IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 4, 2020, accepted September 23, 2020, date of publication September 30, 2020, date of current version October 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027979

A Deep Learning Approach for Human Activities Recognition From Multimodal Sensing Devices

ISIBOR KENNEDY IHIANLE[1], (Member, IEEE), AUGUSTINE O. NWAJANA[2], (Senior Member, IEEE), SOLOMON HENRY EBENUWA[3], RICHARD I. OTUKA[3], (Member, IEEE), KAYODE OWA[1], (Member, IEEE), AND MOBOLAJI O. ORISATOKI[4]

[1]Department of Computer Science, Nottingham Trent University, Nottingham NG11 8NS, U.K.
[2]Faculty of Engineering and Science, University of Greenwich, London SE10 9JR, U.K.
[3]School of Architecture, Computing and Engineering (ACE), University of East London, London E16 2RD, U.K.
[4]Department of Engineering and Design, University of Sussex, Brighton BN1 9RH, U.K.

Corresponding author: Isibor Kennedy Ihianle (isibor.ihianle@ntu.ac.uk)

ABSTRACT Research in the recognition of human activities of daily living has significantly improved using deep learning techniques. Traditional human activity recognition techniques often use handcrafted features from heuristic processes from single sensing modality. The development of deep learning techniques has addressed most of these problems by the automatic feature extraction from multimodal sensing devices to recognise activities accurately. In this paper, we propose a deep learning multi-channel architecture using a combination of convolutional neural network (CNN) and Bidirectional long short-term memory (BLSTM). The advantage of this model is that the CNN layers perform direct mapping and abstract representation of raw sensor inputs for feature extraction at different resolutions. The BLSTM layer takes full advantage of the forward and backward sequences to improve the extracted features for activity recognition significantly. We evaluate the proposed model on two publicly available datasets. The experimental results show that the proposed model performed considerably better than our baseline models and other models using the same datasets. It also demonstrates the suitability of the proposed model on multimodal sensing devices for enhanced human activity recognition.

INDEX TERMS Human activity recognition, deep learning, machine learning, wearable sensors, convolutional neural network, long short-term memory.

I. INTRODUCTION

Human activity recognition (HAR) is a process aimed at recognising what an individual is doing, for example, sleeping, showering, and cooking and the context in which they occur. Research into HAR has been growing recently due to the need to provide support to the elderly and cognitively impaired. These efforts so far have focused on the use of video [1], wearable sensors and wireless sensor networks [2], [3] to capture simple human activities.

Sensor-based HAR approaches recognise activities from sensor signals generated from object use as the result of the interactions of sensor tagged objects in the home environment or from sensor signals generated from wearable sensors and Inertial Measuring Units (IMU) due to body movements. Taking these into consideration, Ihianle *et al.* [4] and Patterson *et al.* [5] relied on machine learning techniques to recognise activities. Irrespective of the sensor modality, these approaches follow a template of handcrafted features which makes them time-consuming and almost impossible to scale up HAR for multiple activity set. These approaches also face the challenges and the limitations to classify discriminative features which are associated with multiple and complex activities due to the variability of body movements which generally could be easily confused. Deep learning presents a promising opportunity to address these challenges and limitations.

Deep learning was applied successfully for image recognition [6], [7], speech recognition [8], [9] and human activity recognition amongst many [10]. Deep learning performs feature extraction and activity classification by the use of non-linear information processing layers. These layers form a

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

hierarchical network of dense layers which work in sequence to extract features and classify activities from raw sensor data automatically. HAR approaches based on deep learning outperformed traditional machine learning techniques achieving high prediction and classification accuracies [10].

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are popular types of deep learning models applicable to HAR. A typical CNN based approach uses convolutional operators in a stack of hierarchical layers for feature extraction. LSTM forms a network of recurrent layers which can automatically learn input features from long-term dependencies between time steps of sequence data. The Bidirectional LSTM (BLSTM) extends LSTM to use their internal states and take full advantage of their forward and backward sequences. Given their characteristic abilities, CNN and BLSTM can be combined to form a feature extraction model to improve HAR.

This paper proposes a deep learning multi-channel CNN Bidirectional LSTM (MCBLSTM) architecture for activity recognition. The multi-channel model comprises of three channels of CNN and BLSTM layers concatenated within the model. Each of the channel or head reads the same sensor data inputs and then processes these for feature extraction. Features extracted are concatenated and then interpreted by another fully connected layer for prediction. The main advantage of the proposed model is that the CNN layers can achieve activity recognition of multiple and complex activities by the direct mapping and abstract representation of raw sensor inputs for feature extraction at different resolutions of the channels. It is then supported by the recurrent BLTSM layer to enhance prediction by fully utilising the forward and backward sequence operations as an advantage over the unidirectional LSTM. The motivation for the proposed model is to improve the performance of existing HAR models and to scale up feature extraction in the recognition of multiple and complex activities due to considerable variability in movements.

The WISDM dataset [11], [12] is a unique example of multiple and complex activities, for example, brushing teeth, eating chip, eating sandwich amongst a set of 18 activities. These activities were the result of the large variability in movements beyond the normal set of activities which primarily are simple exercises. Deep learning models were applied to this dataset and other similar datasets with multiple activities by Benavidez and McCreight [13] and Burns and Whyne [14] achieving good results. However, we believe the results can be improved using the proposed deep learning approach.

The main contributions of this paper are as summarized below:

- We propose a new deep learning architecture for multiple human activity recognition using combined CNN and recurrent Bidirectional LSTM networks capable of feature extraction and the utilization of temporal dependencies.
- We demonstrate that the model can be effectively applied to different sensor modalities to recognise multiple human activities irrespective of the variability of the body movements.
- We evaluate the proposed model using publicly available datasets and show that it outperforms other deep learning models based on the published results and our baseline deep learning models.

II. RELATED WORK

Human activity recognition is an important area of research in pervasive computing due to its significance in the provision of support and assistance to the elderly, disabled and cognitively impaired. In this section, we review related works in HAR and introduce the convolutional neural network (CNN) and Bidirectional LSTM (BLSTM) as baseline models.

Machine learning techniques have been widely applied to recognise activities from multimodal sensing devices some of which includes Dynamic Bayes Networks [15], Hidden Markov Model [5], Naive Bayes [16], Topic model Latent Dirichlet Allocation [17]. Human activity recognition by knowledge-driven ontology models follow web ontology language for the specification of concepts and their relationships. It also involves the use of the ontology language to model activities and the representation of activity concepts to support activity recognition [18]–[20].

The authors [10], [13], [14] applied Deep learning models to recognise human activities of daily living. These models perform feature extractions using non-linear information processing layers and has been extensively used in areas including image pattern recognition and speech recognition. Deep learning networks, just like other neural networks work by imitating the brain neural layers and nodes. These networks of layers take in data at the input layers, which are processed and passed onto the next layers for feature extraction and then classification. The network of layers and their parameters may differ and determine the extent of feature extraction. For the work in this paper, we review Convolutional Neural Networks, recurrent Long Short term Memory and Bidirectional Long Short term Memory networks and their applications specifically to HAR.

A. CONVOLUTIONAL NEURAL NETWORK (CNN)

Deep learning CNN is composed of convolutional layers for feature extraction from input data. Typically, the network is comprised of an input layer, hidden layer(s) and an output layer. The structure of neurons which mimic the human neural network are three-dimensional to have width, height, and depths for the inputs and outputs. Depending on the arrangement, the hidden layers could be convolution, pooling and normalization layers.

The CNN layers perform feature extraction of input data through a convolutional process using filters (kernels). The layers read the distinctive values from input data to generate feature activation maps which highlights relevant features of the data. The feature activation maps are units with parameters which, when activated result to the convolution of the kernel over the data. For example, if the input layer with feature

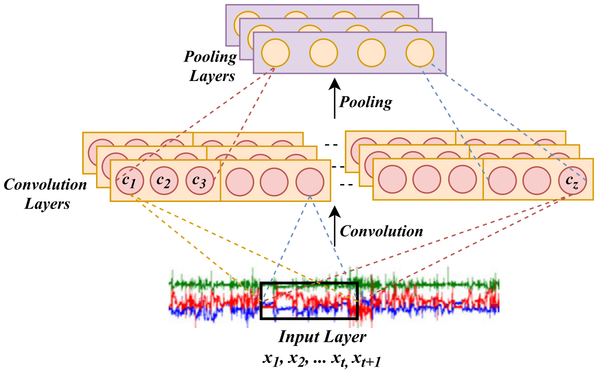


FIGURE 1. An illustration of a one-dimensional convolutional neural network model with convolutional and pooling layers.

signals $X(x_1 \dots x_t, x_{t+1})$ is connected to the convolutional layer with filter (kernel) size K as shown in Figure 1, feature map extraction using one-dimensional operation is given by C_z in the equations 1 to 4 below.

$$C_1 = k_1x_1 + k_2x_2 + k_3x_3 \tag{1}$$

$$C_2 = k_1x_2 + k_2x_3 + k_3x_4 \tag{2}$$

$$C_z = k_1x_{t-1} + k_2x_t + k_3x_{t+1} \tag{3}$$

$$C_z = \sigma\left(\sum_{t=1}^{T} X_t * W_{t,z}\right) \tag{4}$$

where $W_{t,z}$ $(X \times K)$ connects the t^{th} input signal to the z^{th} feature signal, σ is the activation function, $*$ convolutional operator. The pooling layer connected to the convolutional layer, as shown in Figure 1 enhances the extracted feature signals by reducing its dimension using a pooling function. Multiple convolutional layers help feature extraction of input data to greater levels of abstraction.

The dimension of the input data determines the convolutional operator to adopt. 2D convolutional layers are typically used for the temporal sequence of images as a result of the three-dimensional data input. The convolutional process passes a filter (kernel) over the image, continuously inspecting small windows of the image until completely scanned. Feature map activations in the form of the dot product of the pixel values in the current filter window with the weights are defined in the filter. Unlike the 2D CNN, the convolutional operators in 1D CNN are applicable to extract features from fixed-length segments of the entire input data. It is well suited for the recognition of human activities from time-based sequences of multimodal sensor data and audio signals.

With the ability to overcome the limitations of handcrafted features and to automatically learn features, CNN has gained a lot of attention over the years. It has found its application in areas including image classification and recognition [6], [7], text analysis and classification [21], natural language processing [22] and speech recognition [8], [9]. CNN has been applied to HAR Ignatov [23] and Yang *et al.* [24]. Ignatov [23] proposed CNN for local feature extraction together with simple statistical features that preserve information about the global form of time series on two publicly available datasets. Although they achieved good results for walk upstairs, walk downstairs, sit, stand, and laying activity recognition, the result did not reach 100%. Benavidez and McCreight [13] also applied CNN to the WISDM dataset with multiple activities. Their work involved feature extraction for 18 different activities which included hand-oriented and non-hand-oriented activities from phone and watch sensors. The authors achieved significant results from a standalone CNN model. Despite these encouraging results, there is still room to improve recognition accuracy, especially for multiple and complex set of activities.

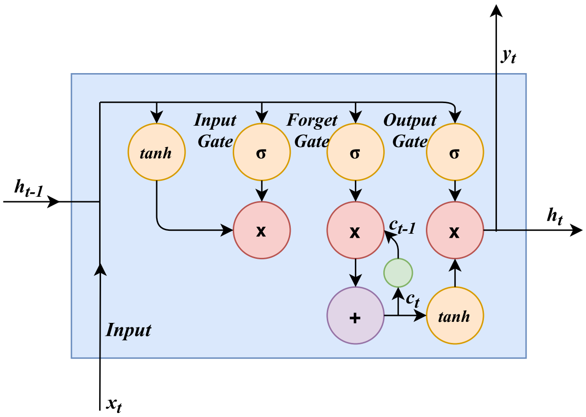


FIGURE 2. An LSTM cell structure showing the Input, Forget and Output gates.

B. LONG SHORT TERM MEMORY (LSTM)

Deep learning LSTM model input features and their temporal dependencies with memory blocks as unique features for internal and outer recurrence [25]. Typically, LSTM layers comprise of memory blocks recurrently connected in a memory unit or cell. These cells are composed of gates to determine when to forget previous hidden states of the memory cell and further update the cells, thereby enabling the network to utilise temporal information.

An LSTM cell as illustrated in Figure 2 with input feature x_t, takes input data x, at time t, so that an input gate i_t controls the flow of the input data to the cell. A forget gate f_t determines when to forget contents of the internal state of the cell, and the output gate o_t to control flow to the output. The cell function in this regard are as follows:

$$i_t = \sigma\left(U_ix_t + W_ih_{t-1} + b_i\right) \tag{5}$$

$$f_t = \sigma\left(U_fx_t + W_fh_{t-1} + b_f\right) \tag{6}$$

$$o_t = \sigma\left(U_ox_t + W_oh_{t-1} + b_o\right) \tag{7}$$

$$g_t = \sigma\left(U_gx_t + W_gh_{t-1} + b_g\right) \tag{8}$$

$$c_t = g_ti_t + f_tc_{t-1} \tag{9}$$

$$h_t = o_t tanh(c_t) \tag{10}$$

The internal recurrence c_t and the current output y_t which is the equal to the current hidden state h_t are both computed in time t using gate parameters U and W (weight matrices) and with b (bias vector) learnt in the process.

LSTMs has been applied to video footage by Zhang *et al.* [26], speech recognition Soltau *et al.* [27] and text analysis Nowak *et al.* [28]. Benavidez and McCreight [13] also applied LSTM to HAR achieving some significantly encouraging results.

C. BIDIRECTIONAL LONG SHORT TERM MEMORY (BI-LSTM)

Bidirectional LSTM proposed by Schuster and Paliwal [29], is an extension to the traditional LSTM. The model includes two parallel LSTM layers to provide a forward and backward loop, as illustrated in Figure 3. The idea is for the network to take advantage of past and future information through the forward and backward sequences to make predictions. In this case, current information has past information as dependencies and also linked to future information. The forward $\overrightarrow{h}$ and backward $\overleftarrow{h}$ sequences respectively are represented by the red and green arrows in Figure 3 and the equations below:

$$\overrightarrow{h}_t = g\left(U_{\overrightarrow{h}} x_t + W_{\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \tag{11}$$

$$\overleftarrow{h}_t = g\left(U_{\overleftarrow{h}} x_t + W_{\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \tag{12}$$

$$y_t = g\left(V_{\overrightarrow{h}} \overrightarrow{h}_t + V_{\overleftarrow{h}} \overleftarrow{h}_t + b_y\right) \tag{13}$$

Wollmer *et al.* [30] applied Bidirectional LSTM network to predict missing words based on contexts, Graves and Schmidhuber [31] to perform phoneme classifications and Graves *et al.* [32] for speech recognition. Bidirectional LSTM has also been used in HAR by Murad and Pyun [33] and Zhao *et al.* [34]. Given the forward and backward sequence advantage, Bidirectional LSTM offered better performance than the traditional LSTM [34].

In addition to the above, research efforts have been made to use a combination of deep learning models. Amongst these are combined Long Short-Term Memory (LSTM) RNN with CNN [35] and the Deep ConvLSTM [10]. Ordóñez and Roggen [10] proposed a combination of deep convolutional and recurrent layers to HAR. Results on publicly available datasets demonstrate the potentials of using convolutional layers with recurrent networks to learn temporal features of multimodal sensor signals.

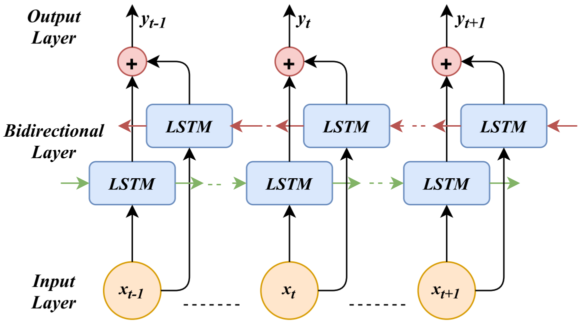


FIGURE 3. Bidirectional LSTM model showing the input and output layers. The red arrows represent the backward sequence track and green the forward sequence track.

This paper proposes a deep learning multi-channel network which significantly differs from the series combinations by Ordóñez and Roggen [10]. Although Yenter and Verma [36] and Hyun *et al.* [37] have both used deep learning multi-channel topologies, it has not been applied to HAR, especially with multiple activity set.

III. PROPOSED APPROACH

The architecture of the proposed multi-channel deep learning model combines CNN for feature extraction and Bidirectional LSTM layers for sequence prediction in three channels which are then fully connected as shown in Figure 4. Each channel or head comprises of three stacked Conv1D layers, each of which followed by a max-pooling layer. The Conv1D layers with a configuration of 64 filters perform direct mapping and abstract representation of sensor inputs for feature extraction. The feature extraction is achieved by the convolution operators applied to the kernels and then the feature maps computed as described in subsection II-A above. To allow for feature extraction at different resolutions, we have used kernel sizes of 3, 5 and 11 for the Conv1D layers of the respective channels. Outputs from each of the Conv1D layers are passed to max-pooling layers, which reduces the sizes of the learned features by summarising them into distinct elements without any loss to accuracy. A Bidirectional LSTM layer with configurations of 128 units and dropout function of 0.25% further receives the outputs of the Conv1D layers and max-pooling layers. The benefit of this is that the Bidirectional LSTM layer is well suited to adapt the internal state taking full advantage of their forward and backward sequences [29], [34]. Typically, the inputs of the Bidirectional LSTM are time sequences and the extracted features from the convolutional process. The dropout layer serves to minimise overfitting and improve model accuracy. The outputs from the three channels are flattened and then concatenated within the model. It further passes through a fully connected layer, capable of generating features interpreted into different classes. The final output of the model is from a dense layer with soft-max activation function that computes the probability distribution over the predicted classes of activities.

IV. EXPERIMENTS

We considered two datasets on HAR to evaluate the performance of the proposed MCBLSTM model. We also compared the performance of the proposed model with CNN and Bidirectional LSTM models, which provided baseline references and with the results reported by other authors on these datasets.

A. DATASETS

Human activities are a result of the large variability in movements resulting in activities like walking, jumping and sometimes complex activities like eating chips and

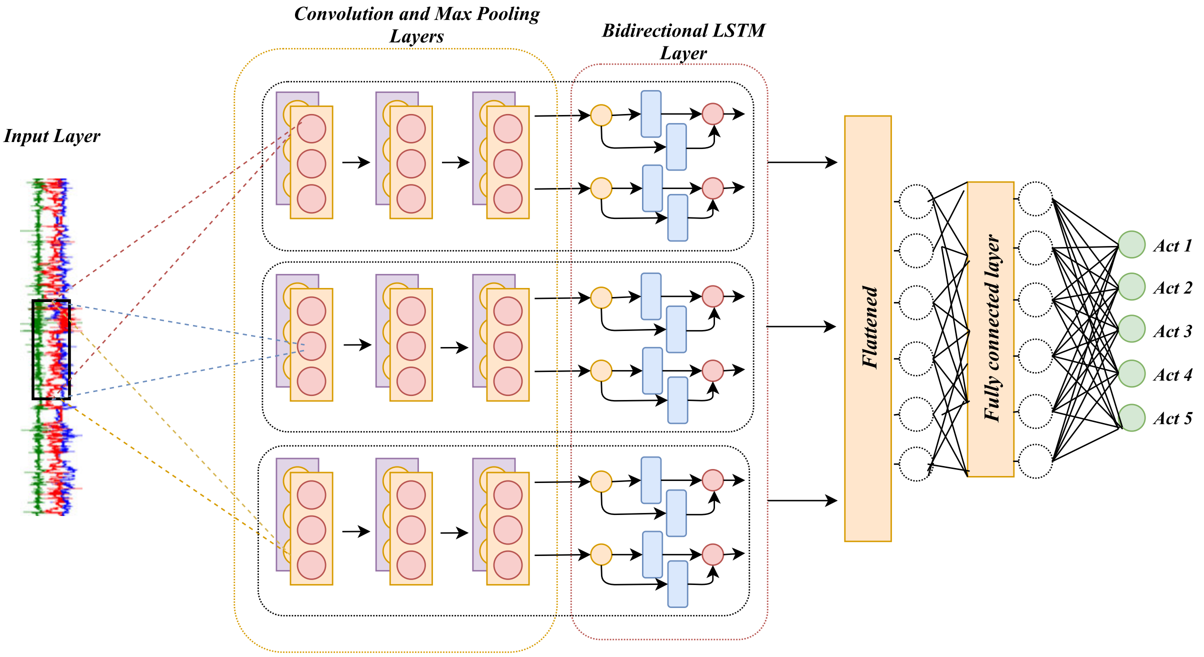


FIGURE 4. An overview of the proposed Multichannel Convolutional Bidirectional LSTM (MCBLSTM). As illustrated, there are three stacked Conv1D layers each followed by a max pooling layer. The CNN and the pooling layers are then followed by a Bidirectional LSTM layer for each channel head before flattening. The channels are then concatenated and fully connected for activity recognition and predictions.

eating sandwich. The recognition of these complex activities tends to be very challenging due to the numerous gestures associated with these activities. They also reflect the human activities of daily living performed as a result of the diverse movements and gestures. Therefore, a benchmark dataset for the recognition of activities should have multiple activities as a reflection of this reality. Several datasets have been published for activity recognition which includes WISDM [11], MHEALTH [38], UCI HAR [39] and OPPORTUNITY [40]. For this paper, we have chosen the WISDM [11] and the MHEALTH datasets [38] due to the complexity of the activities.

1) WISDM DATASET

The WISDM dataset [11], [12] was released late 2019 by the Fordham University. It was collected using accelerometer and gyroscope sensors of a smartphone (Galaxy Nexus or Samsung Galaxy S5) and smartwatch as 51 subjects performed 18 diverse activities of daily living. This dataset is unique and comprises of complex and multiple activities from basic ambulation to diverse movements and gestures. The activities as itemised below are divided into three groups – Non-Hand-oriented, Hand-oriented (General) and Hand-oriented (Eating) activities.

Non-hand-oriented activities:

- Walking
- Jogging
- Stairs
- Standing
- Kicking

Hand-oriented activities (General):

- Dribbling
- Playing catch
- Typing
- Writing
- Clapping
- Brushing teeth
- Folding clothes

Hand-oriented activities (Eating):

- Eating pasta
- Eating soup
- Eating sandwich
- Eating chips
- Drinking

The compressed folder of the WISDM dataset has a hierarchical file structure of raw and processed sensor data. For this study, we have used the raw sensor data. The raw sensor data is a time-series data stored in separate files containing data from one sensor (accelerometer or gyroscope), for one device (smartphone or smartwatch) and the subjects. Each of the subjects, as a result, is associated to four files of tri-axial sensor data sampled at a rate of 20Hz - Phone accelerometer, Phone gyroscope, Watch accelerometer and Watch gyroscope all of which are linked by the timestamp of capture. As part of the data preprocessing, we merged all Phone and Watch data so we can identify the sensor combination which yields the best results, i.e. Phone, Watch and Phone + Watch. Phone and Watch data each has seven features (accelerometer: x, y and z, gyroscope: x, y and z and activity labels respectively), Phone + Watch has thirteen features from a combination of Phone and Watch data. Further, we partitioned the merged data into 10-seconds with an overlap of 50% and then randomly split the dataset into training (70%) and test (30%).

2) MHEALTH DATASET

The MHEALTH dataset [38] is comprised of activities captured using wearable sensors in an out-of-lab environment with no constraints. In the study, 10 participants performed 12 activities (See Table 1)

TABLE 1. The activity for the MHEALTH dataset In brackets are the number of repetitions (Nx) or the duration of the exercises (min).

Activity Keys	Activities
L1	Standing still (1 min)
L2	Sitting and relaxing (1 min)
L3	Lying down (1 min)
L4	Walking (1 min)
L5	Climbing stairs (1 min)
L6	Waist bends forward (20x)
L7	Frontal elevation of arms (20x)
L8	Knees bending (crouching) (20x)
L9	Cycling (1 min)
L10	Jogging (1 min)
L11	Running (1 min)
L12	Jump front and back (20x)

Similar to the WISDM dataset, the MHEALTH dataset generalizes common activities of daily living, given the diversity of body parts involved in each one (e.g., the frontal elevation of arms vs. knees bending), the intensity of the actions (e.g., cycling vs. sitting and relaxing) and their execution speed or dynamicity (e.g., running vs. standing still) which makes this dataset very unique. As part of the process, sensors were attached to the right wrists, left ankles and chests of the participants. These sensors captured measurements from body movements, acceleration, the orientation of the magnetic field and ECG. The sampling rate of the dataset is 50 Hz, which was considered sufficient for capturing human activity. Similarly, for data preprocessing, we have merged all the data from each of the participants and randomly split into training (70%) and test (30%).

B. BASELINE MODELS

We use two baseline models - CNN and Bidirectional LSTM models to compare the performance of the proposed MCBLSTM model. These baseline models are similar to the layers as described in the section III above. With this, the baseline CNN model is comprised of three layers of Conv1D layers of 64 filters, kernel size 5 and each followed by max-pooling layers. The Bidirectional LSTM baseline model is comprised of a single Bi-LSTM layer with 128 units each followed by a dropout layer. The baseline models receive the same input as the proposed MCBLSTM model.

C. PERFORMANCE EVALUATION

We implemented the proposed model in a python 3.7 environment. To evaluate the performance of the proposed model, we used accuracy, F-measure (F_1) and Matthew Correlation Coefficient (MCC). Generally, accuracy is popular and often used as a measure to evaluate the performance of activity recognition models. Due to class imbalance which is normally evident with activity recognition datasets, accuracy tend not to be an appropriate measure of performance evaluation [41], [42]. While accuracy measures the ratio of correctly recognised activities from the total activities in the test dataset, the F-measure (F_1) is also popular and considered ideal for performance evaluation measurement. F_1 is calculated using the average weighted recall and precision as defined by equation 14. In this case, the precision is defined as $\frac{TP}{TP+FP}$ and recall defined as $\frac{TP}{TP+FN}$ with TP, TP, FP and FN corresponding to the true positive, true negative, false positive and false negative respectively. The Matthews Correlation Coefficient (MCC) defined by equation 15 is considered as a much more reliable performance evaluation measure over accuracy, (F_1) and even Cohen Kappa especially for imbalanced multi-class dataset [43], [44].

$$F_1 = \frac{2 \cdot (Precision \cdot Recall)}{(Precision + Recall)} \tag{14}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{15}$$

D. RESULTS

We implemented the proposed MCBLSTM model setting the CNN filter maps to 64 for each of the layers, CNN kernel sizes of 3, 5 and 11 for each of the channels and BLSTM layers to 128 units each followed by dropout set to 0.25. We also used an Adam optimiser for the model and set the learning rate to 1e-3. We conducted a set of experiments for the different sensor modalities Phone (accelerometer + gyroscope), Watch (accelerometer + gyroscope) and Phone + Watch for the WISDM dataset to determine the combination of sensor modality with the best result. We also performed additional experiments on the MHEALTH dataset as part of the model evaluation. We used 10-fold cross-validation for the experiments performed on the datasets.

Figures 5 and 6 show the training progress of the model on the WISDM (Phone + Watch) and MHEALTH datasets, respectively. As shown, the accuracy and loss were out of 1.0. We monitored the accuracy and loss trend for up to 100 training epochs. In this process, we noticed the stability of the model after 15 epochs and no overfitting for both datasets. The experimental results after 100 epochs for the WISDM (Phone), WISDM (Watch) and WISDM (Phone + Watch) are presented in Table 2. We noticed the WISDM (Phone + Watch) outperformed the WISDM (Phone) and WISDM (Watch) modalities. The average weighted F_1 results are 98%, 97% and 99% for WISDM (Phone), WISDM (Watch) and WISDM (Phone + Watch) respectively

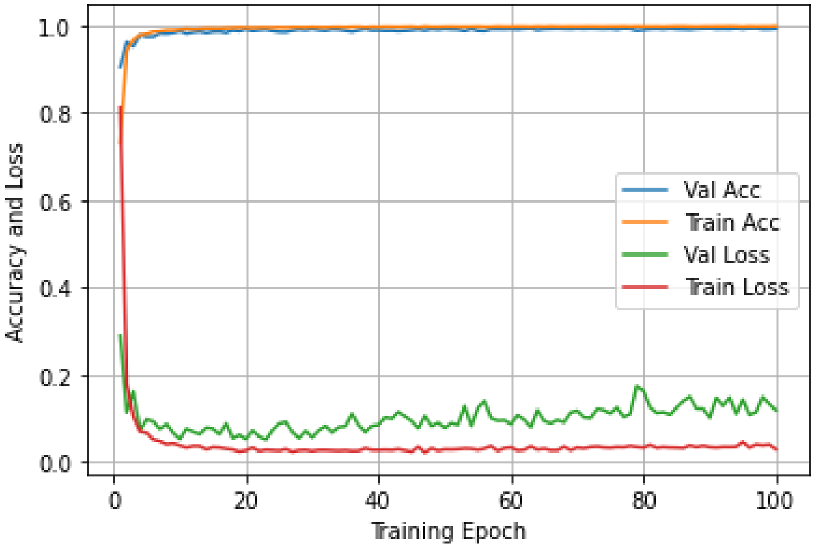


FIGURE 5. Accuracy and loss trends of proposed model for WISDM (Phone + Watch) dataset with Acc as accuracy and Val as validation.

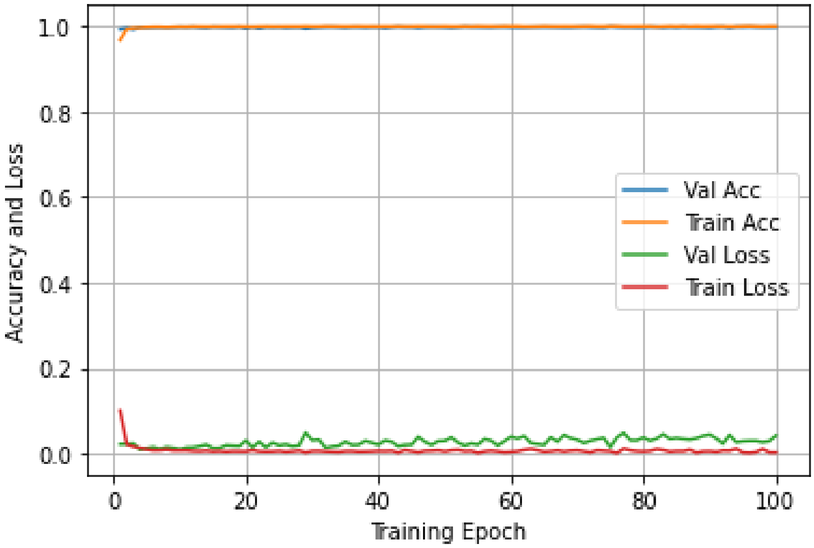


FIGURE 6. Accuracy and loss trends of proposed model for MHEALTH dataset with Acc as accuracy and Val as validation.

in percentage. We also noticed slightly lowered performance for some hand-oriented (Eating) activities for all three results. From these results, it is evident that the combination of the sensor modalities as seen with the WISDM (Phone + Watch) provides rich contextual data for better activity recognition. The results in Table 3 for the MHEALTH dataset also justifies the robustness of the proposed approach, with an average weighted F_1 score of 100%.

Furthermore, Figures 7 and 8 show the confusion matrix for the WISDM (Phone + Watch) and MHEALTH datasets. The results presented also demonstrate the ability of the proposed model to accurately and precisely recognise all the activities. The WISDM (Phone + Watch) confusion matrix also reports a slightly lowered result for Soup and Pasta similar to Table 2. These activities are purely hand-based requiring same or similar movements which go beyond the basic ambulation of body parts. The confusion for these activities can be attributed to the marginal variabilities of these activities resulting in high false positives. Table 4 shows the overall results from the cross-validation, especially the F_1 and MCC metrics which underscores the effectiveness of the proposed model. In general, the results achieved for both

TABLE 2. Precision, Recall and F_1 results for WISDM dataset.

Activity	Phone			Watch			Phone + Watch		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Walking	1.00	0.99	1.00	0.98	0.98	0.98	1.00	1.00	1.00
Jogging	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Stairs	0.99	0.99	0.99	0.97	0.97	0.97	1.00	0.99	1.00
Sitting	0.97	0.98	0.97	0.96	0.96	0.96	0.99	0.99	0.99
Standing	0.98	0.99	0.99	0.98	0.98	0.98	1.00	1.00	1.00
Typing	0.99	0.96	0.98	0.99	0.99	0.99	0.99	0.99	0.99
Teeth	0.99	0.98	0.99	0.99	1.00	0.99	0.99	0.99	0.99
Soup	0.97	0.94	0.96	0.97	0.94	0.96	0.99	0.99	0.99
Chips	0.95	0.96	0.95	0.91	0.92	0.91	0.98	0.99	0.99
Pasta	0.96	0.98	0.97	0.95	0.96	0.95	0.98	0.99	0.99
Drinking	0.96	0.96	0.96	0.93	0.95	0.94	0.99	0.98	0.99
Sandwich	0.95	0.97	0.96	0.90	0.87	0.88	0.99	0.98	0.99
Kicking	0.98	0.97	0.97	0.97	0.97	0.97	0.99	0.99	0.99
Catch	0.97	0.98	0.98	0.98	0.99	0.98	0.99	0.99	0.99
Dribbling	0.98	0.98	0.98	0.99	0.99	0.99	1.00	0.99	1.00
Writing	0.99	0.97	0.98	0.99	0.99	0.99	1.00	1.00	1.00
Clapping	0.99	0.99	0.99	1.00	1.00	1.00	0.99	0.98	0.99
Folding	0.99	0.98	0.99	0.97	0.99	0.98	0.98	0.99	0.99
Weighted Ave.	0.98	0.98	0.98	0.97	0.97	0.97	0.99	0.99	0.99

TABLE 3. Precision, Recall and F_1 results for MHEALTH dataset.

Activity	Precision	Recall	F_1
Standing still	1.00	1.00	1.00
Sitting and relaxing	1.00	1.00	1.00
Lying down	1.00	1.00	1.00
Walking	0.99	1.00	1.00
Climbing stairs	1.00	1.00	1.00
Waist bends forward	1.00	0.99	1.00
Frontal elevation of arms	1.00	0.99	1.00
Knees bending	1.00	1.00	1.00
Cycling	1.00	1.00	1.00
Jogging	1.00	1.00	1.00
Running	0.99	1.00	1.00
Jump front and back	0.99	1.00	0.99
Weighted Ave.	1.00	1.00	1.00

TABLE 4. Overall results (in percentage) with ACC as accuracy.

Dataset	ACC (%)	F_1 (%)	MCC (%)
Phone	97.91±2.007	97.91±2.007	97.85±1.822
Watch	96.60±1.473	96.60±1.473	96.57±1.215
Phone + Watch	99.13 ± 0.455	99.07 ± 0.418	98.91 ± 0.347
MHEALTH	99.73 ± 0.094	99.73±0.094	99.68 ± 0.049

TABLE 5. Performance comparisons with accuracy result (in percentage).

Model	WISDM		
	Phone	Watch	Phone + Watch
CNN [13]	50.0	72.0	-
LSTM [13]	74.0	79.0	-
FCN [14]	-	-	91.3 ± 0.53
Baseline CNN	88.4±2.376	86.8±2.265	90.33 ± 2.368
Baseline BLSTM	92.4±1.213	90.9±1.399	94.31 ± 1.236
MCBLSTM (This paper)	97.91±2.007	96.60±1.473	99.13 ± 0.455

datasets are very significant and most importantly justifying the combination of the sensing devices to achieve better results.

E. COMPARISONS

We compared the models by ranking the overall accuracy results as provided in Table 5. The results (in percentage) show that the proposed MCBLSTM had the highest overall accuracy. It performed better than the models by Benavidez and McCreight [13] and Burns and Whyne [14] for the WISDM dataset. The MCBLSTM model result was also better than the CNN and BLSTM baseline models given their rankings. Further, we compared the baseline models and the proposed MCBLSTM by carrying out paired t-tests to check the similarities between the baseline results and that of the proposed MCBLSTM. The following t-values and p-values are provided in Table 6 for CNN-MCBLSTM, BLSTM-MCBLSTM and CNN-BLSTM. The null hypotheses state that the above paired comparisons significantly reject the similarities between the results of the

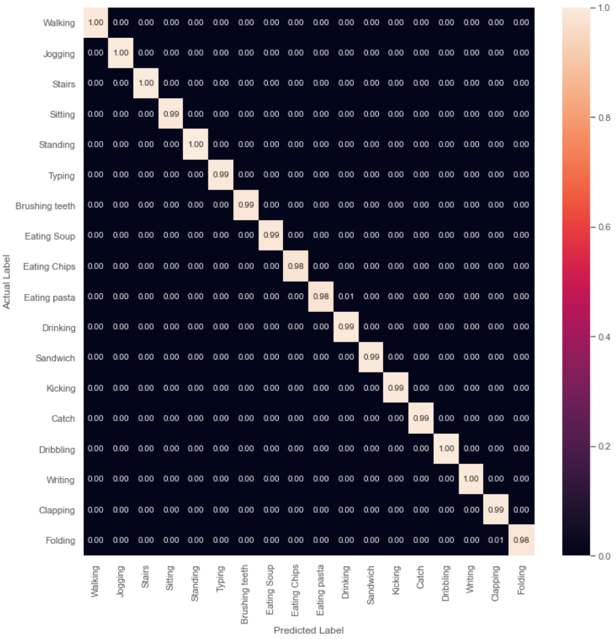


FIGURE 7. Confusion matrix for the WISDM (Phone + Watch) dataset.

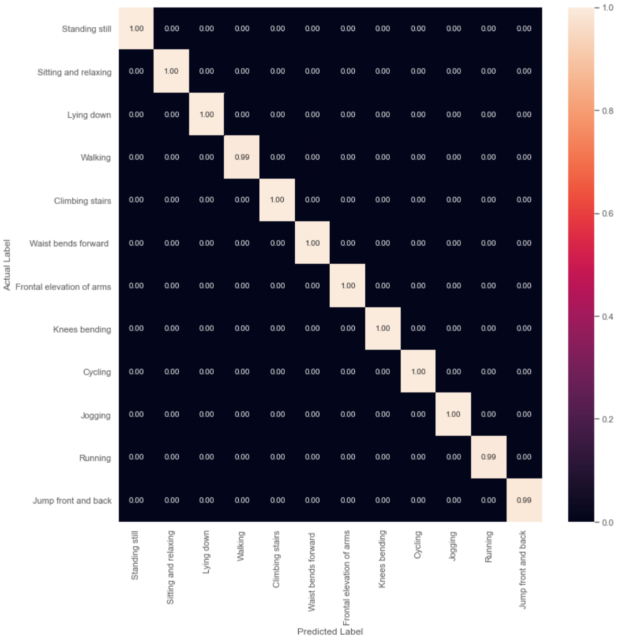


FIGURE 8. Confusion matrix for the MHEALTH dataset.

TABLE 6. Paired t-Test results for model performance comparisons.

Test Models	t-value	p-value
Baseline CNN - MBLSTM	36.4946	7.45290e-90
Baseline BLSTM - MBLSTM	36.5959	4.61942e-90
Baselines CNN - BLSTM	14.8999	3.60695e-34

TABLE 7. Performance results for different window sizes for the WISDM dataset.

Size (sec)	Accuracy (%)
5	98.6 ± 0.682
10	99.1 ± 0.455
15	92.6 ± 0.431
20	82.7 ± 0.513

compared models. Hence, we can rank the proposed MCBLSTM model as performing better than the baseline CNN and BLSTM models. This performance can be attributed to the fact that the MCBLSTM model takes advantage of the backward and forward sequences for time series data to outperform the CNN and BLSTM models. We also noticed that during training, the accuracy trend for both baseline models remained constant with loss increasing. The BLSTM layer of the proposed MCBLSTM helped to achieve a much lower loss sooner and stayed very low in comparison to the baseline models.

F. PARAMETER SETTING

Deep learning models require parameter tuning to achieve better results. There are three methods for parameter tuning–manual based method using the results of validation set and the experience in the domain, random based method involving the use of random values and grid search method which consists of the use of an exhaustive set of values. Whichever the process, they are all very time-consuming. We adopted a grid and search method for the work in this paper for which we set the learning rate to be 1e-3. We did this starting with a value and then a set of parameter values based on the validation result sets as a way of determining the optimal parameter value. As part of this process, we performed some experiments to determine the optimal filter map setting for the convolutional layers. Figure 9 shows the accuracy performance out of 1.0 for the MCBLSTM architecture with different values of filter maps. We also varied the window segments for the datasets to determine the optimal segment sizes. Results in Table 7 shows the suitability of 10 seconds segments for the WISDM dataset. Performance of the model reduces with increased segment sizes.

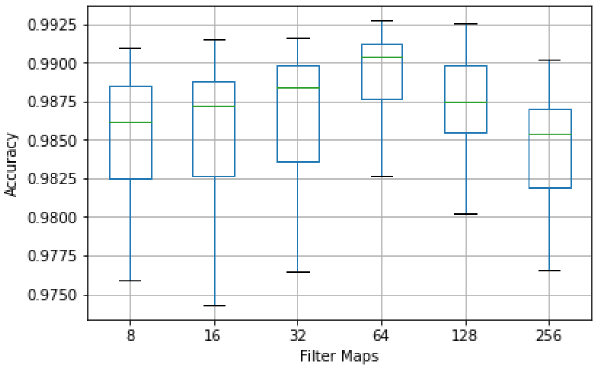


FIGURE 9. Box and whisker plots with overall accuracy results for different filter map selection on the WISDM (Phone + Watch) dataset.

V. CONCLUSION

The research into HAR is strategically significance to healthcare monitoring, surveillance and support for the elderly and cognitively impaired. Deep learning is becoming popular, and the related work demonstrates its suitability to improve and scale-up HAR. Deep learning CNN and Bidirectional LSTM models have been applied to text classification, speech recognition image classification with good results, but yet to combined in a multi-channel architecture for human activity recognition. In this paper, we propose a multi-channel CNN Bidirectional LSTM (MCBLSTM) based deep learning model to recognise activities from multimodal sensing devices. As part of the model, the CNN layers perform direct mapping and abstract representation of sensor inputs for feature extraction. Feature extraction is achieved using convolution operators applied to the kernels and to compute feature maps. The Bidirectional LSTM layer uses the internal state taking full advantage of the forward and backward sequence on the extracted features from the convolutional process. We evaluated the proposed model, using two publicly available datasets. The performance of the model was exceptional as multiple and complex activities from a large variability of movements, and diversity of body parts were recognised. Overall results showed excellent F_1 and MCC performance of the proposed model on the WISDM and MHEALTH datasets. We also showed that the model performs better with data from a combination of multimodal sensing devices. Furthermore, we compared the results of the proposed model with our baseline models – CNN and Bidirectional LSTM and with results reported by other authors on the datasets. The proposed model offered better results in terms of accuracy and other performance metrics. Our future work will be to extend the tuning method used in this work. We hope to explore and investigate better ways to automatically tune and adjust parameters rather than relying on the grid search method to help deep learning methods evolve in the learning process.

REFERENCES

[1] U. Akdemir, P. Turaga, and R. Chellappa, "An ontology based approach for activity recognition from video," in *Proc. 16th ACM Int. Conf. Multimedia MM*, 2008, pp. 27–31.

[2] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surveys*, vol. 46, no. 3, pp. 1–33, Jan. 2014.

[3] W. Ugulino, D. Cardador, K. Vega, E. Velloso, R. Milidiú, and H. Fuks, "Wearable computing: Accelerometers' data classification of body postures and movements," in *Advances in Artificial Intelligence—SBIA*. Springer, 2012, pp. 52–61.

[4] I. K. Ihianle, U. Naeem, and A.-R. Tawil, "Recognition of activities of daily living from topic model," *Procedia Comput. Sci.*, vol. 98, pp. 24–31, Jan. 2016.

[5] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose, "Fine-grained activity recognition by aggregating abstract object usage," in *Proc. 9th IEEE Int. Symp. Wearable Comput. (ISWC)*, Oct. 2005, pp. 44–51.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2016, *arXiv:1409.4842*. [Online]. Available: https://arxiv.org/abs/1409.4842

[8] D. Yu and L. Deng, "Automatic speech recognition: A deep learning approach," in *Signals and Communication Technology*. Springer, 2016.

[9] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*. [Online]. Available: http://arxiv.org/abs/1412.5567

[10] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[11] G. M. Weiss, "Wisdm smartphone and smartwatch activity and biometrics dataset," in *UCI Machine Learning Repository, WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 2019.

[12] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133190–133202, 2019.

[13] S. Benavidez and D. McCreight, "A deep learning approach for human activity recognition," in *Project Category, Other(Time-Series Classification)*. Stanford, CA, USA: Stanford Univ., 2019.

[14] D. M. Burns and C. M. Whyne, "Personalized activity recognition with deep triplet embeddings," 2020, *arXiv:2001.05517*. [Online]. Available: http://arxiv.org/abs/2001.05517

[15] J. Modayil, T. Bai, and H. Kautz, "Improving the recognition of interleaved activities," in *Proc. 10th Int. Conf. Ubiquitous Comput. UbiComp*, 2008, pp. 40–43.

[16] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," in *Proc. 10th Nat. Conf. Artif. Intell.*, San Jose, CA, USA, Jul. 1992, pp. 1–15.

[17] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 1–27, Jan. 2011.

[18] I. K. Ihianle, U. Naeem, S. Islam, and A. R. Tawil, "A hybrid approach to recognising activities of daily living from object use in the home environment," *Informatics*, vol. 5, no. 1, p. 6, 2018.

[19] I. K. Ihianle, U. Naeem, S. Islam, and A.-R. Tawil, "Recognising activities of daily living from patterns of object use," *Int. J. Hybrid Intell. Syst.*, vol. 14, no. 3, pp. 193–208, Mar. 2018.

[20] L. Chen and C. D. Nugent, "Ontology-based activity recognition in intelligent pervasive environments," *Int. J. Web Inf. Syst.*, vol. 5, no. 4, pp. 410–430, 2009.

[21] C. D. Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. 25th Int. Conf. Comput. Linguistics (COLING)*, Dublin, Ireland, 2014, pp. 69–78.

[22] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*. [Online]. Available: http://arxiv.org/abs/1510.03820

[23] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, Jan. 2018.

[24] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Conf. Artif. Intell. (IJCAI)*, Buenos Aires, Argentina, 2015, pp. 3995–4001.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] K. Zhang, W. L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 766–782.

[27] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," 2016, *arXiv:1610.09975*. [Online]. Available: http://arxiv.org/abs/1610.09975

[28] J. Nowak, A. Taspinar, and R. Scherer, "LSTM recurrent neural networks for short text and sentiment classification," in *Artificial Intelligence and Soft Computing* (Lecture Notes in Computer Science), vol. 10246. Springer, 2017, pp. 553–562.

[29] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[30] M. Wollmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 3949–3952.

[31] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, Jul. 2005.

[32] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," *Autom. Speech Recognit. Understand. (ASRU)*, pp. 273–278, 2013.

[33] A. Murad and J.-Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, Nov. 2017.

[34] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, "Deep residual bidir-LSTM for human activity recognition using wearable sensors," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Dec. 2018.

[35] X. Li, Y. Zhang, J. Zhang, S. Chen, I. Marsic, R. A. Farneth, and R. S. Burd, "Concurrent activity recognition with multimodal CNN-LSTM structure," 2017, *arXiv:1702.01638*. [Online]. Available: http://arxiv.org/abs/1702.01638

[36] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," in *Proc. IEEE 8th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2017, pp. 540–546.

[37] S. Hyun, I. Choi, and N. K. Soo, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Detection and Classification of Acoustic Scenes and Events*. Budapest, Hungary, 2016.

[38] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mHealthDroid: A novel framework for agile development of mobile health applications," in *Proc. 6th Int. Work-Conf. Ambient Assist. Living Act. Ageing (IWAAL)*, Belfast, U.K., Dec. 2014, pp. 91–98.

[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. 21st Eur. Symp. Artifcial Neural Netw., Comput. Intell. cMach. Learn., ESANN*, Apr. 2013, pp. 437–442.

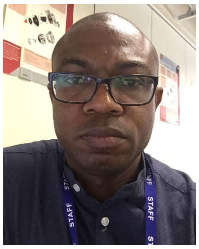
[40] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. D. R. Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sens. Syst. (INSS)*, Jun. 2010, pp. 233–240.

[41] S. H. Ebenuwa, M. S. Sharif, A. Al-Nemrat, A. H. Al-Bayatti, N. Alalwan, A. I. Alzahrani, and O. Alfarraj, "Variance ranking for multi-classed imbalanced datasets: A case study of One-Versus-All," *Symmetry*, vol. 11, no. 12, p. 1504, Dec. 2019.

[42] S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat, "Variance ranking attributes selection techniques for binary classification problem in imbalance data," *IEEE Access*, vol. 7, pp. 24649–24666, 2019.

[43] R. Delgado and X.-A. Tibau, "Why Cohen's kappa should be avoided as performance measure in classification," *PLoS ONE*, vol. 14, no. 9, Sep. 2019, Art. no. e0222916.

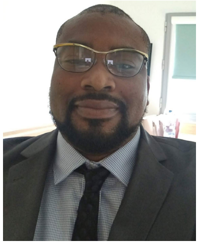
[44] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020.



ISIBOR KENNEDY IHIANLE (Member, IEEE) received the B.Sc. and Ph.D. degrees in computer science from the University of East London. He currently teaches with the Department of Computer Science, Nottingham Trent University, U.K. He has published several conference and journal papers. His research interests include human activity recognition from multi-modal sensing devices, health systems optimization, context awareness, knowledge-based systems with ontology, topic model and its applications, and data analytics. He is also a Fellow of the Higher Education Academy and a member of the British Computing Society.

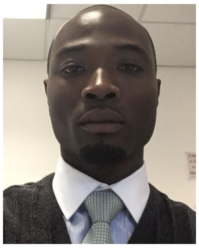
AUGUSTINE O. NWAJANA (Senior Member, IEEE) received the Ph.D. degree in electrical and electronic engineering from the University of East London, U.K., in 2017. From 2005 to 2009, he was a Telecommunications Engineer with Siemens AG, where his experience spanned many countries, including USA, U.K., United Arab Emirates, Germany, South Africa, Ghana, and Nigeria. He is currently an Academician with the University of Greenwich. His current research interests include the analysis and design of RF and microwave devices (including SIW and microstrip filters, diplexers/multiplexers, power dividers/combiners, couplers, and antennas) for modern communication systems. He was a recipient of the Federal Government of Nigeria Scholarship Award. He received the Research CPD Certificate in Practical Antenna Design: From Theory to Practice, from the University of Oxford, U.K., in 2019.

SOLOMON HENRY EBENUWA received the degree in physics, the M.Sc. degree in advanced computing, and the Ph.D. degree in computer science from the School of Architecture, Computing, and Engineering, University of East London (UEL). He is currently a Senior Lecturer with the Newham College of Further Education. He is also a Researcher in data mining, machine learning, decisions-based systems, and programming data structure in relation to machine learning algorithms. He was a Fellow of the U.K. Higher Education Academy. He has received many academic awards for innovative teaching and student support.



RICHARD I. OTUKA (Member, IEEE) received the B.Sc. and M.Sc. degrees in information systems and the Ph.D. degree in computer science from the School of Architecture, Computing, and Engineering, University of East London (UEL). He is currently an IT Lecturer with the Oaklands College of Further Education. He is also a Researcher in cloud computing and semantic technology. He has received many academic awards for innovative teaching and student support.

KAYODE OWA (Member, IEEE) is currently a Senior Lecturer with the Department of Computer Science, Nottingham Trent University, U.K. He has a total of 22 years of experience with expertise in diverse areas such as software engineering, system programming, artificial intelligence, machine learning, neural networks, modeling, data analytics, soft-computing, nonlinear model predictive control (NMPC) strategy, and the Big Data Scientist (Apache Hadoop, Spark, NoSQL, MapReduce, Yarn), probabilistic modeling, predictive analytics, soft computing metaheuristic and evolutionary algorithms, (neural networks, genetic algorithm, and particle swarm optimization), networking, security, forensic, mathematics, and business intelligence. He is also an Associate Fellow of the U.K. Higher Education Academy (AFHEA) and a member of the Institute of Electrical and Electronics Engineers (IEEE) Professional Association and the British Computer Society, (MBCS).



MOBOLAJI O. ORISATOKI received the B.Sc. degree from London South Bank University, in 2010, and the M.Sc. degree from Royal Holloway, University of London, in 2012, and the PGCE Institute of Education-University College London, in 2013. He is currently pursuing the Ph.D. degree with the Department of Engineering and Design, University of Sussex, U.K. He is currently an Associate Lecturer with the Department of Engineering and Design, University of Sussex. He has taught in different colleges across South and East London. His research interests include system optimization and control, system dynamics, and multi-agent systems.