Deep CNN-LSTM With Self-Attention Model for Human Activity Recognition Using Wearable Sensor

Abstract:

Human activity recognition is a complex problem that aims to predict user activities based on their device interactions, which has numerous applications in people's daily lives. Two main methods are currently used to detect human activity: video image recognition and wearable sensors. Wearable activity detectors are widely used in various healthcare applications for tracking fitness activities, but the effectiveness of these methods remains unknown. Therefore, researchers are working to enhance the contribution of inertial sensors for human activity recognition (HAR).

To address this issue, this paper proposes a novel deep learning model, combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks with self-attention, for human activity recognition using wearable sensors. The proposed model is specifically designed to recognize activities such as standing/sitting, regular walking, running, and jogging, using data collected from smartphone sensors.

The proposed model is evaluated using a publicly available dataset, MHEALTH, and achieves remarkable accuracy in recognizing human activities, with an accuracy rate of 99.6%. The model extracts features from time-series sensor data using CNNs and LSTMs and enhances the predictive capabilities of the system with a self-attention mechanism.

Overall, the paper presents a novel approach to human activity recognition using wearable sensors and deep learning techniques. The proposed model accurately recognizes human activities and has potential applications in clinical settings. The paper's findings provide a solid foundation for further research to improve human activity recognition systems' accuracy and efficacy.

Introduction:

Human activity recognition (HAR) is a rapidly growing field that aims to forecast user activities based on device interactions. It has a wide range of applications in areas such as healthcare, fitness tracking, sports injury detection, senior care, rehabilitation, entertainment, and surveillance in intelligent home settings. HAR systems are devised for continuously observing human behavior - primarily in the fields of environmental compatibility, sports injury detection, senior care, rehabilitation, entertainment, and surveillance in intelligent home settings.

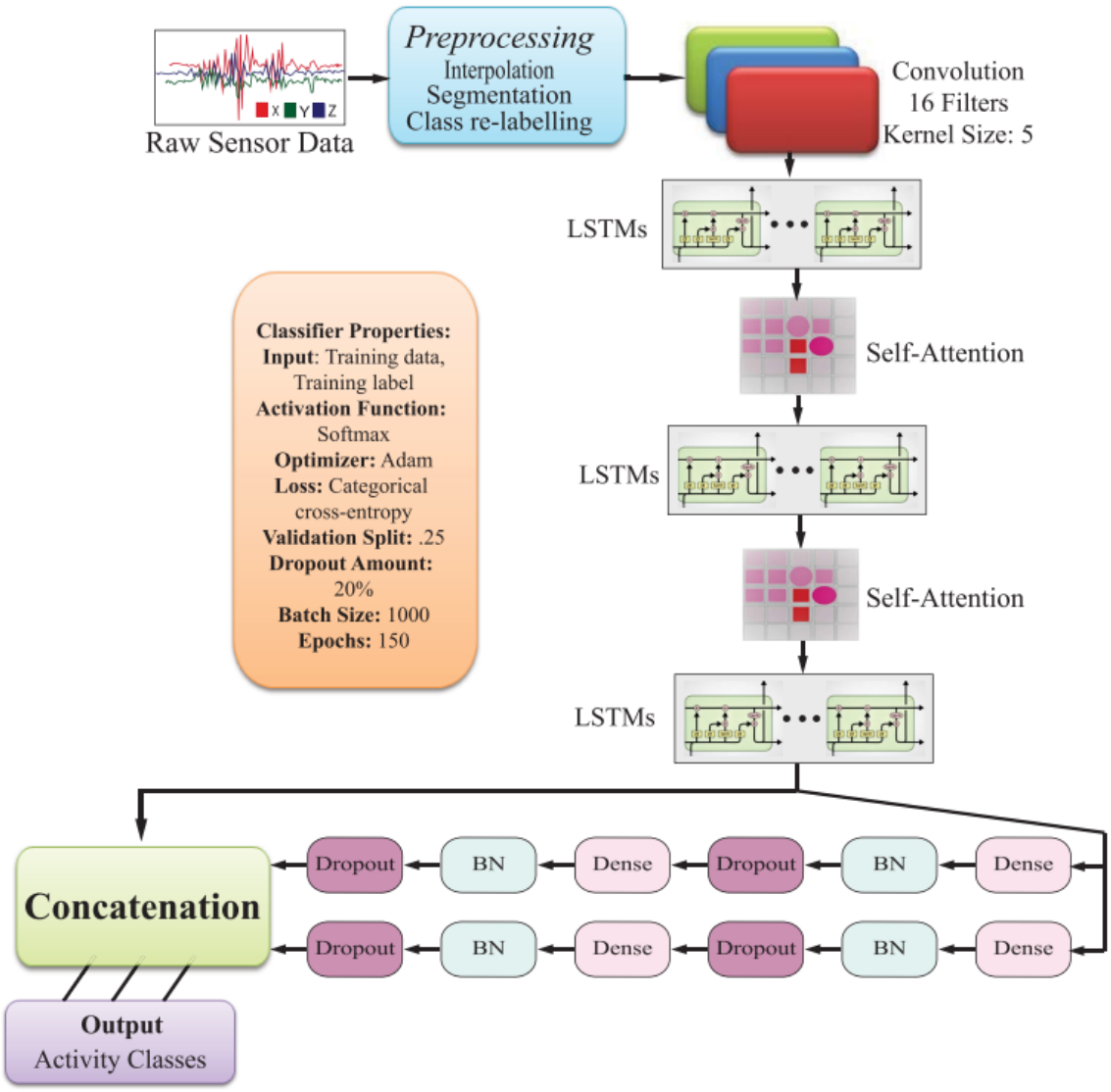
HAR can be achieved using two methods: video image recognition and wearable sensors. Video image recognition involves using cameras to recognize human behavior. This strategy not only necessitates the installation of costly cameras and infrastructure but also creates issues because of the background, lighting, and scale circumstance that make movement detection difficult. Wearable sensors, on the other hand, convert motion into identified signals and provide a new

dimension to moving with fewer environmental constraints than the video-based method while also providing privacy for the user. These sensors are now compacted into smart devices such as smartphones, making activity data acquisition for HAR systems a pressing need.

We present a novel approach to human activity recognition using wearable sensors and deep learning techniques. Our proposed model combines the strengths of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract features from time-series sensor data. CNNs are used to automatically extract spatial features from the data while LSTMs are used to capture temporal dependencies in the data. The self-attention mechanism is used to enhance the predictive capabilities of the system by allowing the model to focus on the most relevant parts of the input sequence. Finally, the dense output layer performs classification using a softmax classifier.

The use of deep learning techniques in human activity recognition (HAR) has several advantages over traditional machine learning methods. One of the key advantages is the ability of deep learning models to automatically learn features from raw data without the need for manual feature engineering. This makes them well-suited for handling complex sensor data, as they can automatically extract relevant features from the data without requiring domain expertise or prior knowledge. Traditional machine learning methods often require manual feature engineering, where domain experts must carefully design and select features that are relevant to the problem at hand. This can be a time-consuming and error-prone process, and the resulting features may not always be optimal. In contrast, deep learning models can automatically learn features from raw data by using multiple layers of non-linear transformations to extract increasingly abstract representations of the data. Another advantage of deep learning models is their ability to handle large amounts of data. Deep learning models can learn complex relationships between inputs and outputs by using multiple layers of non-linear transformations. This allows them to capture subtle patterns and dependencies in the data that may be difficult for traditional machine-learning methods to detect.

In summary, our proposed model achieves high accuracy in recognizing human activities such as standing/sitting, regular walking, running, and jogging using data collected from smartphone sensors. The MHEALTH dataset contains data collected from wearable sensors placed at the left, chest, and right handles of 10 volunteers while performing 12 physical activities. A sample rate of 50 Hz was used to record all the actions. Our proposed model is able to accurately recognize these activities using data from the MHEALTH dataset. This demonstrates the effectiveness of our approach and its potential for use in real-world applications. It also highlights its potential applications in clinical settings.



*The **proposed model architecture** used for human activity recognition*

Related or Existing Work:

Activity recognition has been used in a variety of situations [4]–[6], including individual authentication [7]–[9], medical examination [10], [11], elderly person wellness monitoring, development of wearable and smartphone-based tracking systems, and impersonation attack protection [12]. A significant amount of research has been conducted in the field of human activity recognition (HAR), with a focus on strategies that make use of smartphone sensor data. This is due to factors such as the availability, affordability, and portability of smartphone sensors, which eliminates the need for a sophisticated laboratory setup and expensive equipment.

Related work in the field of human activity recognition (HAR) using wearable sensors and deep learning techniques has focused on developing models that can accurately recognize a wide range of activities. Several studies have proposed the use of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract features from time-series sensor data.

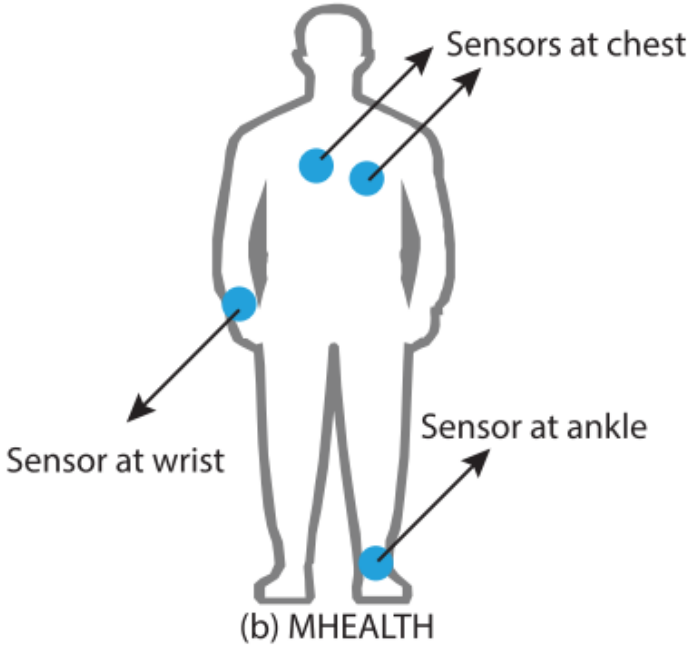
A study by Hammerla et al. [1] proposed the use of a deep CNN-LSTM architecture for HAR using data from wearable sensors. The model was evaluated using the Opportunity dataset and achieved high accuracy in recognizing activities such as walking, standing, and sitting. Another study by Zeng et al. [2] proposed a convolutional recurrent neural network (CRNN) for HAR using data from wearable sensors. The model was evaluated using the PAMAP2 dataset and achieved high accuracy in recognizing activities such as walking, running, and cycling. In addition to CNNs and LSTMs, other deep learning techniques such as autoencoders and restricted Boltzmann machines have also been used for HAR. For example, a study by Wang et al. [3] proposed the use of a stacked denoising autoencoder for HAR using data from wearable sensors. The model was evaluated using the USC-HAD dataset and achieved high accuracy in recognizing activities such as walking, running, and jumping.

Deep learning is a rapidly emerging field that automates these techniques [13]. The deep learning technique employs multiple layers in the system to identify ideal characteristics from raw data without the need for human interaction. According to several research studies, this method can produce very accurate activity classification results [14]–[16]. However, there are limitations and challenges associated with the application of deep learning techniques in HAR. For example, training a deep learning model requires a large amount of data and the model is typically treated as a black box, making algorithm improvement difficult.

In addition to HAR using smartphone sensor data, research has also been conducted on human gait analysis for various clinical and pathological trials of patients with stroke, Parkinson's disease, old-stage walking issues, and other neurological disorders. The researchers in [17] employ several machine learning techniques that necessitate the services of a feature extraction expert. They proposed utilizing cellular automata to forecast human gait state and ELM to classify it.

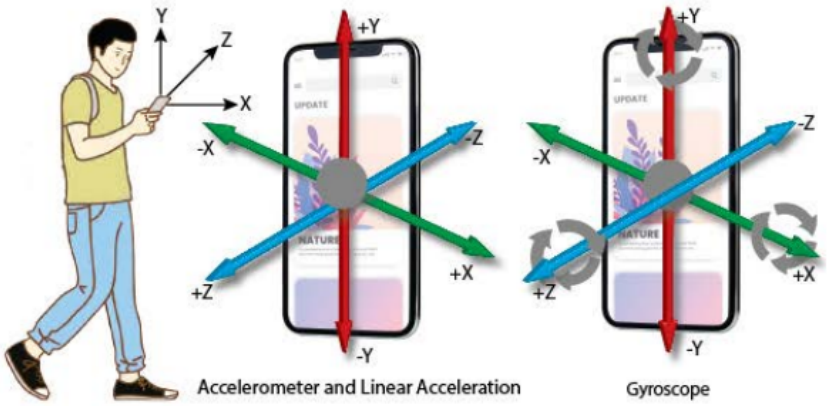
In summary, activity recognition has been used in a variety of situations and deep learning techniques have shown promise in improving activity classification and feature extraction in HAR using smartphone sensor data. Research has also been conducted on human gait analysis using machine-learning techniques and wearable sensors. Overall, these studies demonstrate the effectiveness of deep learning techniques for HAR using data from wearable sensors. Our proposed model builds on this body of work by combining CNNs and LSTMs with a self-attention mechanism to enhance the predictive capabilities of the system.

Methodology:

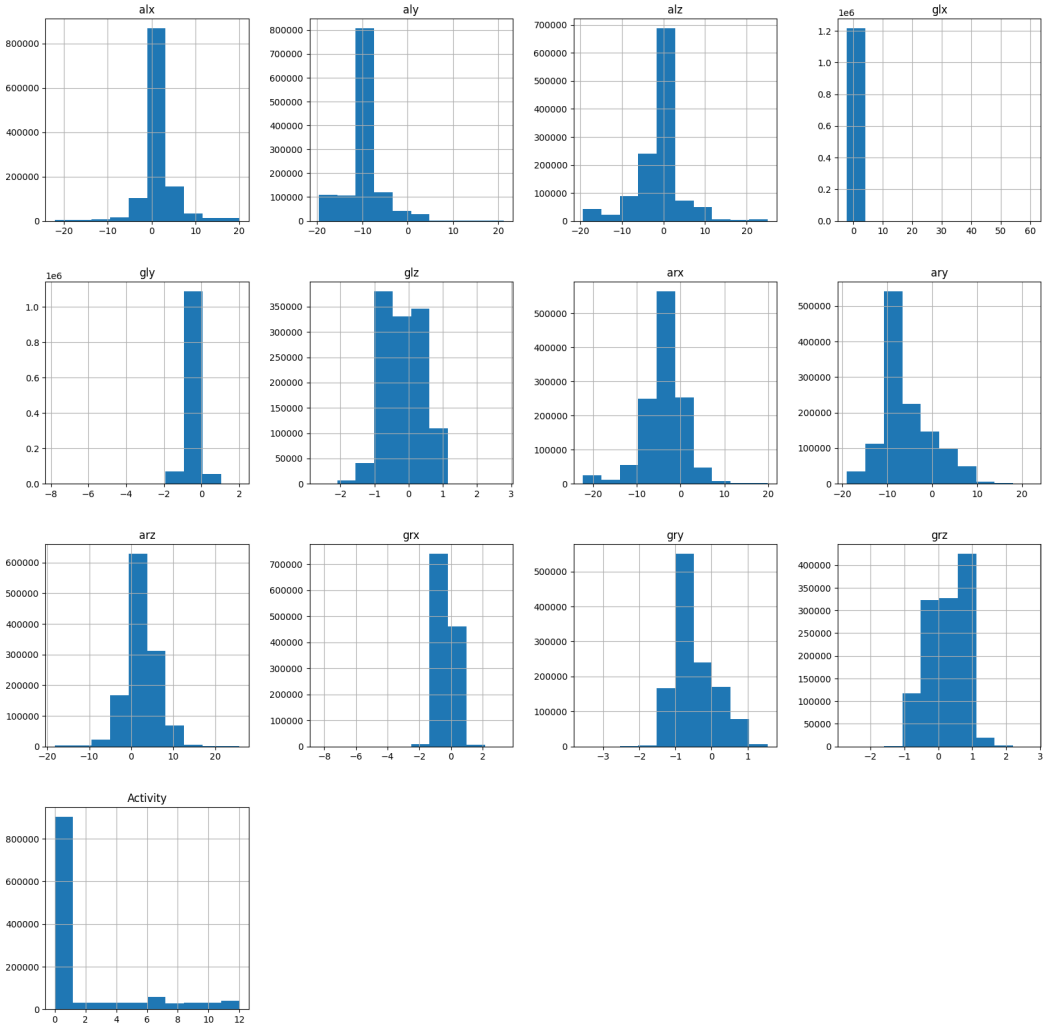


*MHealth data is collected from the accelerometer and gyroscope sensors at **left ankle and right lower arm***

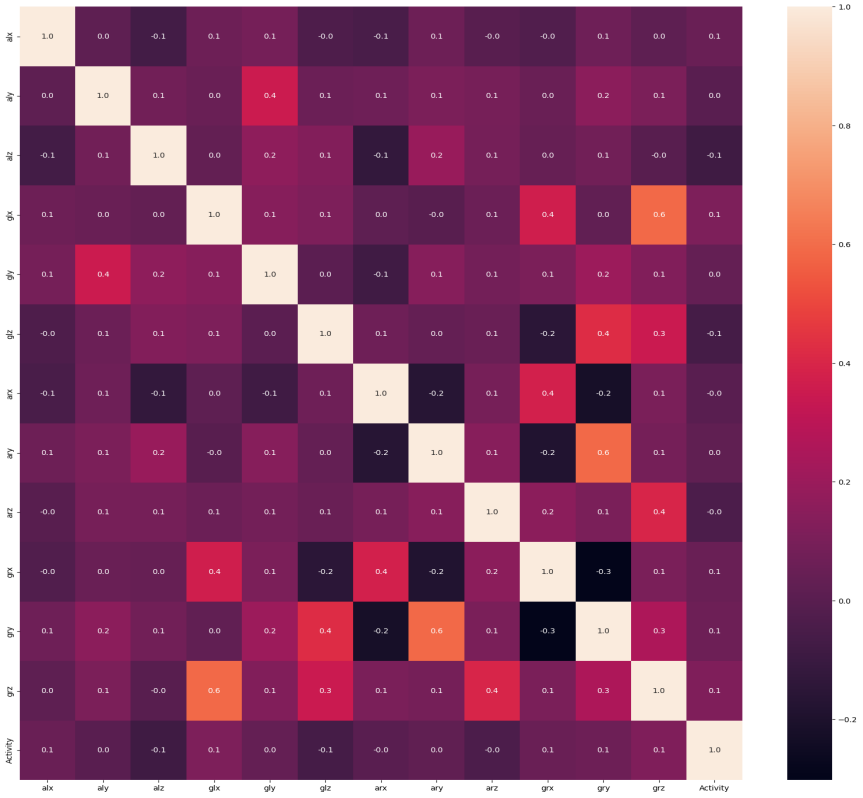
1. **Data gathering:** The MHealth dataset is obtained from Kaggle (https://www.kaggle.com/datasets/gaurav2022/mobile-health) and loaded for further analysis.
2. **Data preprocessing:** First, the data is resampled to ensure that it is evenly distributed. Next, the data is split into train and test sets. Features that fall outside the 98% confidence interval are dropped to remove outliers. Finally, the data is converted into a time series format, represented as a 3D array with shapes (n_samples, n_timesteps, n_features) to prepare for model training.
3. **Model building:** The model architecture comprises several layers including the Input layer, Conv1D layer, Batch normalization layer, Dropout layer, LSTM layer, Attention layer, Dense layer, and Output layer. The Adam optimizer is used to compile the model, and the loss function is set to sparse_categorical_crossentropy.
4. **Model evaluation:** The model is evaluated using the accuracy score and the confusion matrix to assess its performance.



Inertial sensors *in smartphones and the* ***direction*** *of the accelerometer, gyroscope, and linear acceleration*



Data Distribution

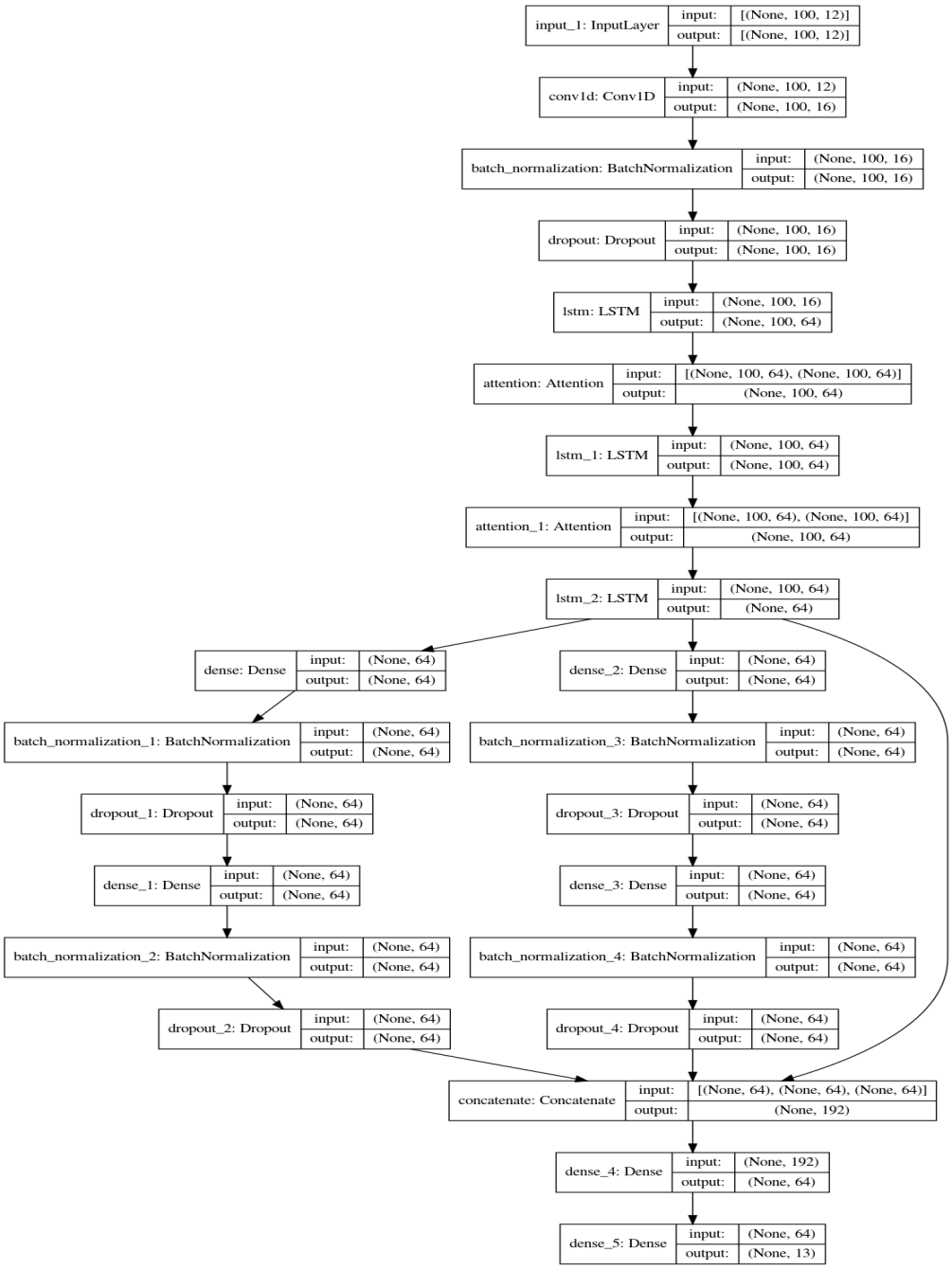


Correlation Matrix

Parameters: These needed to take care of since tuning them affected the model accuracy.

1. **Learning rate:** Affects how quickly the model learns and converges to an optimal solution. This was fixed at **0.001.**
2. **Number of epochs:** The number of times the entire dataset is passed through the model during training. This was set to **50 epochs.**
3. **Dropout rate:** A regularization technique used to prevent overfitting by randomly dropping out (ignoring) some units in the network during training. This was set to **0.2.**
4. **Activation function:** A function applied to the output of a layer to introduce non-linearity into the model. This was selected as **'ReLU'**. It is computationally efficient and can help prevent the vanishing gradient problem, which can occur when using other activation functions such as sigmoid or tanh. ReLU sets negative input values to zero and keeps positive input values as they are, making it well-suited for models with sparse inputs.
5. **Optimizer:** The algorithm used to update the weights of the model during training. The optimizer was selected to be **'Adam'**. It is computationally efficient and has been shown to work well in practice. It combines the benefits of two other optimization methods, Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp), to achieve good results with the relatively little tuning of hyperparameters.

6. **Loss function:** The function used to measure the difference between the predicted and actual outputs, which is used to update the weights of the model during training. The **sparse categorical cross-entropy** loss function is appropriate for multi-class classification problems with integer-encoded labels, as it calculates the cross-entropy loss between the true labels and predicted probability distributions. It can handle multiple classes efficiently and avoids the need for one-hot encoding of the target variables, which can be computationally expensive for large datasets.



Model Architecture

The model consists of an input layer, followed by a Conv1D layer, a batch normalization layer, and a dropout layer. Then, there are two LSTM layers with attention layers, and two branches of dense layers followed by batch normalization and dropout layers. The three branches are concatenated, and then there is another dense layer followed by batch normalization and dropout layer. Finally, there is an output layer with a softmax activation function. The model is compiled with the Adam optimizer and sparse categorical cross-entropy loss. The hyperparameters, such as the learning rate, batch size, and the number of epochs, can be adjusted to improve the model's performance.

The model combines the strengths of **convolutional neural networks (CNNs)** and **long short-term memory (LSTM) networks** to extract features from time-series sensor data. CNNs are used to automatically extract spatial features from the data while LSTMs are used to capture temporal dependencies in the data. The **self-attention mechanism** is used to enhance the predictive capabilities of the system by allowing the model to focus on the most relevant parts of the input sequence. Finally, the dense output layer performs classification using a softmax classifier.

Input: Accelerometer and gyroscope data from left ankle and right lower arm (12 values). The input is a time series data of 100 data points, hence the input shape (100 * 12)

Output: Activity class (out of 13 classes)

Moreover, experiments like **detecting activity class just from 1 pair of sensors** (accelerometer and gyroscope) from the ankle or lower arm have been performed. The accuracy of both these individual experiments is less than the combined accuracy of both the pair of sensors. The accuracy observed from arm data is 90% and from the ankle, data is 75%, whereas combined accuracy is more than 98%.

Future Work: Deploy this model on the smartphone and get the accelerometer and gyroscope from smartphone sensors. The aim is to predict the activity class on the smartphone in real time.

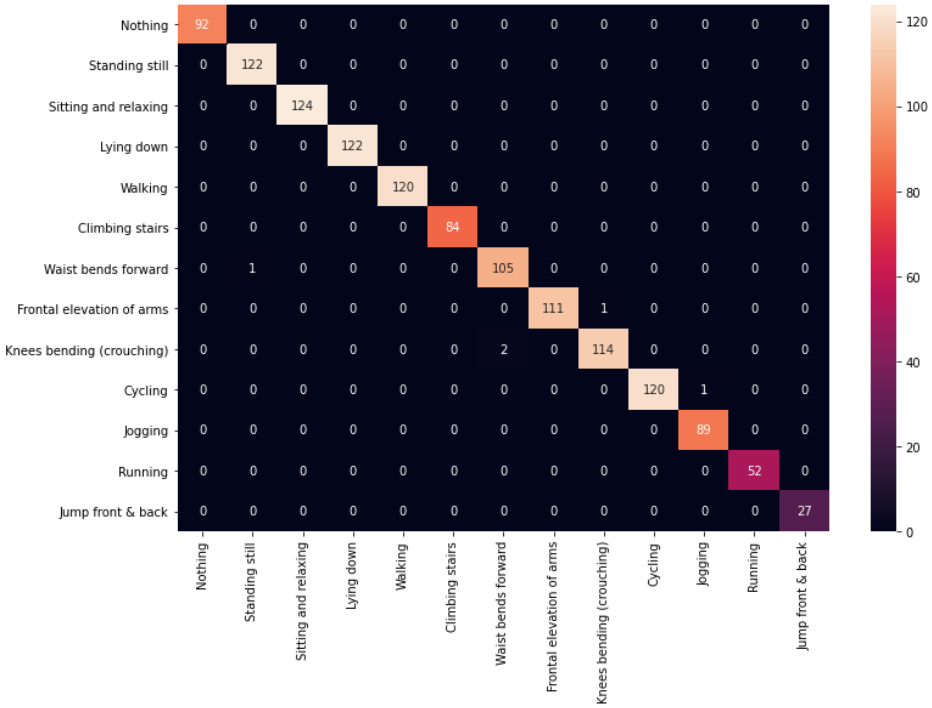
Performance Evalution:

```
model = keras.models.load_model('./mhealth_best.h5')

train_loss, train_acc = model.evaluate(X_train,y_train)
test_loss, test_acc = model.evaluate(X_test,y_test)

print("Train accuracy", round(train_acc*100, 2),'%')
print("Train loss", train_loss)
print("Test accuracy", round(test_acc*100, 2),'%')
print("Test loss", test_loss)
```

```
154/154 [==============================] - 3s 10ms/step - loss: 0.0015 - sparse_categorical_accuracy: 0.9996
41/41 [==============================] - 3s 10ms/step - loss: 0.0321 - sparse_categorical_accuracy: 0.9961
Train accuracy 99.96 %
Train loss 0.0014890251914039254
Test accuracy 99.61 %
Test loss 0.03213530778884888
```

Train: 99.96%, Test: 99.61%

Best model accuracy on various runs



Confusion Matrix of the various classes

	precision	recall	f1-score	support
0	1.00	1.00	1.00	92
1	0.99	1.00	1.00	122
2	1.00	1.00	1.00	124
3	1.00	1.00	1.00	122
4	1.00	1.00	1.00	120
5	1.00	1.00	1.00	84
6	0.98	0.99	0.99	106
7	1.00	0.99	1.00	112
8	0.99	0.98	0.99	116
9	1.00	0.99	1.00	121
10	0.99	1.00	0.99	89
11	1.00	1.00	1.00	52
12	1.00	1.00	1.00	27
accuracy			1.00	1287
macro avg	1.00	1.00	1.00	1287
weighted avg	1.00	1.00	1.00	1287

Precision, Recall, F-1 Score, and Accuracy

```
model1 = keras.models.load_model('./mhealth_ankle_best.h5')

train_loss, train_acc = model1.evaluate(X_train_ankle,y_train_ankle)
test_loss, test_acc = model1.evaluate(X_test_ankle,y_test_ankle)

print("Train accuracy", round(train_acc*100, 2),'%')
print("Train loss", train_loss)
print("Test accuracy", round(test_acc*100, 2),'%')
print("Test loss", test_loss)
```

```
155/155 [==============================] - 13s 68ms/step - loss: 0.5464 - sparse_categorical_accuracy: 0.8342
41/41 [==============================] - 3s 66ms/step - loss: 0.9062 - sparse_categorical_accuracy: 0.7537
Train accuracy 83.42 %
Train loss 0.5464469790458679
Test accuracy 75.37 %
Test loss 0.9062090516090393
```

Accuracy on Ankle Data: *75% (Using data from ankle accelerometer and gyroscope)*

```
model2 = keras.models.load_model('./mhealth_arm_best.h5')

train_loss, train_acc = model2.evaluate(X_train_arm,y_train_arm)
test_loss, test_acc = model2.evaluate(X_test_arm,y_test_arm)

print("Train accuracy", round(train_acc*100, 2),'%')
print("Train loss", train_loss)
print("Test accuracy", round(test_acc*100, 2),'%')
print("Test loss", test_loss)
```

```
155/155 [==============================] - 12s 66ms/step - loss: 0.1108 - sparse_categorical_accuracy: 0.9596
41/41 [==============================] - 3s 65ms/step - loss: 0.2369 - sparse_categorical_accuracy: 0.9060
Train accuracy 95.96 %
Train loss 0.11075565218925476
Test accuracy 90.6 %
Test loss 0.23693278431892395
```

Accuracy on Arm Data: *90% (Using data from arm accelerometer and gyroscope)*

Results and Conclusion:

The model was trained on the **MHealth dataset**, achieving near-perfect accuracy on the test set. Precision, recall, and F1 scores were calculated for each activity class, indicating good performance across all classes. Experiments were conducted to assess the accuracy of detecting activity classes using only one pair of sensors, either from the ankle or lower arm. Results showed that using sensors from the arm yielded an accuracy of 90% while using sensors from the ankle yielded an accuracy of 75%. However, combining both pairs of sensors significantly improved accuracy, highlighting the benefits of using multiple pairs of sensors in activity classification.

In conclusion, we have developed a model using accelerometer and gyroscope data from the left ankle and right lower arm to classify human activities. The model uses a combination of convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and self-attention mechanisms to extract and capture both spatial and temporal features in the data. Our model achieved high accuracy on the test set, demonstrating its effectiveness in classifying human activities. These findings indicate potential applications in healthcare, fitness tracking, and sports analysis. **Future work may focus on deploying this model on a smartphone to predict activity classes in real time.** Overall, this study demonstrates the potential of machine learning techniques for accurately classifying human activities, which has significant implications across various fields.

References:

[1] N. Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, convolutional, and recurrent models for human activity recognition using wearables,” arXiv preprint arXiv:1604.08880, 2016.

[2] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu et al., “Convolutional neural networks for human activity recognition using mobile sensors,” in 6th International Conference on Mobile Computing, Applications, and Services. IEEE, 2014.

[3] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu “Deep learning for sensor-based activity recognition: A survey,” Pattern Recognition Letters 119 (2019): 3-11.

[4] F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, and M. Savvides, ‘‘GaitID on the move: Pace independent human identification using cell phone accelerometer dynamics,’’ in Proc. IEEE 5th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS), Sep. 2012, pp. 8–15.

[5] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, ‘‘Cell phone-based biometric identification,’’ in Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS), Sep. 2010, pp. 1–7.

[6] B. Sun, Y. Wang, and J. Banda, ‘‘Gait characteristic analysis and identification based on the iPhone’s accelerometer and gyro meter,’’ Sensors, vol. 14, no. 9, pp. 17037–17054, Sep. 2014.

[7] J. Le Moing and I. Stengel, ‘‘The smartphone as a gait recognition device impact of selected parameters on gait recognition,’’ in Proc. Int. Conf. Inf. Syst. Secur. Privacy (ICISSP), 2015, pp. 322–328.

[8] H. Abujrida, E. Agu, and K. Pahlavan, ‘‘Smartphone-based gait assessment to infer Parkinson’s disease severity using crowdsourced data,’’ in Proc. IEEE Healthcare Innov. Point Care Technol. (HI-POCT), Nov. 2017, pp. 208–211.

[9] J. Juen, Q. Cheng, V. Prieto-Centurion, J. A. Krishnan, and B. Schatz, ‘‘Health monitors for chronic disease by gait analysis with mobile phones,’’ Telemedicine e-Health, vol. 20, no. 11, pp. 1035–1041, Nov. 2014.

[10] Y. Ren, Y. Chen, M. C. Chuah, and J. Yang, ‘‘User verification leveraging gait recognition for smartphone-enabled mobile healthcare systems,’’ IEEE Trans. Mobile Comput., vol. 14, no. 9, pp. 1961–1974, Sep. 2014.

[11] M. Muaaz and R. Mayrhofer, ‘‘Smartphone-based gait recognition: From authentication to imitation,’’ IEEE Trans. Mobile Comput., vol. 16, no. 11, pp. 3209–3221, Nov. 2017.

[12] F. Cruciani et al., ‘‘Feature learning for human activity recognition using convolutional neural networks: A case study for inertial measurement unit and audio data,’’ CCF Trans. Pervasive Comput. Interact., vol. 2, no. 1, pp. 18–32, Mar. 2020.

[13] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, ‘‘Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey,’’ IEEE Access, vol. 8, pp. 210816–210836, 2020.

[14] S. Ha and S. Choi, ‘‘Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2016, pp. 381–388.

[15] S. Dhanraj, S. De, and D. Dash, ‘‘Efficient smartphone-based human activity recognition using convolutional neural network,’’ in Proc. Int. Conf. Inf. Technol. (ICIT), Dec. 2019, pp. 307–312.

[16] V. Bijalwan, V. B. Semwal, G. Singh, and T. K. Mandal, ‘‘HDL-PSR: Modelling spatio-temporal features using hybrid deep learning approach for post-stroke rehabilitation,’’ Neural Process. Lett., vol. 54, no. 3, pp. 1–20, Jan. 2022.

[17] Y. Liu et al., ‘‘An attention-based category-aware GRU model for the next POI recommendation,’’ Int. J. Intell. Syst., vol. 36, no. 7, pp. 3174–3189, Jul. 2021.