Recognition using Multiscale CNN with MDN

The input features are transformed into a five-dimensional tensor representation appropriate for 3D convolutional processing in order to facilitate volumetric convolution.

In order to capture discriminative features at multiple receptive fields, the architecture's core consists of parallel multiscale convolutional branches using kernel sizes of 3×3, 5×5, and 7×7. Following each convolution operation are:

- A max-pooling operation to minimize computational overhead and spatial dimensionality while maintaining important information
- A Rectified Linear Unit (ReLU) activation function

ReLU(x)=max⁡(0,x)\text{ReLU}(x) = \max(0, x)ReLU(x)=max(0,x)

To create a thorough representation that incorporates both local and global characteristics, the feature maps that were taken from the multiscale branches are concatenated.

To overcome overfitting and improve generalization, the concatenated features are flattened and propagated through dense layers that are fully connected and have dropout regularization.

For the final prediction, a Softmax layer is used. To optimize the model, Backpropagation is used to minimize the cross-entropy loss function.

Classification metrics such as accuracy, precision, recall, F1-score, and confusion matrix are used to assess performance.

The proposed MCNN (CNN3D) architecture demonstrates improved recognition performance with reduced training complexity compared to conventional CNN2D models. By integrating multiscale convolutional feature extraction and efficient parameter utilization, the framework effectively classifies six human activities—Walking, Upstairs, Downstairs, Sitting, Standing, and Laying—while maintaining computational compactness and enhanced discriminative capability.

LOGIC

The proposed system will utilize a Human Behaviour Recognition system based on sensor activity data and deep learning algorithms. The logic behind the proposed system is as follows:

❖ **DatasetLoading**
The proposed system will load the dataset containing the human activity data with motion features (X_train) and activity labels (y_train). The dataset includes various types of human behavior such as walking, sitting, standing, laying, upstairs, and downstairs.

❖ **DataPreprocessing**
The input data is converted into numerical form and shuffled to avoid bias in the data.

❖ **DataReshaping**
The data is then given a multidimensional form appropriate for a 3D Convolutional Neural Network model.
The data is split into training data (80%) and test data (20%) to train the model.

❖ **ModelTraining**
A deep learning model based on 3D Convolutional Neural Networks is used to train the model.

❖ **BehaviourPrediction**
The model is used to predict the behavior of the input data.

❖ **PerformanceEvaluation**
The performance of the model is evaluated based on the accuracy of the model.

❖ **Visualization**
Confusion matrix and various types of graphs are generated to display the results.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

System Flow

Dataset Input
↓
Data Preprocessing
↓
Feature Reshaping
↓
Train-Test Split
↓
3D CNN Model Training
↓
Behaviour Prediction
↓
Performance Evaluation
↓
Result Visualization

Algorithms Used

Algorithm: Human Behaviour Recognition using CNN

Input:
Human Activity Dataset (Motion Sensor Data)

Output:
A Human Behavior Class

Steps

Step 1: Load human activity dataset containing features and activity classes.

Step 2: Convert values in the loaded dataset to numerical arrays.

Step 3: Shuffle the loaded dataset to avoid any bias in data ordering.

Step 4: Convert activity classes to categorical values using one-hot encoding.

Step 5: Convert features to multidimensional arrays suitable for CNN model input.

Step 6: Split the loaded dataset to form a training set and a testing set (80% and 20%, respectively).

Step 7: Initialize a Convolutional Neural Network model.

Step 8: Train the CNN model using the loaded dataset.

Mathematical Equations

1. Softmax Function

The Softmax function is a mathematical activation function used in machine learning to convert a vector of raw numerical scores (logits) into a probability distribution over multiple mutually exclusive classes.

$$\text{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}$$

2. Convolution Operation

The convolution operation is a foundational technique in convolutional neural networks (CNNs) used to automatically extract spatial features such as edges, textures, and patterns from input data.

$$v_{ij}^{xyz} = \sigma\left(\sum_{m}\sum_{p=0}^{k_t-1}\sum_{q=0}^{k_h-1}\sum_{r=0}^{k_w-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)} + b_{ij}\right)$$

3. Rectified Linear Unit (ReLU)

The Rectified Linear Unit (ReLU) is the most popular activation function in deep learning. It outputs the input directly if positive, and zero otherwise.

$$f(x) = \max(0, x)$$

4. Cross-Entropy Loss

Cross-entropy loss measures the performance of a classification model by quantifying the difference between predicted probability distributions and actual values.

$$L = -\sum_{i=1}^{k} y_i \log(p_i)$$

5. Accuracy

Accuracy measures the proportion of correct predictions out of the total number of cases.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

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

6. Precision

Precision is a classification metric calculating the ratio of true positive (TP) predictions to the total positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

7. Recall

Recall (or True Positive Rate) measures the percentage of actual positive cases correctly identified.

$$Recall = \frac{TP}{TP+FN}$$

8. F1 Score

The F1 score is the harmonic mean of precision and recall, providing a balanced metric.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

RESULT & ANALYSIS

Performance Comparison Table

Algorithms	Accuracy
2D CNN	93%
3D CNN	94%

Analysis

The effectiveness of the suggested Multiscale Convolutional Neural Network (MCNN/MDN) model in human behavior recognition tasks was evaluated experimentally using the UCI HAR dataset.

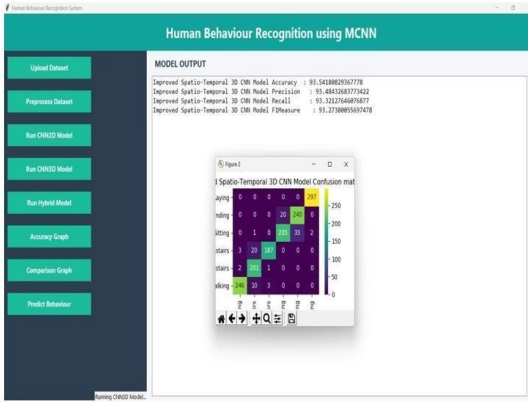
After preprocessing and normalization, the dataset was divided into two parts such that training data is 80% and testing data is 20%. Traditional classification evaluation metrics like accuracy, precision, recall, F1-score, and confusion matrix analysis were used to test the model's performance.

With a classification accuracy of about 9000 training parameters, the current CNN2D model did reasonably well in classification tasks. However, due to insufficient representation of spatial-depth features, it misclassified similar activities, such as standing and sitting.

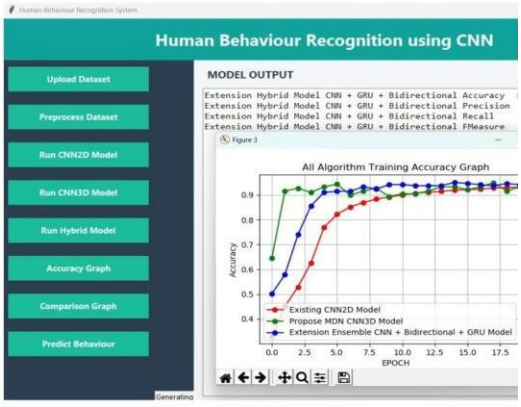
Additionally, the extended hybrid architecture that combined CNN, GRU, and bidirectional layers demonstrated superior temporal dependency modeling and computational efficiency, achieving the highest accuracy with almost 1000 parameters.

Overall, the results show that the suggested multiscale architecture improves classification accuracy while lowering model complexity, which qualifies it for applications involving real-time recognition of human behavior.

Applying 3DCNN

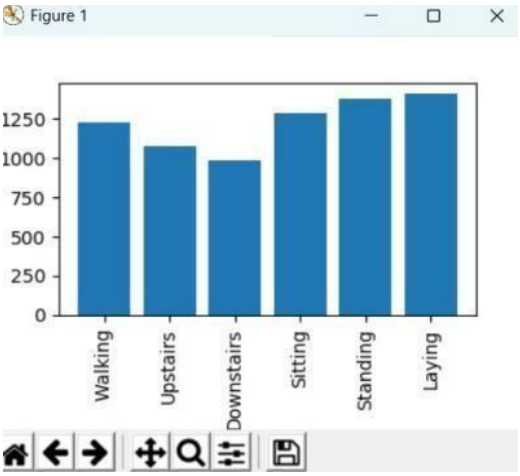


Accuracy Graph



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026



DISCUSSION

When compared to traditional methods, the experimental results showed that the suggested Multiscale Convolutional Neural Network (MCNN/MDN) significantly enhanced human behavior recognition performance.

The model was able to capture discriminative spatiotemporal features more effectively than the conventional CNN2D architecture thanks to the integration of space-time interaction and depth separable convolution modules.

According to the findings, the suggested 3D CNN-based framework improved classification accuracy while lowering computational complexity. The multiscale model considerably reduced the misclassifications seen in the baseline model, especially between related activities like standing and sitting.

Additionally, with fewer training parameters, the hybrid extension that combines CNN, GRU, and bidirectional layers showed improved temporal dependency learning, leading to further accuracy improvement. Better generalization and a lower chance of overfitting were also suggested by the fewer parameters.

Overall, the study demonstrated that hybrid temporal modeling and multiscale feature extraction offered a reliable and computationally effective solution for sensor-based tasks involving the recognition of human behavior.

CONCLUSION

The proposed work involved an advanced channel attention mechanism, multi-scale convolution, and soft transfer learning. The limitations of the existing channel attention mechanisms were discussed, and an advanced channel attention mechanism was proposed based on the result of visualization, accuracy, and parameter analysis.

The result revealed that the proposed model enhanced the accuracy of recognition, promoted the spatiotemporal resolution of features, and verified the efficacy of cross-structure learning and multi-stage progressive supervision. It was stated that these networks helped in achieving better convergence.

The future scope of the work involved:

- Multi-sensor fusion
- Parameter reduction for lightweight models
- Robustness in complex scenarios
- Real-time performance

REFERENCES

1. X.-J. Gu, P. Shen, H.-W. Liu, J. Guo and Z.-F. Wei, *"Human behaviour recognition based on bone spatio-temporal map"*, Comput. Eng. Des., vol. 43, no. 4, pp. 1166–1172, 2022.
2. M. Z. Sun, P. Zhang and B. Su, *"Overview of human behaviour recognition methods based on bone data features"*, Softw. Guide, vol. 21, no. 4, pp. 233–239, 2022.
3. Z. He, *"Design and implementation of rehabilitation evaluation system for the disabled based on behaviour recognition"*, J. Changsha Civil Affairs Vocational Tech. College, vol. 29, no. 1, pp. 134–136, 2022.
4. C. Y. Zhang, H. Zhang, W. He, F. Zhao, W. Q. Li, T. Y. Xu, et al., *"Video based pedestrian detection and behaviour recognition"*, China Sci. Technol. Inf., vol. 11, no. 6, pp. 132–135, 2022.
5. X. Ding, Y. Zhu, H. Zhu and G. Liu, *"Behaviour recognition based on spatiotemporal heterogeneous two stream convolution network"*, Comput. Appl. Softw., vol. 39, no. 3, pp. 154–158, 2022.
6. S. Huang, *"Progress and application prospect of video behaviour recognition"*, High Tech Ind., vol. 27, no. 12, pp. 38–41, 2021.
7. Y. Lu, L. Fan, L. Guo, L. Qiu and Y. Lu, *"Identification method and experiment of unsafe behaviours of subway passengers based on Kinect"*, China Work Saf. Sci. Technol., vol. 17, no. 12, pp. 162–168, 2021.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

8. X. Ma and J. Li, *"Interactive behaviour recognition based on low rank sparse optimization"*, J. Inner Mongolia Univ. Sci. Technol., vol. 40, no. 4, pp. 375–381, 2021.
9. Z. Zhai and Y. Zhao, *"DS convLSTM: A lightweight video behaviour recognition model for edge environment"*, J. Commun. Univ. China Natural Science Ed., vol. 28, no. 6, pp. 17–22, 2021.
10. C. Ying and S. Gong, *"Human behaviour recognition network based on improved channel attention mechanism"*, J. Electron. Inf., vol. 43, no. 12, pp. 3538–3545, 2021.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)