**Literature Review – Seminar**

1. Abobakr, A., Hossny, M., Abdelkader, H., & Nahavandi, S. (2018). RGB-D Fall Detection via Deep Residual Convolutional LSTM Networks. 2018 Digital Image Computing: Techniques and Applications (DICTA).

Introduction - A robust fall detection system can be deﬁned as an assistive device that aims at detecting and alerting fall incidents. Hence, its main objective is to distinguish between fall events and normal activities of daily living. The signiﬁcant similarities of some ADL activities to falls challenge the robustness of fall detection systems. Accelerometer devices are the most commonly used wearable sensors for fall detection. Readings from the accelerometer attached to the human body are evaluated using thresholding or machine learning methods to detect fall events. Despite the effectiveness of wearable devices, they have limitations such as battery lifetime, being easily disconnected, and being forgotten. Moreover, wearing electronic devices is not preferable for the aging societies. Context-aware systems, on the other hand, rely on sensors deployed in the environment such as ﬂoor vibration sensors, microphones, and cameras to detect falls.

This paper proposes a vision-based integrable and automated fall detection system. The fall events are detected using an end-to-end deep machine learning model composed of convolutional and recurrent neural networks. The deep convolutional network extracts visual features from input sequences of depth frames. We use a ConvNet architecture that follows the residual learning approach to optimize the visual representations.

Dataset - The UR Fall Detection Dataset is used for training and evaluating the performance of the proposed method. So, the experiments have been conducted on the URFD dataset only. We split each of these activity classes into 80% for training and 20% for validation.

This dataset contains 70 (30 falls + 40 activities of daily living) sequences. Fall events are recorded with 2 Microsoft Kinect cameras and corresponding accelerometric data. ADL events are recorded with only one device (camera 0) and an accelerometer. Sensor data was collected using PS Move (60Hz) and x-IMU (256Hz) devices.  
The dataset is organized as follows. Each row contains a sequence of depth and RGB images for camera 0 and camera 1 (parallel to the floor and ceiling-mounted, respectively), synchronization data, and raw accelerometer data. Each video stream is stored in a separate zip archive in form of a png image sequence.

* Synchronization data contains frame number, time in milliseconds since sequence start, and interpolated accelerometric data.
* Raw accelerometric data contains time in milliseconds since sequence start and accelerometer data.

Proposed method – This paper proposes an end-to-end ConvLSTM model for fall event detection. This model combines a deep residual convolutional network ResNet with a recurrent LSTM neural network module. We use a ConvNet architecture that follows the residual learning approach (ResNet) to learn discriminative features from articulated body postures. The motivation for this combination is twofold. First, the ResNet model has powerful capabilities to learn and extract deep hierarchical visual features from raw input images. Second, using the extracted body features, the LSTM module can learn long-term temporal dynamics that can discriminate sequential input data, e.g., fall events.

* Preprocessing depth sequences

It has been concluded that depth images provide weak local gradient information of objects which makes it difﬁcult for deep learning models to generalize and biases the ConvNet model towards detecting objects silhouettes. Therefore, several depth encoding methods have been proposed to make efﬁcient use of depth measurements and provide a better learning signal for the deep network. combined the depth map as an additional channel with the RGB image forming an RGB-D modality. In, depth representation using the HHA encoding was proposed. Learning from HHA encoded maps have demonstrated better results than using raw depth data. This method spreads depth measurements over three RGB color channels. The values of RGB color components vary according to the distance from the depth camera, and hence, provide a more powerful input signal to the ConvNet model, ResNet in this work. Finally, an RGB color map is applied to produce the colorized depth image.

* Deep feature extraction using ResNet

ConvNets are the basic building block for the state-of-the-art methods for visual perception tasks such as object recognition, localization and detection, and semantic segmentation. This stack of layers learns multiple levels of feature extractors with increasing levels of abstraction. Incorporating ConvNets with the residual learning paradigm has led to the ResNet architecture which ensures better and faster generalization performance, easier optimization, and makes efﬁcient use of network depth. In this work, we use the ResNet model for learning and extracting visual features from input depth videos.   
  
The ResNet model is composed of residual blocks where each block learns a residual mapping regarding its input, instead of learning a direct unreferenced mapping. The layers of a residual block are formulated as CONV layers. This approach provides a strong initialization that reduces the effect of overﬁtting due to the small number of videos in fall detection datasets.

* LSTM modeling temporal dynamics

In addition to the hidden state, it incorporates a memory unit or cell state that is continuously modiﬁed using non-linear gating functions, which are learned. These gating functions manipulate the memory unit through forget and update operations to allow storing only relevant information.

1. Bulbul, E., Cetin, A., & Dogru, I. A. (2018). Human Activity Recognition Using Smartphones. 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).

Introduction - Smartphones are the most useful tools in our daily life and with the advancing technology they get more capable day by day to meet customer needs and expectations. An accelerometer has been standard hardware for almost all smartphone manufacturers. Since there is a meaningful difference in characteristics between data retrieved from these sensors, many features could be generated from these sensors' data to determine the activity of the person that is carrying the device. In this study, a dataset consisting of signals from the accelerometer and gyroscope of a smartphone carried by different men and women volunteers while doing different activities are classified using different machine learning approaches.

Dataset - Dataset consists of signals from a smartphone carried by 9 individuals performing 6 different activities. Activities performed are listed below with their corresponding codes.

* WALKING
* CLIMBING UP THE STAIRS
* CLIMBING DOWN THE STAIRS
* SITTING
* STANDING
* LAYING

Signals are recorded with a sampling rate of 50Hz and stored as time series for each dimension so 6 different signals were obtained (3 are from the accelerometer and the other 3 are from the gyroscope). The noise was filtered using median and 20Hz Butterworth filters to get more precise results. A second 3hz Butterworth filtering was applied to eliminate the effect of gravity in accelerometer signals. Values then normalized to (-1,1) interval. Euclid magnitudes of the values of 3 dimensions were calculated to merge 3-dimensional signals into one dataset. Finally, class codes (activity codes) given above for each row are added at the end of them among with the number that is given to each individual. In the end, the dataset consists of 2947 records with 561 features.

Proposed method - Supervised machine learning is used to recognize activity from dataset records. Different supervised machine learning models are designed using different classification approaches. Designed models first trained with training data that consists of 80% of the total dataset and then tested with the rest. Classification precision of models is tested and observed using 5-fold cross-validation. Methods used for classification are as follows:

* Decision Trees
* Support Vector Machines
* K-nearest neighbors (KNN)
* Ensemble classification methods
  + Boosting
  + Bagging
  + Stacking

1. Wang, H., Zhao, J., Li, J., Tian, L., Tu, P., Cao, T., … Li, S. (2020). Wearable Sensor-Based Human Activity Recognition Using Hybrid Deep Learning Techniques. Security and Communication Networks, 2020, 1–12.

Introduction - Human behavior recognition (HAR) is the detection, interpretation, and recognition of human behaviors, which can use smart health care to actively assist users according to their needs. Human behavior recognition has wide application prospects, such as monitoring in smart homes, sports, game controls, health care, elderly patients care, bad habits detection, and identification. It plays a significant role in in-depth study and can make our daily life smarter, safer, and more convenient. This work proposes a deep learning-based scheme that can recognize both specific activities and the transitions between two different activities of short duration and low frequency for health care applications.

Dataset - This paper adopts the international standard Data Set, Smartphone-Based Recognition of Human Activities, and Postural Transitions Data Set to conduct an experiment, which is abbreviated as HAPT Data Set. The data set is an updated version of the UCI Human Activity Recognition Using popularity Data set. It provides raw data from smartphone sensors rather than preprocessed data and collects data from accelerometer and gyroscope sensors. In addition, the action category has been expanded to include transition actions. The HAPT data set contains twelve types of actions. A total of 815,614 valid pieces of data are used here.

Proposed method - The overall architecture diagram of the method proposed in this paper contains three parts. The first part is the preprocessing and transformation of the original data, which combines the original data such as acceleration and gyroscope into an image-like two-dimensional array. The second part is to input the composite image into a three-layer CNN network that can automatically extract the motion features from the activity image and abstract the features, then map them into the feature map. The third part is to input the feature vector into the LSTM model, establish a relationship between time and action sequence, and finally introduce the full connection layer to achieve the fusion of multiple features. In addition, Batch Normalization (BN) is introduced, in which BN can normalize the data in each layer and finally send it to the Softmax layer for action classification.

1. Agarwal, P., & Alam, M. (2020). A Lightweight Deep Learning Model for Human Activity Recognition on Edge Devices. Procedia Computer Science, 167, 2364–2373.

Introduction - Here the architecture for the proposed Lightweight model is developed using Shallow Recurrent Neural Network (RNN) combined with Long Short Term Memory (LSTM) deep learning algorithm. then the model is trained and tested for six HAR activities on resource-constrained edge devices like RaspberryPi3, using optimized parameters. The experiment is conducted to evaluate the efficiency of the proposed model on the WISDM dataset containing sensor data of 29 participants performing six daily activities: Jogging, Walking, Standing, Sitting, Upstairs, and Downstairs. And lastly, the performance of the model is measured in terms of accuracy, precision, recall, f-measure, and confusion matrix and is compared with certain previously developed models.

Dataset - Here Android smartphone having an inbuilt accelerometer is used to capture tri-axial data. The dataset consists of six activities performed by 29 subjects. These activities include walking, upstairs, downstairs, jogging, standing, and sitting. Each subject performed different activities by carrying a cell phone in the front leg pocket. A constant Sampling rate of 20 Hz was set for the accelerometer sensor. A detailed description of the dataset is given in table 1 below.

Total no of samples: 1,098,207

Total no of subjects: 29

Activity Samples: Percentage

Walking 4,24,400 38.6%

Jogging 3,42,177 31.2%

Upstairs 1,22,869 11.2%

Downstairs 1,00,427 9.1%

Sitting 59,939 5.5%

Standing 48,397 4.4%

Proposed method - The working of the Lightweight RNN-LSTM-based HAR system for edge devices. The accelerometer reading is partitioned into fixed window size T. The input to the model is a set of readings (x1, x2, x3,…….,xT-1, xT) captured in time T, where xt is the reading captured at any time instance t. These segmented window readings are then fed to the Lightweight RNN-LSTM model. The model uses the sum of rule and combine output from different states using a softmax classifier to one final output of that particular window.

1. Das, S., Partha, S. B., & Imtiaz Hasan, K. N. (2020). Sentence Generation using LSTM Based Deep Learning. 2020 IEEE Region 10 Symposium (TENSYMP).

Introduction - Sentence generation means generating new words or sequences according to a specific context. This process can be considered a subset of Natural Language Generation. In sentence generation, preserving the context is strictly maintained. This preservation of context assures the process of maintaining both synaptic and semantic rules.

Thus the resulting sequence secures the correct format as well as the relevant meaning of it. Sentence generation can be useful in text summarization, providing suggestions in the search section, and also in producing predicted replies for an automated system or chatbot. The proposed system is trained with the relevant data and the objective of the system is to understand the given input sequence and generate relevant new words.

Dataset – The dataset consists of 4515 examples and contains Author\_name, Headlines, Url of Article, Short text, and Complete Article. The summarized news is gathered from various sources such as Inshorts and only scraped the news articles from Hindu, Indian times, and Guardian. The period ranges from February to august 2017.

Proposed method – The system is trained with a suitable dataset containing a large number of words. The unique words are used to form a vocabulary. Consequently, the labels and features are retrieved and Long Short-Term Memory architecture is used to generate new words according to the context.

* Data collection and description

Collect data and prepare model

* Processing data for model

Process collected data and train the model

* + Creating vocabulary

Here the words are split and a vocabulary of unique words is created

* + Tokenization

Each word gets tokenized meaning each of the words can be uniquely identified using a number

* + N-grams generation
  + Padding

N-grams generated may be of different lengths, here we use padding to acquire uniformity in length

* + Retrieving labels and features
  + One-hot encoding
* Proposed model

In this paper, the proposed model has a total of seven sequential layers

* + Embedding layer

In this layer, a set of words is mapped into vector forms to improve the ability of neural networks because working on numerical data is much easier

* + LSTM
  + Bidirectional LSTM

Bidirectional LSTM has two hidden states working in opposite directions and hence can work on both past and future states at a time

* + Flatten layer

This layer takes the output of the previous layer and puts the value in a single vector

* + Dense layer

The dense layer in our model uses a softmax activation function which is nonlinear and uses adam optimizer to fit features and labels

* + Generating new word

After training the model, input is given to it in tokenized form. The model predicts a new word after each iteration maintaining the context. After generating every new word, the word also is added to the previous input and the new combination is considered the next input.

1. Shojaei-Hashemi, A., Nasiopoulos, P., Little, J. J., & Pourazad, M. T. (2018). Video-based Human Fall Detection in Smart Homes Using Deep Learning. 2018 IEEE International Symposium on Circuits and Systems (ISCAS).

Introduction - The concept of a “smart home” is a major step towards wellness and improved quality of life and a hot interdisciplinary research topic bringing together artificial intelligence, cloud computing, communications and networks, psychology, and healthcare. Monitoring the well-being of the residents is an expected service to be provided by a smart home. Wearable devices are relatively inexpensive and can directly measure kinematic quantities. Nowadays, inexpensive depth cameras, such as the Microsoft Kinect, can address some of the privacy issues and under proper implementation, conditions could be a promising and feasible option for human fall detection in the context of a smart home.

Dataset – NTU RGB+D is a large-scale dataset for RGB-D human action recognition. It involves 56,880 samples of 60 action classes collected from 40 subjects. The actions can be generally divided into three categories: 40 daily actions (e.g., drinking, eating, reading), nine health-related actions (e.g., sneezing, staggering, falling), and 11 mutual actions (e.g., punching, kicking, hugging). These actions take place under 17 different scene conditions corresponding to 17 video sequences (i.e., S001–S017). The actions were captured using three cameras with different horizontal imaging viewpoints, namely, −45∘,0∘, and +45∘. Multi-modality information is provided for action characterization, including depth maps, 3D skeleton joint position, RGB frames, and infrared sequences. The performance evaluation is performed by a cross-subject test that split the 40 subjects into training and test groups, and by a cross-view test that employed one camera (+45∘) for testing, and the other two cameras for training.

Proposed method - Because the 3D locations of major body joints carry most of the body kinematic information required for discriminating different actions, keeping track of the body joints, as shown by our evaluations, proves to be sufficient for action recognition and fall detection, while it is computationally much cheaper. Since the existing algorithms to extract skeletons from the depth map, such as the one provided by the Microsoft Software Developer Kit, process the video sequence frame by frame, the use of body skeleton information by our model can be implemented in real-time. As actions usually take place within a long sequence of frames, vanilla RNN encounters the vanishing gradient issue, so we specifically take LSTM, which can go deep in time.

As a deep neural network, LSTM requires abundant training data. We first train a multi-class LSTM on the abundant samples of human regular actions. Then, we transfer all the learned weights, except for those of the last layer, to a two-class LSTM that is designed for fall detection. Finally, we train the last layer of the two-class LSTM on scarce human fall samples in combination with part of the regular action samples. To prevent the LSTMs from getting biased toward training data. These include the depth of the LSTM in time, the number of layers, the number of the hidden units in each layer, and the dropout ratio.

1. Alemayoh, T. T., Hoon Lee, J., & Okamoto, S. (2019). Deep Learning-Based Real-time Daily Human Activity Recognition and Its Implementation in a Smartphone. 2019 16th International Conference on Ubiquitous Robots (UR).

Introduction - Human activity recognition is a broad area of study mainly concerned with identifying specific movement or action of a person based on given input data. Mostly, input data signals are obtained from videos, where video frames are taken for analysis or multi-axis time-series IMU devices. Comparably, wearable IMU sensors became a popular and convenient way of data collection mechanisms without an extensive installation of equipment and privacy issues.

The processed version of the data was used to fit statistical and machine learning models such as SVM as in Anguita et al. Deep learning methods have shown the capability and even achieve state-of-the-art results by automatically learning high-level and meaningful features from raw data. In large-scale data classification, CNN is competent to extract features from signals and it has demonstrated excellent performance in image classification, speech recognition, and sentence classification.

Dataset – The smartphone used was attached tightly to the waist of the subjects. Out of the various motion-related sensors of a smartphone, the 3-axis Acc and 3-axis Gyro were chosen for a better result. Motion data of eight activities were collected. The activities are: walking, jumping, running, bicycle riding, stairs ascending, and descending, laying down, and still.

Proposed method - A CNN is applied to the activities’ one-dimensional virtual images prepared. Each convolutional layer performs a 2D convolution on its inputs followed by a non-linear activation function, ReLU (Rectifier Linear Unit). To reduce the effect of internal covariance shift of activations, batch normalization was utilized, which forces each mini-batch input of a layer to have similar distribution throughout the hidden layers. Besides, it allows the use of larger learning rates to speed up the optimization process. After the output of the second pooling is flattened to form a

long 1D feature map vector, the classification is decided by the probability distribution of an eight-class softmax layer. All the parameters of the network are updated by Adam optimizer using back optimization.

1. Ullah, M., Ullah, H., Khan, S. D., & Cheikh, F. A. (2019). Stacked Lstm Network for Human Activity Recognition Using Smartphone Data. 2019 8th European Workshop on Visual Information Processing (EUVIP)

Introduction - With the exponential growth of computing technology, wear-able electronics are widespread in human communities for daily usage. With the daily usage of the smartphone, embedded sensors like accelerometers and gyroscopes produce a large amount of useful data that can be used to automatically predict and classify human activities. Potentially, human activity recognition can be used in elderly houses, especially in the countries where the average old population is on the rise. In essence, the human activity algorithm can be divided into the following two broad categories – Vision based & Sensor based

Dataset - The network is evaluated on a public domain UCI dataset and quantitative results are compared against six state-of-the-art methods. The performance is calculated in terms of precision-recall and average accuracy.

Proposed method – The method mainly consists of two parts i.e. a single layer neural network and a network of stacked LSTM cells. Initially, sensor data is obtained from the smartphone that is worn by a human subject. We used two types of sensor data i.e. accelerometer and the gyroscope. The raw sensor data is passed through a single-layer neural network which acts as a pre-processing and normalized input data for the succeeding network. The normalization is achieved through a linear discriminant function and ReLU activation. After that, the data is fed to the stacked LSTM network. The network consists of five LSTM cells that have learned the temporal dependencies of the sensor sequential data. The output of the stacked LSTM network is given to a six-way softmax which gives the individual probability of the six human behavior i.e. (walking, walking upstairs, walking downstairs, sitting, standing, lying).

1. Sun, B., Liu, M., Zheng, R., & Zhang, S. (2019). Attention-based LSTM Network for Wearable Human Activity Recognition. 2019 Chinese Control Conference (CCC).

Introduction - Human activity recognition is an important area of research in ubiquitous computing, human behavior analysis, and human-computer interaction. It can be used widely, including in health monitoring, smart homes, and human-computer interactions. HAR focuses on the motion data from smart sensors such as accelerometers, gyroscopes, Bluetooth, light sensors, and so on. Although the video-based recognition method excels other recognition methods in indoor activity, it has several restrictions such as space limitation and interference from the environment. With the development of sensor technology and computing power, the sensor-based HAR is becoming more promising with privacy well-protected.

In this paper, we describe an LSTM network with an attention mechanism, which can automatically focus on the time series that has a decisive effect on classification, to capture the most important temporal dependencies from the input, without using extra handcrafted features and human domain knowledge. The attention mechanism allows a model to learn a set of weights over raw sensor data, which we leverage to weight the temporal context.

Dataset - The Opportunity data set for HAR from wearable, object, and ambient sensors is a data set devised to benchmark human activity recognition algorithms. The data sets contain- s activities from 4 subjects and each has 6 different runs. Five of them, termed activity of daily living, are composed of temporally unfolding situations in which a large number of action primitives occur. The remaining one, a drill run, is designed to generate a large number of scripted sequence instances. Notably, we use 17 mid-level gesture classes for predictions. This group contains the “NULL” class, which is common, for a total of 18 classes.

Proposed method - The model contains four components - Input layer: input sensor data to this model, LSTM layer: utilize LSTM to get high-level features, Attention layer: produce a weight vector, and merge features from each time step into a temporal feature vector to find relevant temporal context by multiplying the weight  
vector and Output layer: the temporal feature is finally used for activity recognition.

1. Deep, S., & Zheng, X. (2019). Hybrid Model Featuring CNN and LSTM Architecture for Human Activity Recognition on Smartphone Sensor Data. 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT).

Introduction - The proliferation of smartphones with various embedded sensors have eased the method of gathering human activity data in recent time. With the development of unprecedented characteristics of sensors such as accelerometers and gyroscopes, sensor-based human activity recognition has received extensive concerns. In wearable-based HAR, sensors or other external devices are attached to the human body. HAR is a method of predicting activities from the data obtained from sensors. The process involves extracting motion features and classifying the activities into different categories. The data collected from the sensors are a sequence of time series data and traditional machine learning algorithms may not exploit the temporal correlations of input data.

In this paper, we use a combination of CNN and LSTM for HAR. Furthermore, we also apply LSTM for activity recognition tasks in the same dataset and compare the results with the CNN-LSTM model.

Dataset - To evaluate the effectiveness of the CNN-LSTM model, we experiment on the UCI HAR dataset. The dataset consists of time series data collected from 30 volunteers of the 19-48 age group. Each volunteer performed six activities (walking, walking upstairs, walking downstairs, sitting, standing, laying) with a smartphone attached to their waist. The 3-axial linear acceleration (tAcc-XYZ) from the accelerometer and 3-axial angular velocity (tGyro-XYZ) from gyroscope data were collected. The data were collected with a constant sampling rate of 50Hz. The activities were video-recorded for ground truth and data were manually labeled. The dataset is randomly divided into 70% training and 30%  
testing data.

Proposed method - We used a kernel size of 6 and several filters 128 for both the convolutional layers. The output data size is then passed through the dropout layer. This output data is further used to pass through the LSTM layer. The output 3D data is fed to the LSTM layer.

The LSTM layer is used in this hybrid architecture because it works well with the time series data and is designed to handle time dependence problems. Each LSTM layer in this architecture produces hidden cell information. The LSTM layer is followed by a dense layer, hyperbolic tangent activation, and a soft-max layer at the end. It is then passed through a dense layer with hyperbolic tangent activation and used Adam optimizer which ends the LSTM networks in this hybrid model.