Comparative Analysis of CNN, RNN, and SVM for Smartphone-Based Human Activity Recognition

Clarissa Man and Merjem Memic
Department of Statistics, University of Michigan, USA
{merjemm, yiman}@umich.edu

Abstract

A pool of 30 volunteers performed six activities (WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS, SITTING, STANDING, LAYING) with a smartphone at the waist. The smartphone's embedded systems recorded triaxial acceleration and angular velocity, which form the basis for a multi class classification prediction experiment. A Support Vector Machine, Convolutional Neural Network, and a Recurrent Neural Network were trained on the resulting data and created predictions to varying success. In accuracy and F1 Score, the SVM outperforms the other models. Each faced issues distinguishing between different sets of seemingly similar activities.

1 Introduction

Human Activity Recognition (HAR) is the task of classifying different human activities based on data collected from wearable sensors, such as accelerometers and gyroscopes embedded in smartphones. This problem has significant importance in various domains, including health monitoring, fitness tracking, elder care, and smart home systems, where accurately identifying activities can enable personalized interventions, improve quality of life, and enhance safety. Despite the growing interest in HAR, achieving high accuracy remains challenging due to the variability in activity patterns, sensor noise, and user-specific differences.

In this paper, we address the problem of classifying six human activities—walking, walking upstairs, walking downstairs, sitting, standing, and lying—based on multivariate time-series data from smartphone sensors. Our models take as input sequences of raw accelerometer and gyroscope signals, represented as time-series data with nine channels corresponding to the x, y, and z axes of the body acceleration, body gyroscope, and total acceleration signals. The output of the models is the predicted activity label for each sequence, representing the most likely activity being performed. This classification task involves identifying both spatial patterns in the data (e.g., feature correlations across axes) and temporal dependencies (e.g., repetitive motion over time). By comparing the performance of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units, and a baseline Support Vector Machine (SVM), we aim to demonstrate how modern deep learning techniques can improve upon traditional approaches for this critical problem.

2 Related work

Human Activity Recognition has been a growing topic over the past years because of the widespread application to health and fitness sectors. Technology has evolved to allow easier collection of human activities through wearable devices with sensors, the most common being a smartphone. There are a variety of approaches to analyze and accurately predict movements based on raw data. In the initial papers that investigate Human Activity Recognition and its relation to smartphones, more traditional methods like decision trees or Support Vector Machines are used to create models [4, 5, 6].

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

They attempt to use the data to predict people's activities, but acknowledge the limits to their data and the impact of the variation in a phone's specifications [6]. SVM struggles to process complex models compared to CNN and RNN, but it can still perform well on relatively smaller datasets. In addition, it depends on labeled data and trouble to generalize their data beyond their experiments. Other research has been done to explore deep learning approaches to Human Activity Recognition including CNN, RNN, and BLSTM [2, 3, 9]. One paper used already accessible datasets and multiple deep learning and found CNN to outperform the additional approaches [3]. CNNs excel at capturing spatial patterns in data, offering strengths in precision, recall, and F1-score. There has been success found in combining models [3, 11, 12]. For example, a paper published in 2022 by Zhu et al. [11] introduced a hybrid classifier combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for spatial-temporal pattern extraction in human activity recognition (HAR) using radar data. This approach achieves a classification accuracy of approximately 90.8% for nine-class HAR, validated through K-fold cross-validation and leave-one-person-out methods. The current cutting edge development in HAR based on smartphone data is the use of Few-Shot Learning. In a paper published in early October of 2024, researchers found high accuracy in a novel approach newly named MADA which "combines meta-learning with automated data augmentation." [9] This approach aimed to allow the model to be versatile and adapt to different people's manner of doing activities and the other variable elements of the individual's environment, of which all of the other models struggled with [7]. This approach was robust and achieved higher accuracy on all chosen HAR datasets compared to CNN and other models employed.

3 Dataset and features

The dataset used in this study is the Human Activity Recognition (HAR) dataset, which contains time-series data recorded from wearable sensors. These sensors captured signals during six activities: walking, walking upstairs, walking downstairs, sitting, standing, and lying. The training dataset consists of 7,352 examples, and a validation set was created by splitting 20% of the training data. The test dataset contains 2,947 examples. Each sample consists of 128 time steps and nine sensor features, which include accelerometer readings (*body_acc_x*, *body_acc_y*, *body_acc_z*), gyroscope readings (*body_gyro_x*, *body_gyro_y*, *body_gyro_z*), and total acceleration readings (*total_acc_x*, *total_acc_y*, *total_acc_z*). Preprocessing involved normalizing the features across all samples and time steps to achieve zero mean and unit variance, which aids in model convergence during training. Additionally, the activity labels were adjusted to be zero-based, facilitating the use of sparse categorical cross-entropy as the loss function. The resulting data shape was (7352, 128, 9) for the training set, with similar dimensions for the test set (2,947, 128, 9). This dataset provides a rich set of features for recognizing human activities from wearable sensors.

4 Methods

Depending on the model, the data needed to be read in as a different format and within a different container. Originally, it was read in using functions from the pandas library. Later, we would convert it using library functions to convert it to numpy arrays. Meaningful column labels weren't used because they wouldn't improve or impact our machine learning models.

The first algorithm we implemented was a Support Vector Machine since it was the model mentioned in the paper where the data was sourced. Using GridSearchCV from the sklearn selection model library, we try to find the best kernel for predicting on the training data using a set degree, c, and gamma hyperparameters. By doing so, it was shown that a poly kernel works best. Then, the process is repeated using only the poly kernel and the degrees and C values of [1, 3, 5, 7, 10]. This process results in finding that a degree of 3 and a C of 10 work best on the data. Because of the computational power needed to do GridSearchCV on the data, we didn't testing it on a larger amount of hyperparameters. When attempting to run it with a range of 10 values for the degree and C, it didn't complete after running for 5 minutes. To do the first computation to find the best kernel, it took over 3 minutes to fully run, while it took a similar amount of time to find the best degree and c. This results in an SVM with a poly kernel and degree of 5 and a C of 10 as the final first model. By using 'GridSearchCV', there is more confidence in the strength of these model hyperparameters compared to the others. We wanted to test different kernels and see which would fit best according to the validation scores rather than get fixed on certain sets.

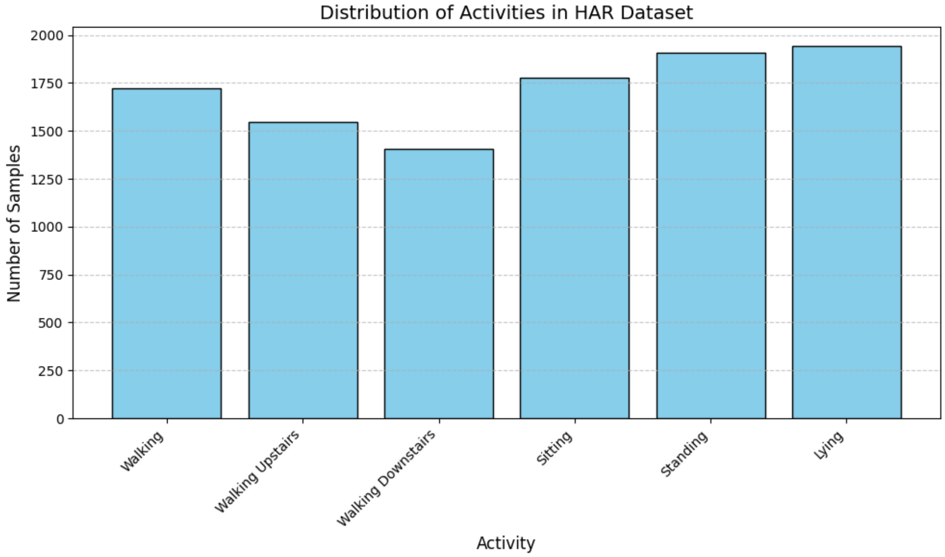

Figure 1: Distributions of Activities in HAR Dataset

The second learning pipeline employed in this study is centered around a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) layers, which are well-suited for sequential data. The input to the model has the shape (timesteps, features), corresponding to the 128 time steps and nine sensor features per sample. The model architecture includes an initial LSTM layer with 64 units, which outputs a sequence to enable further temporal processing. This is followed by a 50% dropout layer to prevent overfitting. A second LSTM layer with 32 units is used, this time outputting a single vector, which is passed through another dropout layer. The dense layers include a fully connected layer with 32 units and a ReLU activation, followed by a final dense layer with six units and a softmax activation to output the predicted probabilities for each activity class. The model was trained with the Adam optimizer, a learning algorithm that adapts the learning rate during training, and the sparse categorical cross-entropy loss function, which is ideal for multi-class classification tasks. Training was conducted over 50 epochs with a mini-batch size of 64 to balance training efficiency and model stability. LSTMs, the backbone of the architecture, employ a gating mechanism that allows them to retain important long-term dependencies in the sequence data while discarding irrelevant information. This design choice makes LSTMs particularly effective for recognizing patterns in time-series data like those in the HAR dataset.

The third learning pipeline to be implemented was a Convolutional Neural Network. The input needed to be of certain sizes so the testing and training data is reshaped. In addition, the predicted labels needed to be one hot encoded. These changes to the data were made accordingly. Multiple different layers are used to train the model. A Convo1D layer is used first because of the sequential nature of the data and to find patterns within small subsets of the data. Then, a MaxPooling1D layer is followed by a DropOut layer to reduce the parameters and to improve generalizability and prevent overfitting. Then, a second Convo1D layer is applied to learn the more subtle patterns. Then, we use a Flatten layer to prepare the model to be fed into two Dense layers, which have the model connect back to our original predictions and output the probabilities for our testing data. All hyperparameters are chosen based on the hope to balance predictable power and prevention of overfitting.

The fourth pipeline integrates both CNN and LSTM models into a hybrid architecture. This hybrid model combines the strengths of both approaches: the ability of CNNs to capture local spatial or temporal patterns and the LSTM's capability to retain long-term dependencies in sequence data. The input data is reshaped into three-dimensional arrays to match the requirements of Conv1D and LSTM layers. The model begins with a Conv1D layer containing 64 filters and a kernel size of 3, followed by a MaxPooling1D layer to downsample the features. Another Conv1D layer with 128 filters and a kernel size of 3 is added to capture more abstract features, followed by another MaxPooling1D layer. These CNN layers help in learning hierarchical features from the input data. The extracted features are then passed to an LSTM layer with 100 units to model temporal dependencies across the sequence.

To prevent overfitting, a dropout layer with a 50% rate is applied. The output from the LSTM layer is flattened and fed into dense layers, including a fully connected layer with 64 units and ReLU activation, followed by a final softmax layer with six units for multi-class classification. The model was compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. This hybrid approach leverages the ability of CNNs to reduce computational complexity and LSTMs to model temporal correlations, offering a comprehensive method for handling time-series data effectively.

5 Experiments/Results [1]

After completing all the models, the SVM emerged as the best-performing model, achieving an accuracy of 96.3% and a weighted F1 score of 96.29%, demonstrating its strong predictive power. This performance is likely due to the nature of the problem being a multi-class classification task. The SVM excelled in identifying data from individuals lying down but struggled with distinguishing between standing and sitting activities. Specifically, it misclassified 49 instances of sitting as standing, likely due to the similar body positions and lack of movement in these activities.

The CNN model also showed competitive performance, achieving an accuracy of 95.18% and an F1 score of 95.2%. It demonstrated perfect classification for sitting data but faced challenges in differentiating between sitting and standing, as well as between walking upstairs and downstairs. This performance suggests that while the CNN effectively captures certain patterns, further refinements may be necessary to enhance its ability to distinguish between similar activities.

The RNN achieved promising results with a training accuracy of 87.02% and a test accuracy of 91.00%, indicating its strong generalization capability to unseen data. A detailed analysis of the confusion matrix revealed excellent performance in classifying static activities such as lying down, where it correctly classified 527 out of 537 samples. However, it showed some confusion in differentiating dynamic activities, such as walking upstairs versus downstairs, likely due to the similar sensor signals generated by these activities. The weighted F1 score of 0.91 underscores the model's robust performance across all classes.

The hybrid CNN-LSTM model achieved a test accuracy of 91.18% and a weighted F1 score of 0.91. This shows its ability to balance feature extraction and temporal sequence modeling effectively. The confusion matrix highlights its strengths and challenges. The model performed exceptionally well in classifying static activities such as lying down, correctly predicting 478 out of 496 instances, and dynamic activities like walking, where it classified 444 out of 471 samples of walking downstairs. Despite its challenges categorizing climbing up and down the stairs, the hybrid model's ability to generalize across both static and dynamic activities is evident. Its performance aligns closely with that of the standalone RNN, leveraging the strengths of CNN layers to capture spatial patterns and LSTM layers to retain temporal dependencies, making it a robust option for time-series classification tasks.

To ensure a comprehensive evaluation, both accuracy and F1 scores were computed. Accuracy reflected the overall correctness of predictions, while the F1 score balanced precision and recall, offering a nuanced measure of performance. The confusion matrices provided qualitative insights into each model's strengths and areas for improvement, particularly in distinguishing between similar activities. Additional visualizations, such as loss and accuracy curves over epochs, could further aid in understanding model convergence and detecting potential overfitting.

6 Conclusion/Discussion

The study focuses on a Human Activity Recognition (HAR) dataset collected using wearable sensors embedded in smartphones. These sensors, typically including accelerometers, gyroscopes, and magnetometers, record motion and orientation data. The data was gathered while participants engaged in six common activities such as walking, sitting, standing, and laying down, in a controlled environment. This setup ensured consistent data recording across participants, reducing variability unrelated to the activities themselves. Since the study focuses on using an SVM as their training

[1] All of the code produced by the authors can be found at <https://github.com/merjemmm/HAR-MachineLearning>

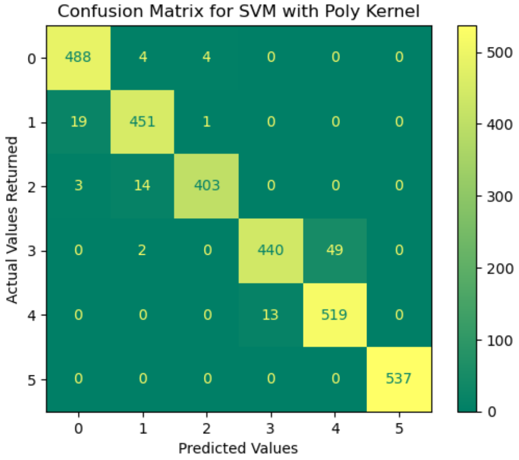

Figure 2: Confusion Matrix Plot for SVM Model

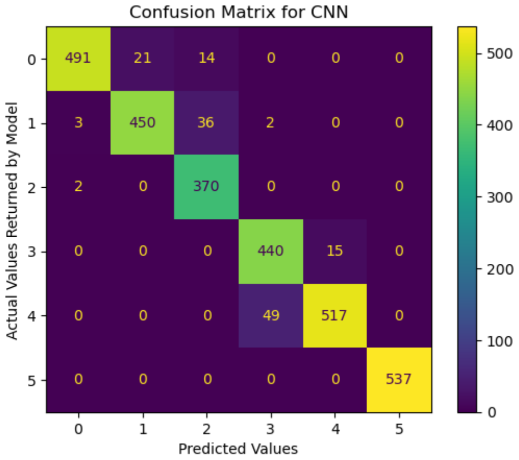

Figure 3: Confusion Matrix Plot for the CNN Model

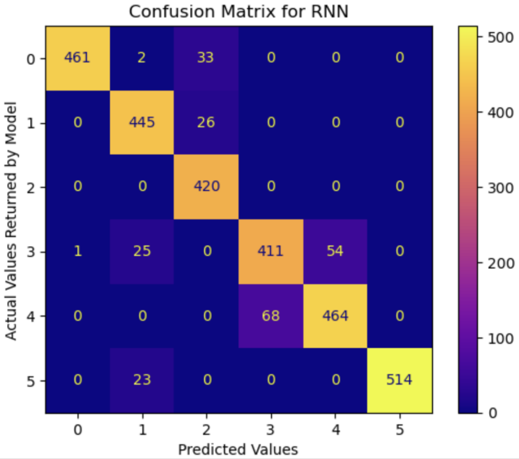

Figure 4: Confusion Matrix Plot for the RNN Model

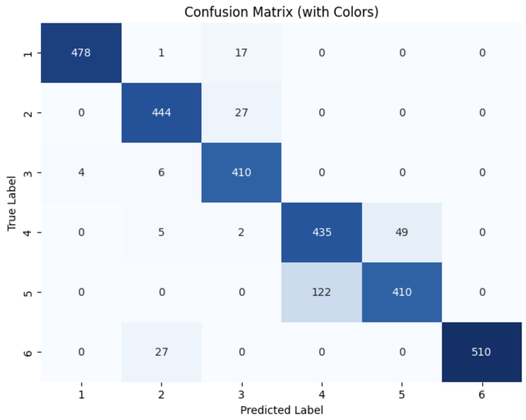

Figure 5: Confusion Matrix Plot for the Hybrid CNN-LSTM Model

model, we wanted to extend the use of the data to more modern Machine learning techniques and gauge the differences in performance.

Four machine learning models were trained and compared: a Support Vector Machine (SVM), a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN), and a hybrid model of CNN-LSTM. The SVM employed a polynomial kernel of degree 5 to map input features into a higher-dimensional space, enhancing separability. A regularization parameter C of 10 was used, indicating a low tolerance for misclassification errors and a focus on tightly fitting the training data. The decisions for the final layers of the RNN and CNN were based on common practice, the aspects of our data, and the resulting accuracy scores. The hybrid CNN-LSTM combined the spatial pattern detection of CNNs with the sequential modeling capabilities of LSTMs, aiming to leverage the strengths of both architectures for improved activity recognition.

The CNN, known for its ability to detect spatial hierarchies in data, was used to identify patterns in short time windows of the sensor data. Similarly, the RNN, designed for processing sequential data, attempted to model the temporal dependencies within the dataset. The CNN-LSTM ocmbined model tried to take the best of both worlds. However, these neural network-based models might have been disadvantaged by factors such as the dataset's size or limited temporal complexity.

The SVM outperformed both all of the other models. This could be due to several reasons. Firstly, the HAR UCI dataset was relatively small and controlled, which favored SVM's simplicity and robustness. Unlike CNNs and RNNs, which require large datasets to generalize effectively, SVMs excel in smaller datasets by focusing on maximizing the margin between classes, ensuring better generalization without overfitting. The dataset's features, which were either pre-engineered or inherently distinguishable, aligned well with SVM's strengths. For example, activities like "laying down" likely had distinct sensor readings that the SVM could separate efficiently with its polynomial kernel. In contrast, CNNs, RNNs, and the hybrid model of CNN-LSTM, designed for complex tasks like extracting spatial or temporal patterns, might have been unnecessarily complicated for this task, leading to overfitting due to their high capacity. Additionally, SVMs inherently handle activity classification tasks well, as their margin-maximizing approach makes them powerful for distinguishing between classes with non-linear boundaries. Overall, the controlled nature of the dataset, the alignment of its features with SVM capabilities, and SVM's robustness against overfitting contributed to its superior performance compared to the more complex neural network models.

Interestingly, all models performed exceptionally well in identifying the activity of "laying down." This suggests the sensor readings for this activity, characterized by static accelerometer data and a lack of movement, were distinct and easily separable from those of other activities. However, the models struggled to differentiate between activities involving similar body movements, such as walking on an incline versus walking up or down stairs. These activities likely produced overlapping sensor patterns, making them more challenging to distinguish.

Given the SVM's strong performance, it appears to be the most promising model for further refinement. Future improvements could involve more advanced feature engineering, such as extracting higherorder derivatives, spectral features, or additional time-domain statistics to better capture activity nuances. Hyperparameter optimization, exploring alternative kernel functions, polynomial degrees, or regularization values, could further enhance the SVM's accuracy.

Because of the amount of data, there are limitations to the generalizability of the findings. There are many overlapping features that were collected because of the complexity of sensors and the three-dimensional aspects of human activities, which resulted in measurements for each x, y, and z directions. More variation in the the types of features might improve predictions and the learning capabilities of the models. Furthermore, each person is unique and having a pool of only 30 people puts limits on our ability to claim our models would perform strongly on larger amounts of data. In addition, for privacy and anonymity, there were no details provided about the people involved in the pool. Their fitness levels and physical characteristics would play a major role in the performance of these activities.

Dataset augmentation, such as collecting data in uncontrolled environments, could improve the model's robustness and generalizability for future studies. Additionally, hybrid approaches combining feature extraction through CNN layers with SVM classification might harness the strengths of both methods. While the SVM already excels at distinguishing certain activities, these enhancements could improve its performance on more ambiguous tasks, such as differentiating between similar body movements.

7 Contributions

Clarissa took responsibility for implementing the RNN model the hybrid CNN-LSTM model. Merjem focused on implementing the SVM and CNN models. Both team members contributed detailed explanations in the paper for the models they were responsible for, providing clarity and depth to their respective sections.

8 References

[1] Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, 108, 107561.

[2] Dua, N., Singh, S. N., & Semwal, V. B. (2021). Multi-input CNN-GRU based human activity recognition using wearable sensors. Computing, 103(7), 1461-1478.

[3] Gupta, S. (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. International Journal of Information Management Data Insights, 1(2), 100046.

[4] Kusuma, W. A., Minarno, A. E., & Safitri, N. D. N. (2022, July). Human activity recognition utilizing SVM algorithm with gridsearch. In AIP Conference Proceedings (Vol. 2453, No. 1). AIP Publishing.

[5] Khatun, M. A., Yousuf, M. A., Ahmed, S., Uddin, M. Z., Alyami, S. A., Al-Ashhab, S., ... & Moni, M. A. (2022). Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor. IEEE Journal of Translational Engineering in Health and Medicine, 10, 1-16.

[6] Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. IEEE communications surveys & tutorials, 15(3), 1192-1209.

[7] Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., & Parra, X. (2013). Human Activity Recognition Using Smartphones [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C54S4K.

[8] Stisen, A., Blunck, H., Bhattacharya, S., Prentow, T. S., Kjærgaard, M. B., Dey, A., ... & Jensen, M. M. (2015, November). Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In Proceedings of the 13th ACM conference on embedded networked sensor systems (pp. 127-140).

[9] Vettoruzzo, Anna, Mohamed-Rafik Bouguelia, and Thorsteinn Rögnvaldsson. "Efficient Few-Shot Human Activity Recognition Via Meta-Learning and Data Augmentation." Available at SSRN 4980432.

[10] Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. mobile networks and applications, 25(2), 743-755.

[11] Zhu, S., Guendel, R. G., Yarovoy, A., & Fioranelli, F. (2022). Continuous human activity recognition with distributed radar sensor networks and CNN–RNN architectures. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-15.

[12] Mutegeki, R., & Han, D. S. (2020, February). A CNN-LSTM approach to human activity recognition. In 2020 international conference on artificial intelligence in information and communication (ICAIIC) (pp. 362-366). IEEE.