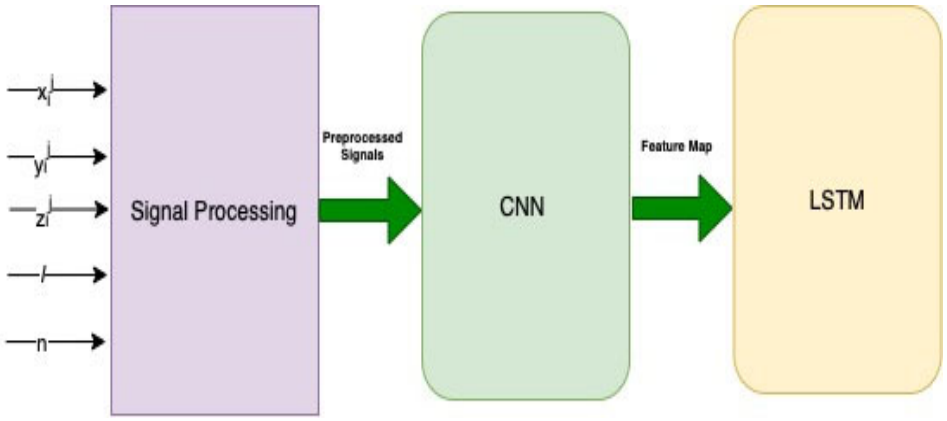

Figure S1. An illustration of inputs from multiple sensors of the proposed HAR hybrid model.

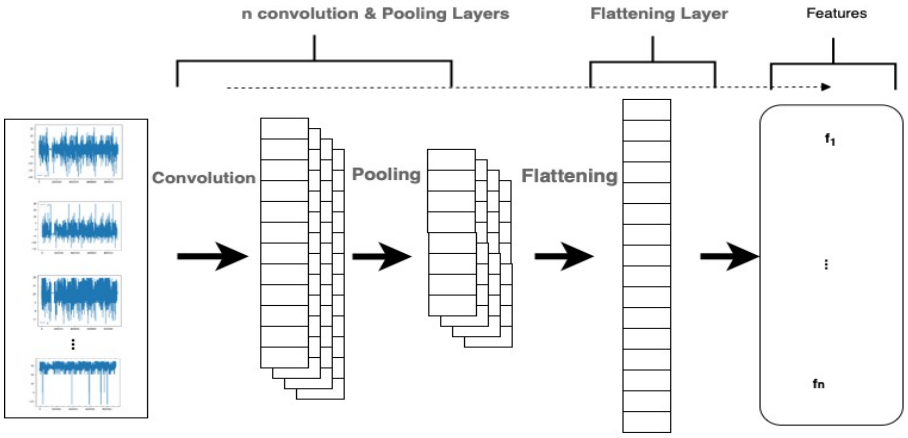

Figure S2. Automatic Feature Extraction Using CNN

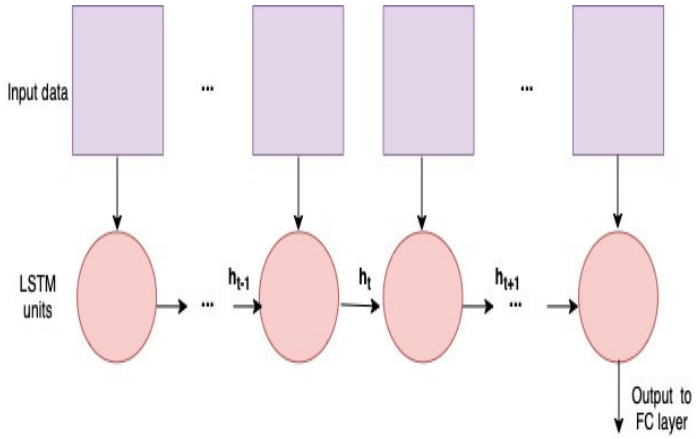

Figure S3. A single layer of LSTM Network

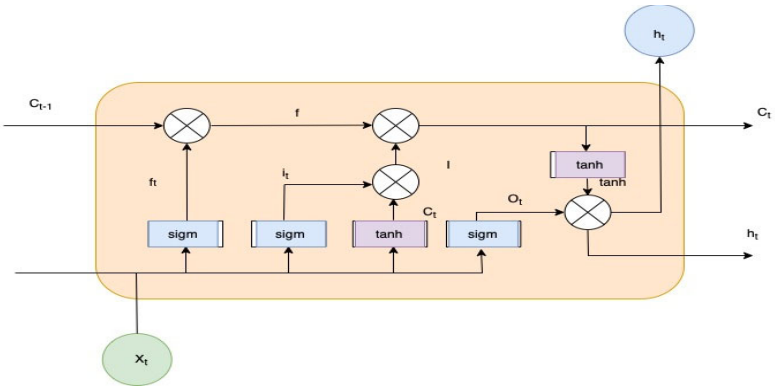

Figure S4. An overview of the architecture of the LSTM cells

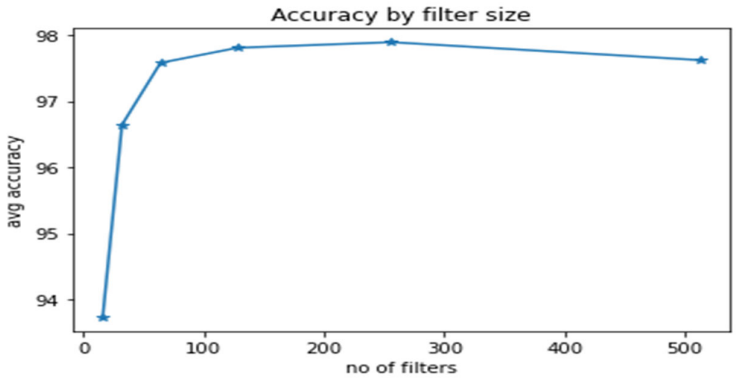

Figure S5. Average Accuracy by Number of Filters

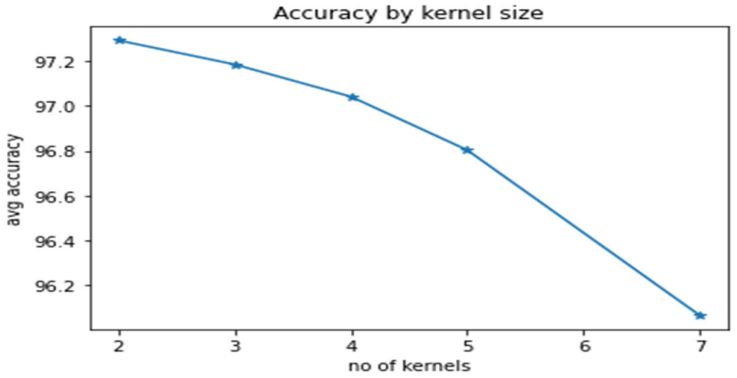

Figure S6. Average Accuracy by Kernel Size

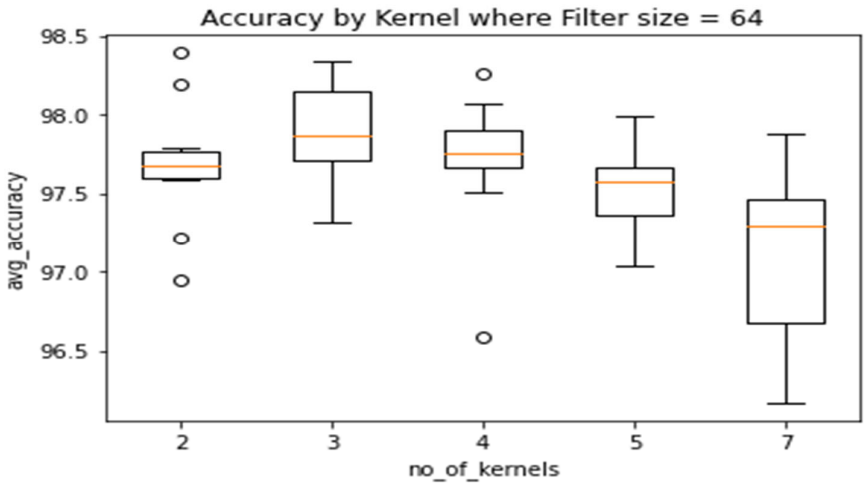

Figure S7. Box Plot for different kernels with filter size of 64

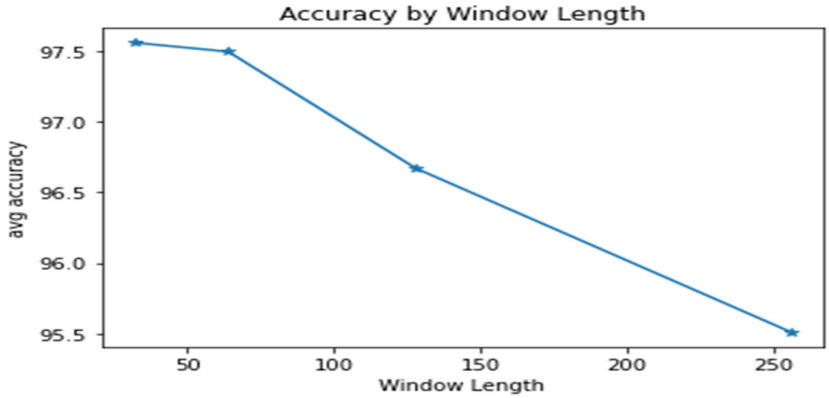

Figure S8. Performance by Window Length

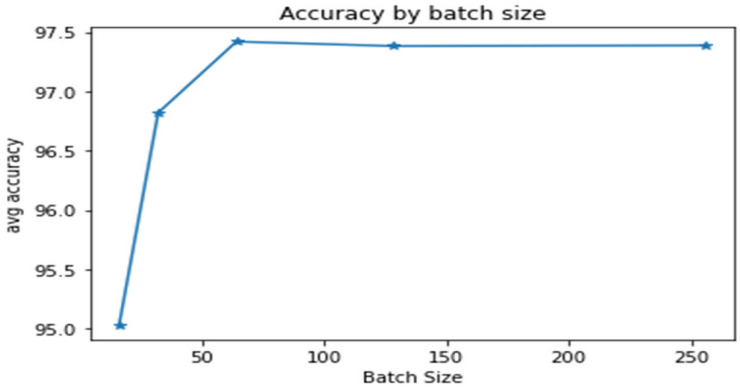

Figure S9. Performance by Batch Size

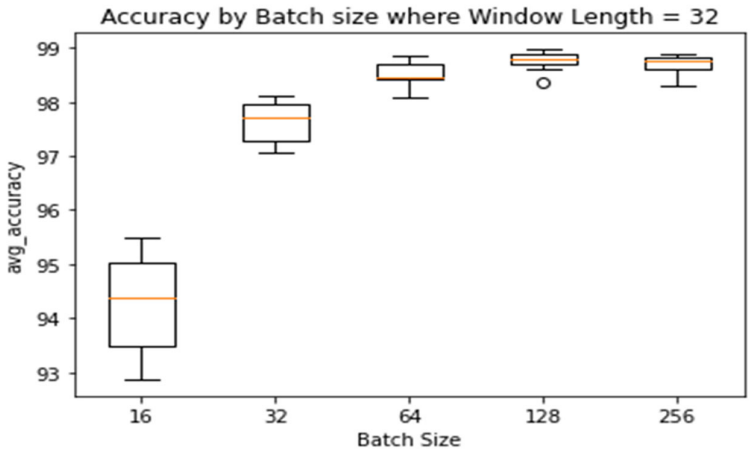

Figure S10. Box plot of different batch sizes for window length 32

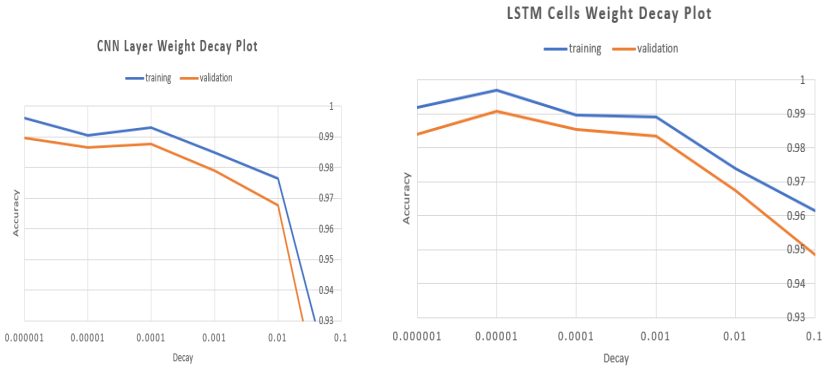

Figure S11. Line plot showing Accuracy over the weight decay between $1e^{-1}$ and $1e^{-6}$ for (a) the CNN layer and (b) the LSTM cells.

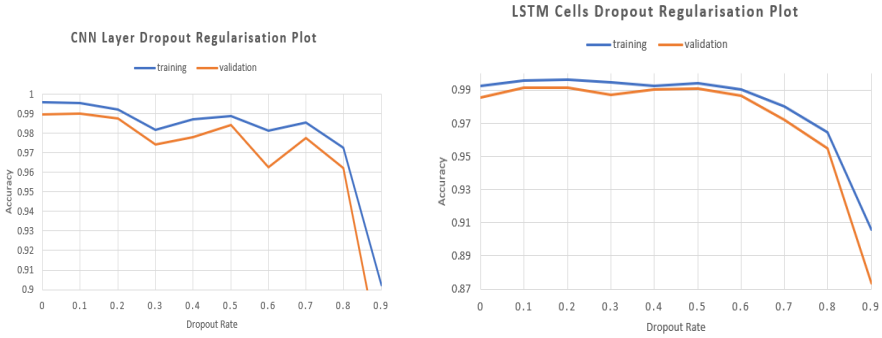

Figure S12. CNN-LSTM Model Dropout regularization plots. (a) shows the CNN layer accuracy for training and validation over dropout rate between 0 and 0.9 (b) shows the LSTM cells accuracy for training and validation over dropout rate between 0 and 0.9

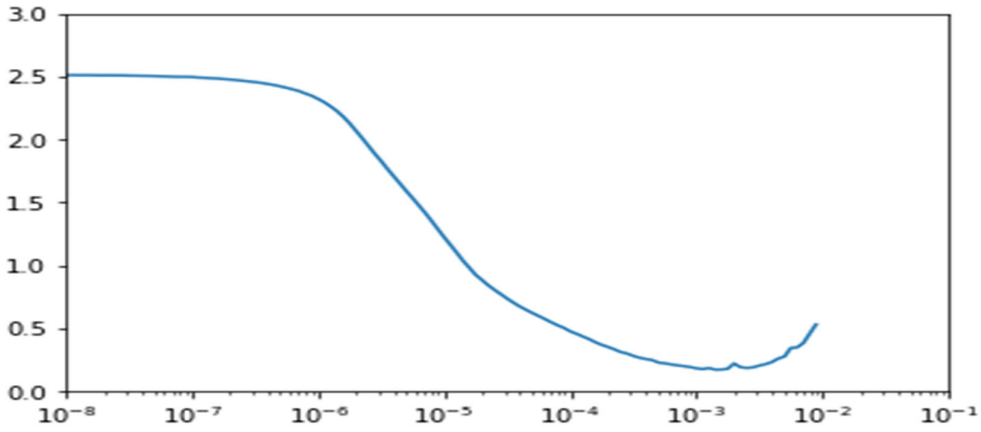

Figure S13. Line plot of loss over learning rate

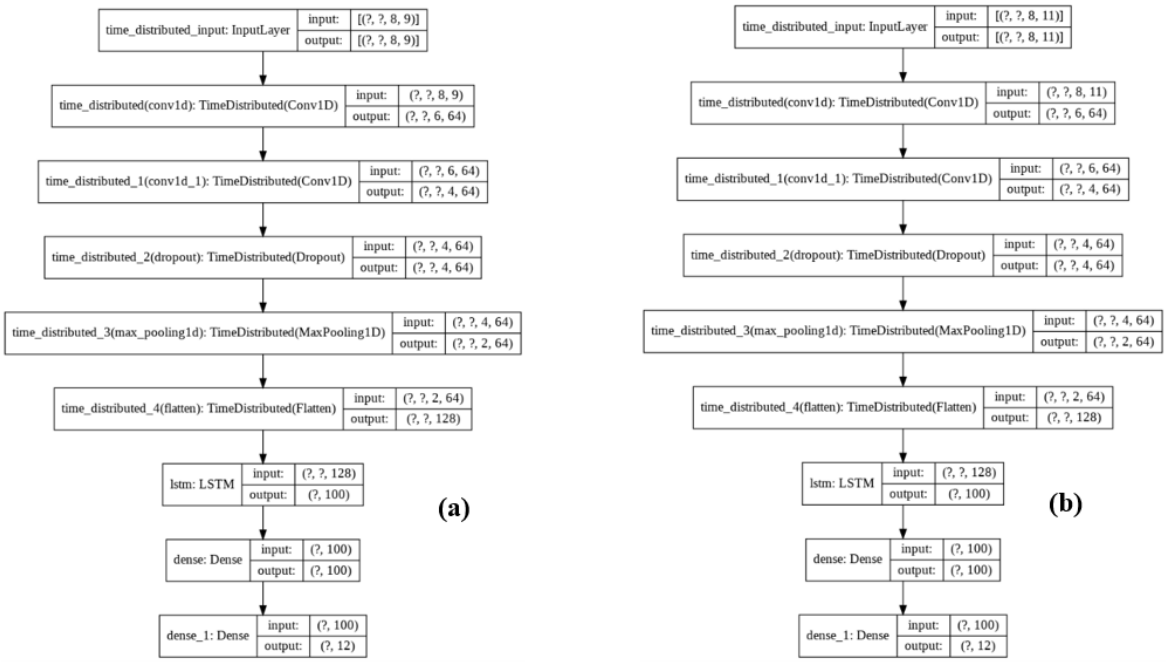

Figure S14. Low level details of the Baseline CNN-LSTM model(a) and Context-aware CNN-LSTM Model(b). The difference between these models is at the TimeDistributed input layer where the baseline model trained on 9 inputs (using the traditional tri-axial inertial sensors and the context-aware model trained on 11 inputs i.e., adding 2 inputs for context data from microphone and light sensor.