Mae Mai Muay Thai Style Classification in Movement Appling Long-Term Recurrent Convolution Networks

Shujaat Ali Zaidi[1] and Varin Chouvatut[2*]

[1]Research Scholar, Department of Computer Science, Faculty of Science, Chiang Mai University, Thailand. shujaatxaidi@gmail.com, ORCID: https://orcid.org/0000-0002-5100-3805

[2*]Assistant Professor, Department of Computer Science, Faculty of Science, Chiang Mai University, Thailand. varin.ch@cmu.ac.th, ORCID: https://orcid.org/0009-0008-9125-4650

Received: December 22, 2022; Accepted: February 20, 2023; Published: February 28, 2023

Abstract

The research community has become more interested in human activity recognition due to improvements in technology and machine learning algorithms. In particular, when we discuss the automated assessment of athletic talents, which has been the most active study topic over the last decade. The highly competitive nature of sports necessitates collecting accurate data on an athlete's performance to evaluate their activities while competing accurately. In this article, we present a method for identifying seven typical styles of Mae Mai Muay Thai (MMMT) using continuously collected boxing sequences as the data source. Long-term Recurrent Convolution Networks (LRCNs) is used to handle MMMT recognition. Additionally, we combined Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNNs) classifiers for experimental testing. According to experimental testing, our training strategy outperformed the use of both CNNs and LSTM classifiers. Experiments were conducted utilizing the MMMT dataset with four professional boxers as participants. The LRCNs classifiers were able to achieve a 99 percent accuracy rate that proves the LRCNs algorithm is appropriate for assessing the boxer's abilities while competing. Furthermore, we will determine the overall usefulness of the model by using a confusion matrix in our analysis. Additional performance metrics included in the investigation were accuracy, precision, recall, and the F1-score.

Keywords: Computer Vision, Machine Learning, Long-term Recurrent Convolution Networks, Convolutional Neural Network, Long Short-term Memory, Image Classification, Boxing.

1 Introduction

Sports video based action recognition is an exciting and challenging field to study, especially when finding and identifying Mae Mai Muay Thai (MMMT) actions in a video stream (Shujaat et al., 2022). Video action recognition has many uses (Costa et al., 2019), such as surveillance systems for privacy and security, content-based video recovery, activity recognition, and human-computer interaction (Keshavarzian et al., 2019). Due to the rapid expansion of digital content in the modern era, robust artificial intelligence-based is required for surveillance purposes in order to monitor and recognize boxer behaviors and activities. The purpose of MMMT action recognition is to recognize and identify techniques, their style. Action recognition still presents many challenges when it comes to ensuring the

Journal of Internet Services and Information Security (JISIS), volume: 13, number: 1 (February), pp. 95-112.
DOI: *10.58346/JISIS.2023.I1.010*

*Corresponding author: Assistant Professor, Department of Computer Science, Faculty of Science, Chiang Mai University, Thailand.

security and safety of residents (Spolaror et al., 2020) because of substantial modifications in camera angles, occlusions, complex backgrounds, and variations in illumination (Aggarwal et al., 2021). These challenges include manufacturing surveillance, violence sensing, virtual reality, cloud environments, and person identification. Temporal and spatial information are critical in detecting various MMMT in videos. In the previous decade, most strategies used handcrafted engineering to represent the spatial aspects of dynamic states for describing the related action in videos. Because of how people move and how complicated the background is, the hand-made features method for action identification is mainly database-based and cannot work in all situations. In this way, representational motion characteristics and traditional techniques get better as they move from two dimensions to three dimensions in order to get accurate data. Such strategies convert spatial characteristics to 3D spatiotemporal characteristics to collect real-time information in frames (Li et al., 2019).

Deep learning is the primary and frequently used approach for learning high-level racist and salient discriminatory features and end-to-end building systems in video-based action and MMMT identification (Dai et al., 2019). Deep learning systems for human action recognition using convolutional neural network procedures in convolution operation to learn characteristics from video frames using per-trained models. These convolutional layers retrieve and analyze spatial information to train a classification model. In comparison, the conventional CNN models perform worse than handwrought features in sequential data, Alex Net, and other standard CNN models from a single input picture, Res Net and VGG learn spatial features (Dai et al., 2019). These models are adequate for gathering spatial data. However, they are ineffective for time series data, which is a disadvantage. A crucial component in capturing motion data for MMMT in a video series, for example, employed the long short-term memory (LSTM) for activity recognition learned using a CNN with spatiotemporal information. High-level human action recognition approaches based on video need a two-stream architecture. Design distinct modules that learn spatially and combine methods to detect temporal characteristics in video sequences. Dynamic information may be captured in time series data (Dai et al., 2020). Recurrent neural networks (RNNs) have been used to address spatiotemporal difficulties. The LSTM is specially built for long-term video sequences to learn and process temporal characteristics for human action recognition in video surveillance (Kwon et al., 2018). Most researchers have devised a two-stream strategy for MMMT action recognition that combines spatial and temporal characteristics for joint feature training to address MMMT action recognition's present constraints and limits. Based on these facts, exact MMMT action detection in real-world recordings remains challenging due to a lack of information regarding motion, style, and backdrop clutter for the appropriate identification of MMMT styles. Traditional techniques failed to solve these challenges because of concerns with managing continuous actions, simulating crowded situations due to occlusion, and noise sensitivity (Baccouche et al., 2011). Similarly, existing approaches for MMMT action recognition handled the sequence learning issue utilizing RNNs, LSTMs,Long-term Recurrent Convolutional Networks (LRCNs) and gated recurrent units but without concentrating on information sources in sequences, which is critical for maintaining a link between previous and future frames.

The difficulty of MMMT action recognition is problematic not only because of the subtle differences between MMMT techniques explain by (Muaysena, 2022) and (Muayboran, 2022), which require the careful selection of characteristics, but also because of the boxer's stance and whether the style is used to the body or the head. This necessitates the careful selection of characteristics. The boxers' position patterns in each MMMT category are altered as a result of this factor. Boxers' elbow flexion angles are influenced by the length and angle of the strikes they deliver, whether they are short or lengthy. It is crucial to have a thorough understanding of the sequence in which a boxer's body stances are performed, particularly the placements of their hands and elbows. It is essential to have a video so that you can

observe the boxers' arms without being hindered in your ability to differentiate the blows that they deliver as they are going about the ring.

Despite research on estimating human body position (Si et al., 2019), and Motion detection (Guo et al., 2014), more research needs to be done on MMMT action. Even while it is vital to recognize the style, recognizing identification with little obstruction. Recent Kinect research showed that skeletal estimation performed effectively in experimental circumstances but was less effective if the human body was either not erect or partly covered (Xia et al., 2013). Although lot of work has been done to enhance action recognition, but MMMT recognition performance is currently difficult to determine from videos because of the complexity of the human body while fighting.

In this paper, we will recognize and classify seven MMMT techniques (Muayboran, 2022) out of 15: Salab Fan Pla, Paksa Waek Rang, Chawa Sad Hok, Inao Thang Grit, Yo Khao Prasumeru, Ta Then Kam Fak, Mon Yan Lak. Unfortunately, there have not been many efforts made by the scientific community to find a solution to this unsolved issue. The following are the study's main contributions:

- Our main objective is to categorize and identify the MMMT style utilizing the Long-term Recurrent Convolution Networks architecture.
- The complex architecture of MMMT has never been tried to be understood.
- We detect and categorize the MMMT from animation.
- The proposed method to classify the MMMT style deviates from the works of Kasiri-Bidhendi (2015), Shen (2017) and Chantaprasert (2019).

1.1. Mae Mai Muay Thai (MMMT)

Mae-Mai the term "Muay Thai" refers to the core fighting methods that are said to be used in Muay Thai combat (Muayboran, 2022). The trainee must first master and practice the necessary skills before being taught Look-Mai or other less critical techniques. Several body parts, including the punch, the foot, the knee, and the elbow, are employed in this combat sport. Since Thai literature and the study of nature served as their primary sources of inspiration, the names of these disciplines have taken on poetic and profound overtones. The Mae- Mai Muay Thai style has 15 variations (Chankuna, 2006). Nevertheless, this article will talk about seven of them. Because these seven boxing styles are challenging to discern which boxing philosophies a fighter employs during a match, we choose to utilize them. However, this article will just talk about seven of them. They are as follows:

1.1.1 Salab Fan Pla

Salab Fan Pla (Figure 1) is the critical move to defend against or get away from an opponent's straight fist by stepping out of the circle of arms and letting the fist go by the face (Chankuna, 2006). The attacker strikes the defender with a straight left fist while stepping forward with the left foot. For defense, step the right foot obliquely right side 1 step and constantly move the body to the right side with the load on the right foot, the right leg bent a bit, to move the head and body away from the attacker's fist.

Figure 1: Salab Fan Pla

1.1.2 Paksa Waek Rang

The instructor uses Paksa Waek Rang (Figure 2) as a move-in technique and other maneuvers (Chankuna, 2006). The attacker moves with their left foot forward after throwing a straight left punch to the defender's face. In order to defend attacker, take a quick step forward obliquely to the left side of the opponent's left arm, and place your weight on your left foot. Next, bend both arms to block the upper and lower portions of the attacker's arm, with your fists close to one another and your elbows slightly apart. Next, cover your head and face with both of your arms as you cast a glance.

Figure 2: Paksa Waek Rang

1.1.3 Chawa Sad Hok

The primary move (Chankuna, 2006) to avoid a straight punch by stepping out and countering with an elbow is Chawa Sad Hok (Figure 3). The assailant comes forward with his left foot and delivers a straight left hand into the defensive's face. Stepping, turning the body to the right, weighing the right foot, bending the left arm, and striking the attacker's ribs with the elbow are all defensive maneuvers.

Figure 3: Chawa Sad Hok

1.1.4 Inao Thang Grit

The primary fundamental method of Inao Thang Grit (Figure 4) employs an elbow near the torso while deforming a straight fist (Chankuna, 2006). The attacker moved forward after throwing a straight left punch into the defensive's face. To defend oneself, one swiftly advances with the left foot while maintaining a 60-degree angle to the approximately left side of the body and placing weight on the left foot. Next, one bends their right elbow to be parallel to the ground and throws it at the attacker's ribs.

Figure 4: Inao Thang Grit

1.1.5 Yo Khao Prasumeru

Yo Khao Prasumeru (Figure 5) defended himself by bending his body and using a straight punch. Close up, let the fist travel over the head before throwing it up to the chin (Chankuna, 2006). The attacker advances with the right foot while simultaneously throwing a straight right fist in the direction of the defender. To defend, take a rapid stride with the left foot, bending the body down to the front with weight on the left foot and throwing up the right fist beneath the attacker's chin.

Figure 5: Yo Khao Prasumeru

1.1.6 Ta Then Kam Fak

The primary fundamental (Chankuna, 2006) utilized to defend against fists to the chin is Ta Then Kam Fak (Figure 6). Technique for using the arm to push the attacker's fists away. The attacker strikes the defender with a straight left fist. Simultaneously moves the left foot forward. In order to defend, the defender closes the distance with the attacker by stepping forward with their left foot and bending their right arm to push their left fist out.

Figure 6: Ta Then Kam Fak

1.1.7 Mon Yan Lak

Mon Yan Lak (Figure 7) is a crucial master skill (Chankuna, 2006). This Mae-Mai used a kick to the top of the chest or the abdomen to block fists. The assailant comes forward with his left foot and delivers a straight left hand. Always turn your right foot outward and 45 degrees to the right while defending. Bends both arms to protect the face while simultaneously tossing the left foot to the top of the attacker's chest or abdomen to push him away.

Figure 7: Mon Yan Lak

The rest of this article is organized as follows. Section 2 reviews various publications on the classifiers used in our investigation, including recurrent neural networks (RNN), long term recurrent convolution networks (LRCNs), Convolutional Neural Networks (CNNs), and long short-term memory (LSTM). Section 3 explains the whole technique that we use in our work. Section 4 displays the experiment and results of our strategy. In Section 5 discussion and comparison. Section 6 concludes with research suggestions.

2 Background and Related Work

Our approach is inextricably tied to recurrent neural networks (RNNs), long-term recurrent convolution networks (LRCNs), Convolutional Neural Networks (CNNs), and long short-term memory (LSTM) classifiers. The following are the corresponding literature reviews for these areas.

2.1. Recurrent Neural Networks (RNNs)

Recurrent neural networks are a kind of artificial neural network that uses a unique looping topology to enable ongoing information connected to prior knowledge. They are used in several contexts involving data with sequences. Directed cycles are present in RNN in addition to having the feed-forward neural network's structure. Because of the network's ability to circulate information, each time's output is tied to both the current and earlier timestamps input (Dhruv et al., 2020). RNNs are used to analyze time series data in the pursuit of discovering patterns, such as predicting the stock market, modeling languages, classifying images, and generating text and speech, among other things (Schmidhuber, 2015). Through the feedback links, it picks up on the dependencies in a time series of data. The findings on picture identification, segmentation, detection, and retrieval are state-of-the-art (Fan et al., 2021).

RNNs were initially established in the 1980s, but they have lately garnered interest owing to several scientific and technological breakthroughs that make them computationally efficient for training (Fan et al., 2021). This is the case since these developments have made RNNs more effective. RNNs are not the same as feed-forward networks since they have an effect on a certain kind of neural layer known as recurrent layers. These layers enable the network to save its state between instances of the network being used. RNNs may be constructed using a wide variety of architectural styles. The feedback inside the network is one of the primary distinctions between the two topologies. RNNs can typically be unfolded in time and trained using back-propagation through time. This is a form of training in which the same set of weights is used for a layer across various timesteps, and the weights are updated using gradients that are analogous to the back-propagation algorithm (Lokhande et al., 2015).

2.2. Long-term Recurrent Convolutional Networks (LRCNs)

Long-term Recurrent Convolutional Networks were a concept that a group of researchers developed in the year 2016 (Donahue et al., 2015). They proposed a class of trainable architectures that could be used from beginning to finish for visual recognition and description. The fundamental concept is to utilize CNNs to learn visual characteristics from video frames and LSTMs to turn a series of image embeddings into a class label, phrase, probability, or anything else that may be required. This may be accomplished by combining the two types of neural networks. Therefore, the unprocessed visual data is run through a CNN, and the results of that processing are then put together into the stack of recurrent algorithms. The long-term Recurrent Convolutional Network (LRCNs) model for tasks requiring sequential data, whether visual, linguistic, or otherwise, combines a deep hierarchical feature extractor with a model that can learn to detect and synthesize temporal dynamics.

2.3. Convolutional Neural Networks (CNNs)

Convolutional neural networks have significantly advanced science (Li et al., 2021). It has grown to be one of the best-known and most renowned deep-learning neural networks (Rodriguez et al., 2022). The use of machine vision powered by CNN's algorithms has enabled people to do tasks that were previously unthinkable. The layered architecture of CNN consists of both primary layers and auxiliary layers. The fundamental layers used by CNN include convolution layers, activation layers, pooling layers, flatten layers, and dense layers (Kandel et al., 2020). A CNN may become more resilient against overfitting and generalizable by adding more layers. Regularization, dropout, and batch normalization are a few of these layers (Yang et al., 2021).

2.3.1 Convolutional Layer

In a CNNs, the convolution layer is the first and most important since it can dynamically retrieve visual information without requiring explicit specification.

2.3.2 Activation Layer

Convolution layers are often followed by activation layers, a nonlinear layer that is essential as a selection criterion for choosing which neurons will fire. The activation layer receives an actual number as input from a nonlinear function. Activating layers are essential because they enable networks to understand nonlinear mappings and increase their resistance to difficult processes. Sigmoid, ReLU, Tanh, softmax, and LeakyReLU are the most often used activation layers in CNNs.

2.3.3 Pooling Layer

We might decrease the amount of processing and the number of variables the network needs to handle by sandwiching a pooling layer between two convolution layers. A pooling layer reduces the feature map created by a convolution layer by only including useful information. Filter size and stride are the two most crucial pooling layer variables. The pooling system must operate with both maximum and average pooling.

2.3.4 Flattening Layer

The outputs of the pooling layer are then simplified to a one-dimensional matrix since the dense layers that come after it can only work with data in this shape.

2.3.5 Dense Layer

The network's output is often close to dense layers, also known as fully linked layers, which get feature extraction results. The main goal of the dense layer is to classify the original image by considering all the data gathered from the previous layers.

2.4. Long Short-term Memory (LSTM)

The Long Short-Term Memory (LSTM) architecture is a variant of the recurrent neural network that was developed by (Hochreiter et al., 1997) to address gradient explosion (Gers et al., 2002) or decline brought on by lengthy backpropagation delays during the training of RNN designs (Lipton et al., 2015). One approach to visualize LSTM is using a framework of LSTM units (see Figure 8). Each LSTM unit has three gates that regulate the information flow:

- Input gates evaluate whether or not an input is important enough to remember.
- Forget gates, which govern whether the unit should remember the value or not.
- Output gates decide whether or not the value should be extracted by the unit.

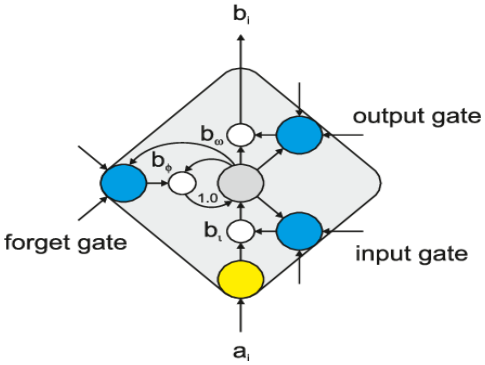

Figure 8: LSTM Gates (Sundermeyer et al., 2012)

3 Methodology

Here, we describe the proposed technique for MMMT classification utilizing Long-term Recurrent Convolutional Networks.

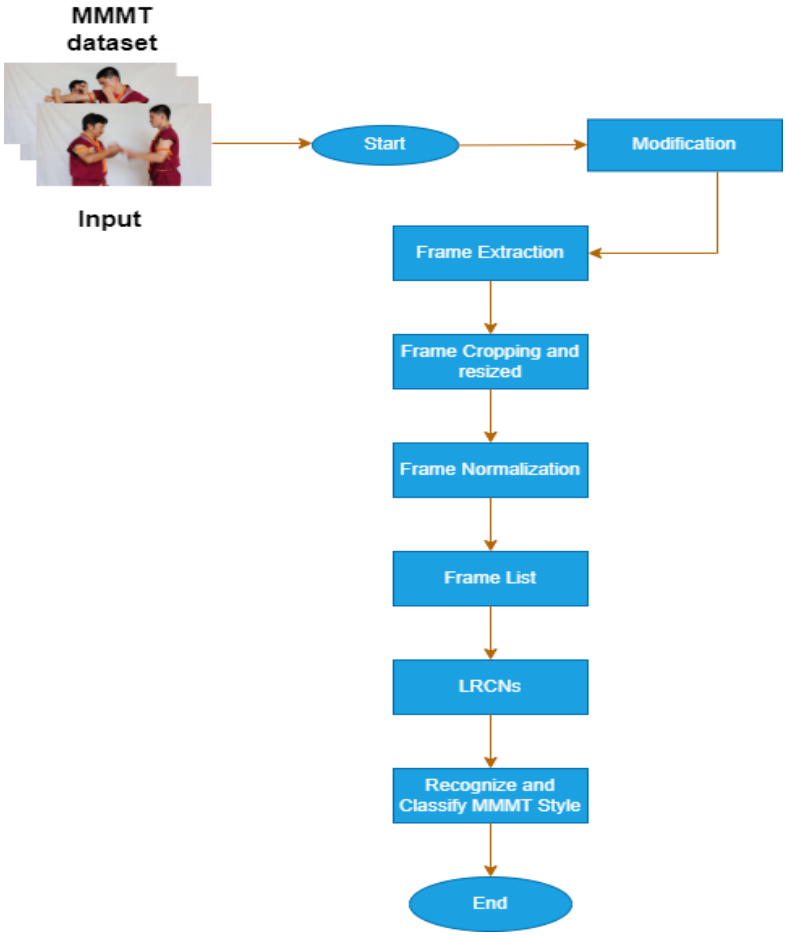

Figure 9: A Flowchart Depicting the Suggested Procedure

Figure 9 illustrates a significant achievement in the progression of our technique. In the beginning, we are going to load the Mae Mai Muay Thai video dataset, and in order to do so, we have to identify the methods that are being used in the video, as shown in Figure 9. After that, we do some preliminary processing on the MMMT RGB videos dataset and produce sequences of frames based on the videos presented here (Figure 9). After the frames have been extracted, then we resized all the extracted frames. In addition, when we have adjusted the frame size, then we also normalized the frames. The final frame list is prepared after completing the normalizing and cropping processes. There is a total of 15 MMMT techniques. Salab Fan Pla, Paksa Waek Rang, Chawa Sad Hok, Inao Thang Grit, Yo Khao Prasumeru, Ta Then Kam Fak, and Mon Yan Lak are the seven MMMT methods that will be recognized and classified. We used 75% of the data in each category for training, and we used the remaining 25% of the data for testing the network. The data were divided into a training set and a test set. Following that, we train the network using a Long-term Recurrent Convolution Networks (LRCNs) architecture, which is seen in Figure 9. Following the implementation of the LRCNs model, we are able to categorize and identify the MMMT style. After that, we will discuss the criteria that we utilized to evaluate the efficacy of our technique.

3.1. LRCNs Architecture

In the suggested study, MMMT classification in videos was accomplished using the Long-term Recurrent Convolution Networks (LRCNs) architecture. The network has four Time Distributed and Conv2D layers. And four maximum pooling layers. The LRCNs architecture is more profound than regular CNN and LSTM. A dropout of 0.25 percent is added to the fully linked layers to prevent the data from being overfitting. The following are the elements of the architecture:

- 1st Layer Time Distributed and Conv2D with 16 convolutions and 3 (width, height, channel number) dimensions shape.
- Rectified Linear Unit (RELU) Activation function.
- Time Distributed and Max Pooling2D (4*4 Kernal size)
- 2nd Layer Time Distributed and Conv2D with 32 convolutions and 3 (width, height, channel number) dimensions shape.
- Rectified Linear Unit (RELU) Activation function.
- Time Distributed and Max Pooling2D (4*4 Kernal size)
- 3rd Layer Time Distributed and Conv2D with 64 convolutions and 3 (width, height, channel number) dimensions shape.
- Rectified Linear Unit (RELU) Activation function.
- Time Distributed and Max Pooling2D (2*2 Kernal size)
- 4th Layer Time Distributed and Conv2D with 64 convolutions and 3 (width, height, channel number) dimensions shape.
- Rectified Linear Unit (RELU) Activation function.
- Time Distributed and Max Pooling2D (2*2 Kernal size)
- LSTM with 32 nodes
- SoftMax layer

3.2. Dataset

The first step of this research was to gather information on the challenges of identifying and categorizing the MMMT styles. These styles were performed by four professional MMMT boxers, and each style

was repeated sixty times by the four boxers. The difficulties with this dataset stem from the wide variation in object appearance, boxer pose, and viewpoint. The proposed MMMT dataset is more comprehensive, covering the MMMT styles in a boxing ring, and it is the largest MMMT dataset of RGB videos, which includes 420 sequences. These RGB video samples have been saved using the MP4 file format with a resolution of 1920 by 1080 and a fixed frame rate of 25 frames per second. Canon EOS RP camera is utilized to capture the dataset. The recording of the MMMT dataset takes place inside a gymnasium. Our dataset is a time series. Seventy five percent of the samples from this set are put into the learning process, while the remaining twenty five percent are utilized to check the model's accuracy.

3.3. Experimental Environment

The validity of the findings was checked by utilizing a machine with a 1.80 GHz Intel Core i5 CPU, 12 GB of RAM, and a pre-installed copy of Windows 10; this ensured that the results were accurate. Python libraries TensorFlow, sci-kit-learn for attribute classification, and Keras, a wrapper and a high-level neural network framework built on top of Tensor- Flow, are used to create the approach that is being recommended. TensorFlow is a popular framework for machine learning applications such as neural networks, and these libraries are developed on top of that framework. Jupyter notebook is an environment that may be used for the construction of programs as well as the analysis of data.

4 Experiment and Results

4.1. Evaluation Procedures

Evaluating whether or not a model is working effectively in terms of trust and validity is critical, and empirical evidence is sometimes the only method to do so. There are four possible results when making classification predictions. Table 1 gives a list of the variables used in this investigation. When evaluating a model, the four most important properties are its accuracy, precision, recall, and F1 score. The evaluation measurement variables are summarized below.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - Measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (4)$$

Table 1: The Factors that were Considered to Determine the Assessment Results

Variables	Meaning
True Positive (TP)	The observed value validated the model's prediction as being accurate.
True Negative (TN)	The matching "real" or "actual" result was "No," which was also the value that the model anticipated.
False Negative (FN)	The model predicted. No, but the actual result was Yes.
False Positive (FP)	The model predicted. Yes, but the actual result was No.

4.2. Process of Validation

Our model has been validated utilizing classification criteria, batch sizes, and epochs. In addition, we put our model to the test by boxing with dummies in different styles. When we tested our model using a fake picture, they informed us they did not have any of these styles. We evaluated our model using a total of one hundred faked images of boxing styles. They got the answers correct 98 percent of the time. The results were 98% accurate. The accuracy was poor when dealing with smaller batch sizes and epochs, but good when working with larger batch sizes and epochs. The findings of the tests are summarized in Table 2.

Table 2: Our Algorithm's Results

Epochs	Batch Size	F1-Score	Recall	Precision	Accuracy
4	8	0.58	0.65	0.57	0.63
4	16	0.66	0.70	0.65	0.72
4	32	0.77	0.78	0.73	0.80
4	64	0.83	0.85	0.83	0.85
4	128	0.86	0.87	0.87	0.87
8	8	0.89	0.90	0.90	0.90
8	16	0.90	0.91	0.92	0.91
8	32	0.92	0.92	0.92	0.92
8	64	0.93	0.94	0.93	0.93
8	128	0.94	0.95	0.94	0.94
12	8	0.94	0.94	0.94	0.94
12	16	0.95	0.95	0.95	0.95
12	32	0.94	0.95	0.94	0.95
12	64	0.95	0.95	0.95	0.95
12	128	0.95	0.95	0.95	0.95
16	8	0.94	0.95	0.95	0.95
16	16	0.95	0.95	0.95	0.95
16	32	0.96	0.96	0.96	0.96
16	64	0.97	0.97	0.97	0.97
16	128	0.97	0.97	0.97	0.97
20	8	0.98	0.98	0.98	0.98
20	16	0.98	0.98	0.98	0.98
20	32	0.97	0.98	0.98	0.97
20	64	0.98	0.98	0.98	0.98
20	128	0.98	0.98	0.98	0.98

4.3. Classification Results

The confusion matrices displaying the classification results achieved by the suggested method for the MMMT dataset are shown in Figure 10. According to the examination of the confusion matrices, the classification rates are relatively high for all of the MMMT classes. Figure 11 depicts the model training and validation accuracy. Figure 12 shows the model training and validation loss.

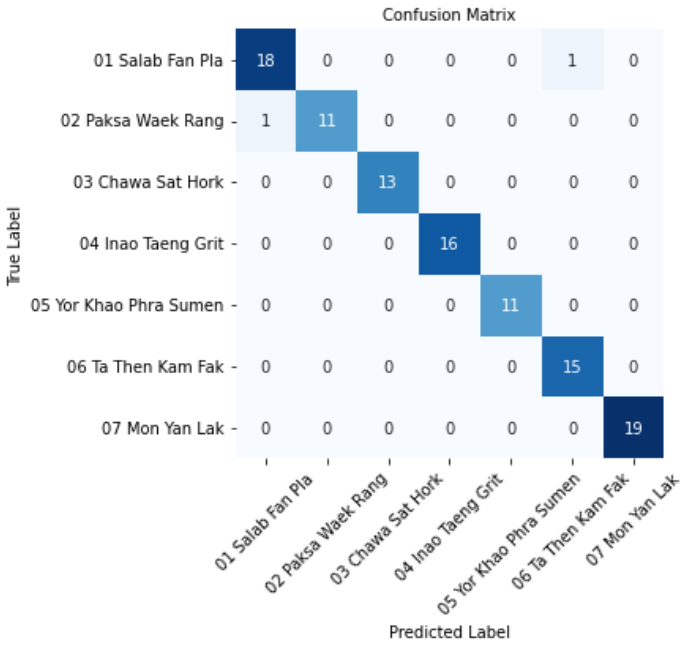

Figure 10: Confusion Matrices for MMMT Classification

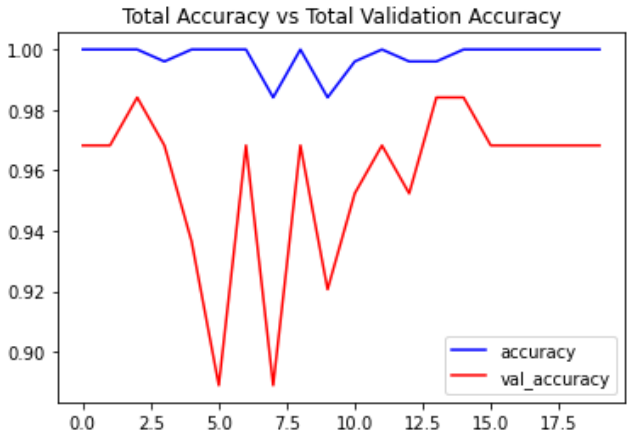

Figure 11: Graphs of Validation and Training Accuracy

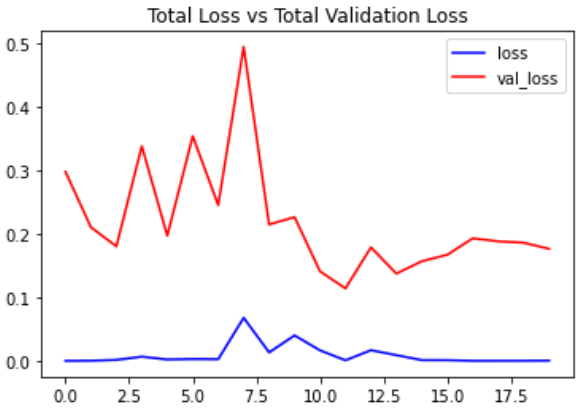

Figure 12: Graphs of Validation and Training Loss

5 Discussion

5.1. Comparison with Convolutional LSTM

We compare the long-term recurrent convolutional neural networks with Convolutional LSTM. The Convolutional LSTM results are summarized in Table 3. Additionally, we did some testing using a variety of classification methods. The accuracy values connected to the MMMT validation set have been determined for each classifier and are shown in Table 4. Therefore, we are able to identify which of the many categorization methods for MMMT action recognition is the most effective.

Table 3: Convolutional LSTM Results

Epochs	Batch Size	F1-Score	Recall	Precision	Accuracy
4	8	0.67	0.67	0.67	0.67
4	16	0.74	0.75	0.80	0.75
4	32	0.71	0.73	0.77	0.72
4	64	0.79	0.81	0.81	0.80
4	128	0.73	0.75	0.79	0.75
8	8	0.82	0.83	0.84	0.83
8	16	0.85	0.86	0.86	0.86
8	32	0.88	0.89	0.89	0.89
8	64	0.84	0.85	0.86	0.84
8	128	0.88	0.89	0.89	0.89
12	8	0.84	0.85	0.86	0.85
12	16	0.85	0.86	0.86	0.86
12	32	0.87	0.89	0.88	0.88
12	64	0.86	0.88	0.88	0.87
12	128	0.84	0.85	0.86	0.84
16	8	0.76	0.76	0.81	0.76
16	16	0.79	0.80	0.82	0.79
16	32	0.85	0.87	0.87	0.85
16	64	0.82	0.84	0.84	0.82
16	128	0.87	0.89	0.89	0.87
20	8	0.87	0.89	0.89	0.87
20	16	0.89	0.88	0.89	0.88
20	32	0.90	0.92	0.91	0.90
20	64	0.90	0.91	0.90	0.90
20	128	0.91	0.92	0.91	0.90

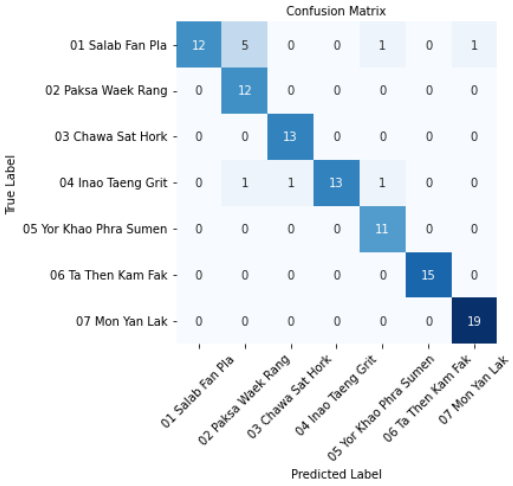

Figure 13: Confusion Matrices for MMMT Classification using Convolutional LSTM

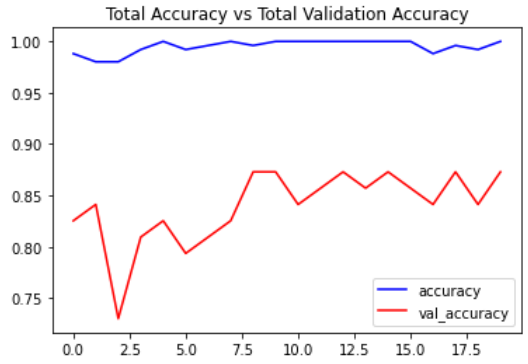

Figure 14: Graphs of Validation and Training Accuracy using Convolutional LSTM

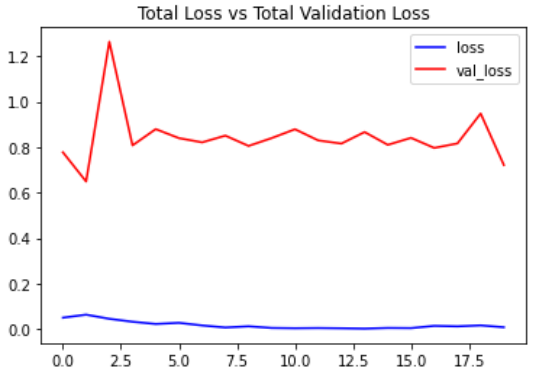

Figure 15: Graphs of Validation and Training Loss using Convolutional LSTM

Table 4: The Outcomes of Each Classification Method

Classifiers	**Accuracy**
Convolutional Neural Network (CNN)	86%
Long short-term Memory (LSTM)	89%
Convolutional LSTM	90%
Long-term Recurrent Convolution Networks (LRCNs)	98%

5.2. Comparison with Another Method Used by Other Researchers

We compared the classification outcomes achieved using our recommended strategy to those of another method to evaluate it. Compared to long-term recurrent convolutional neural networks, which have an accuracy of 98.00%, Our strategy exceeds Chantaprasert technique (2019), which only achieves an accuracy of eighty percent. In addition, they are only familiar with the most fundamental Thai boxing skills. They do this by using an angular dynamin time-wrapping strategy in conjunction with a kinetic sensor to identify the user's action (Chantaprasert et al., 2019). Through the use of overhead depth imaging, Kasiri-Bidhendi (2015) was able to categorize the straight punches that are used in boxing. During their inquiry, they used both random forest and support vector machine models. In addition, when we compare our technique to the one developed by dos S Silva (2021), we find that they determine the action by using a complementary Naive Bayes classifier. In addition, they had a mean accuracy of 62.03 and an average of 62.03 (mAP). Shen (2017) employed a hidden Markov model to determine action-based and posture-based graphs to depict boxing skills. This was done so that they could compare the two types of graphs. The boxers can only get a reaction from them when they make both essential

and necessary movements. In Table 5, the compares and enumerates several other methodologies, including the proportion of boxers, the accuracy of categorization achieved by our method, and a few others. The classification accuracy of our methodology and a few other methodologies is compared and summarized in the following table: 5. Although our method has been shown to be very accurate and precise, it does have certain limitation like the existing technique needs more time to train and identify the MMMT style.

Table 5: Comparison with Another Methods Used by other Researchers

No	Citation	Classifier	Modality	Accuracy
1	(Kasiri-Bidhendi et al., 2015)	SVM and Random Forest Classifier	Simple Boxing Images	96%
2	(Shen et al., 2017)	Hidden Markov Model (HMM)	Shadow Boxing	-
3	(Chantaprasert et al., 2019)	Angular Dynamic Time Wraping	Microsoft Kinect Sensor	80%
4	(dos S Silva et al., 2021)	Naive Bayes Classifier	Human Action from Videos	62%
5	Our Proposed Method	Long-term Recurrent Convolution Networks	MMMT from videos	98%

6 Conclusion

We propose an innovative method for visual MMMT identification based on long-term recurrent convolutional neural networks. With an accuracy rate of 98.00%, the practical results were attained using a long-term recurrent convolutional neural network. Our approach demonstrates that it is an excellent tool for locating activities immediately. Therefore, our approach is appropriate for applications that need speedier processing, particularly in situations that occur in real-time, such as videos. This methodology can be applied to a variety of different domains, as well as future studies in the sports of boxing, karate, and wrestling.

In the future, we want to provide an end-to-end framework for boxer skeleton extraction. Another intriguing research is the creation of a deep learning-based system for extracting boxer skeletons from videos. we will focus on other crucial areas of skeleton extraction. Experimenting with different MMMT aesthetics so that we can enhance performance.

References

[1] Aggarwal, J.K. & Ryoo, M.S. (2011). Human activity analysis: A review. *Acm Computing Surveys, 43*(3), 1-43.

[2] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C. & Baskurt, A. (2011). Sequential Deep Learning for Human Action Recognition. *In International Workshop on Human Behavior Understanding*, 29-39.

[3] Bush, R.R., & Mosteller, F. (1955). Stochastic models for learning. *John Wiley & Sons, Inc*.

[4] Chankuna, D. (2006). *Video Analysis for the Causes of Head and Face Injuries in Amateur Muaythai Boxers Doctoral dissertation*, Chulalongkorn University.

[5] Chantaprasert, B., Chumchuen, P. & Wangsiripitak, S. (2019). Comparison of Gesture In Thai Boxing Framework Using Angular Dynamic Time Warping. *In 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON),* 601-604.

[6] Costa, K.A.P., Papa, J.P., Lisboa, C.O., Munoz, R. & Albuquerque, V.H.C. (2019). Internet of Things: A Survey on Machine Learning-based Intrusion Detection Approaches. *Computer Networks, 151*, 147-157.

[7] Dai, C., Liu, X. & Lai, J. (2020). Human Action Recognition Using Two-stream Attention-based LSTM Networks. *Applied Soft Computing, 86*.

[8] Dai, C., Liu, X., Lai, J., Li, P. & Chao, H.C. (2019). Human Behavior Deep Recognition Architecture for Smart City Applications in the 5G Environment. *IEEE Network, 33*(5), 206-211.

[9] Dhruv, P. & Naskar, S. (2020). Image Classification using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). *A Review Machine learning and Information Processing*, 367-381.

[10] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. & Darrell, T. (2015). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625-2634.

[11] dos S Silva, F.H., Bezerra, G.M., Holanda, G.B., de Souza, J.W.M., Rego, P.A., Neto, A.V.L., de Albuquerque, V.H.C. & Rebouc¸as Filho, P.P. (2021). A Novel Feature Extractor for Human Action Recognition in Visual Question Answering. *Pattern Recognition Letters, 147*, 41-47.

[12] Fan, J., Ma, C. & Zhong, Y. (2021). A Selective Overview of Deep Learning. Statistical Science: *A Review Journal of the Institute of Mathematical Statistics, 36*(2), 264-290.

[13] Gers, F.A., Schraudolph, N.N. & Schmidhuber, J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research*, 115-143.

[14] Guo, G. & Lai, A. (2014). A Survey on Still Image Based Human Action Recognition. *Pattern Recognition, 47*(10), 3343-3361.

[15] Hochreiter, S. & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation, 9*(8), 1735-1780.

[16] Kandel, I. & Castelli, M. (2020). Transfer Learning with Convolutional Neural Networks for Diabetic Retinopathy Image Classification. *A review Applied Sciences, 10*(6), 2021.

[17] Kasiri-Bidhendi, S., Fookes, C., Morgan, S., Martin, D.T. & Sridharan, S. (2015). Combat Sports Analytics: Boxing Punch Classification Using Overhead Depth Imagery. *In IEEE International Conference on Image Processing (ICIP)*, 4545-4549.

[18] Keshavarzian, A., Sharifian, S. & Seyedin, S. (2019). Modified Deep Residual Network Architecture Deployed on Serverless Framework of IoT Platform Based on Human Activity Recognition Application. *Future Generation Computer Systems, 101*, 14-28.

[19] Kwon, H., Kim, Y., Lee, J.S. & Cho, M. (2018). First Person Action Recognition Via two-Stream Convnet with Long-term Fusion Pooling. *Pattern Recognition Letters, 112*, 161-167.

[20] Li, Y., Li, Q., Huang, Q., Xia, R. & Li, X. (2019). Spatiotemporal Interest Point Detector Exploiting Appearance and Motion-variation Information. *Journal of Electronic Imaging, 28*(3).

[21] Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. (2021). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and learning systems*, 1-21.

[22] Lipton, Z.C., Kale, D.C., Elkan, C. & Wetzel, R. (2015). Learning to Diagnose with LSTM Recurrent Neural Networks, 1-18.

[23] Lokhande, B.P. & Gharde, S.S. (2015). A Review on Large-scale Video Classification with Recurrent Neural Network (RNN). *International Journal of Computer Science and Information Technologies, Jalgaon, India.*

[24] Mae Mai, (2022). https://muaysena.com/
[25] Mae Mai Look Mai, (2022). http://www.muaythai.it/techniques/mae-mai-look-mai/
[26] Rodriguez, E., Valls, P., Otero, B., Costa, J.J., Verd´u, J., Pajuelo, M.A. & Canal, R. (2022). Transfer-Learning-Based Intrusion Detection Framework in IoT Networks. Sensors, 22(15), 1-17.
[27] Schmidhuber, J., (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks, 61*, 85-117.
[28] Nowakowski, P., Zórawski, P., Cabaj, K., & Mazurczyk, W. (2021). Detecting Network Covert Channels using Machine Learning, Data Mining and Hierarchical Organisation of Frequent Sets. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 12*(1), 20-43.
[29] Shujaat, A.Z., & Varin, C. (2022). Mae Mai Muay Thai Layered Classification Using CNN and LSTM Models. *In 26th International Computer Science and Engineering Conference (ICSEC),* 351-356.
[30] Shen, Y., Wang, H., Ho, E.S., Yang, L. & Shum, H.P. (2017). Posture-based and Action-based Graphs for Boxing Skill Visualization. *Computers and Graphics, 69*, 104-115.
[31] Si, C., Chen, W., Wang, W., Wang, L. & Tan, T. (2019). An Attention Enhanced Graph Convolutional lstm Network for Skeleton-based Action Recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1227-1236.
[32] Spolaor, N., Lee, H.D., Takaki, W.S.R., Ensina, L.A., Coy, C.S.R. & Wu, F.C. (2020). A Systematic Review on Content-based Video Retrieval. *Engineering Applications of Artificial Intelligence, 90*.
[33] Sundermeyer, M., Schl¨uter, R. & Ney, H. (2012). LSTM Neural Networks for Language Modeling. *In Thirteenth Annual Conference of the International Speech Communication Association.*
[34] Xia, L. & Aggarwal, J.K. (2013). Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition using Depth Camera. *In Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2834-2841.
[35] Yang, X., Zhang, Y., Lv, W. & Wang, D. (2021). Image Recognition of Wind Turbine Blade Damage Based on a Deep Learning Model with Transfer Learning and an Ensemble Learning Classifier. *Renewable Energy, 163*, 386-397.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Funding Statement: The author received no specific funding for this study.

Author's Information

Shujaat Ali Zaidi is currently doing master's in computer science at the Chiang Mai University, Chiang Mai, Thailand. He completed his bachelor's degree in computer science from BZU Multan, Pakistan. Later briefly working in different technology-based research roles, his interests lie at the field of computer vision, image processing, computer graphics, 2D- and 3D-motion processing, medical imaging analysis, artificial intelligence, and machine learning fields.

Assistant Professor Varin Chouvatut, Ph.D. She received her Bachelor of Engineering (B. Eng) in Computer Engineering, Master of Engineering (M. Eng) in Computer Engineering, and Philosophy of Doctoral Degree (Ph.D.) in Electrical and Computer Engineering (International Program) from the Computer Engineering Department, Faculty of Engineering, King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand. She is now an assistant professor at the Department of Computer Science, Faculty of Science, Chiang Mai University (CMU), Chiang Mai, Thailand. Her research of interest includes computer vision, image processing, computer graphics, 3D reconstruction, 2D- and 3D-motion processing, augmented reality (AR) and virtual reality (VR) system, medical imaging analysis, artificial intelligence (AI), and machine learning (ML) fields.