ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Unveiling the Power of Ensemble Learning with Multiple Convolutional Neural Networks for Human Activity Recognition

[1] Azim, [2] P. Mrunalini Rupa,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 16-06-2025 Accepted: 28-06-2025

Abstract—

A subfield of machine learning known as "Human Activity Recognition" focuses on identifying specific human actions from sensor data, particularly one-dimensional time series data. In the past, activity detection machine learning models were built using characteristics that were manually created. But that's no easy achievement; it calls for extensive knowledge of the topic as well as feature engineering. Models can now automatically learn characteristics from raw sensor data, making it more simpler and leading to better classification results, all thanks to deep neural networks. Ensemble learning of several convolutional neural network (CNN) models is introduced in this research as a new method for human activity identification. Using the freely accessible dataset, we train three separate CNN models and then construct several ensembles of these models. Outperforming approaches found in the literature, the combined results of the first two models achieve an accuracy of 94%.

Index Terms—

Convolutional neural networks, deep learning, ensemble learning, activity recognition in humans.

I. INTRODUCTION

A wide range of industries and purposes have found uses for Human Activity Recognition (HAR), such as security, automated surveillance, and smart healthcare systems. Firstly, unlike video sensors, which might compromise users' privacy, on-body sensors like gyroscopes, proximity sensors, magnetometers, temperature sensors, etc., are ideal for HAR applications. Second, a whole-body sensor network (BSN) enables more precise signal collection system deployment, and third, they lessen the constraints imposed by the surrounding environment and fixed camera locations, similar to video-based datasets. HAR has defined an essential role in ubiquitous computing [1] because to the streamlined data collection procedure made possible by embedded sensors, particularly with the proliferation of smartphones over the last decade. Hence, smartphones may effectively function as an array of sensors worn by the body to gather data. Feature extraction from these gathered data points used to be a laborious, task-dependent, and heuristic procedure. Feature extraction has traditionally relied on features that are purpose-built and application-specifically chosen. In order to prepare the raw signals for classification algorithms, statistical metrics like variance and mean, as well as transform coding metrics like Fourier transforms, were retrieved. One drawback of these approaches is that they are only useful for certain classification tasks. Another issue with feature selection by hand is the potential loss of raw signal information [2]. The availability of more data and more computing power, together with developments in deep learning, have greatly facilitated the acceleration and enhancement of this process. Because deep learning models can automatically extract features and apply them to various classification tasks, the feature selection process is no longer task-dependent. Maximizing the

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

utilization of this cutting-edge research for focused HAR implementation is a natural flow with the availability of deep learning. Consequently, this work's objective is to provide a methodical procedure for a more simplified feature representation in order to enhance activity recognition by means of ensemble learning of convolutional neural networks (CNNs) [3]. Convolutional neural networks are heavily used in the current HAR research. By using its unique unification layer to combine feature maps at the conclusion, Deep CNN in [4] is able to adjust automatic feature learning from raw inputs. In [5], the author outlines a less complex CNN-based method that uses three distinct subsets of the same dataset to train the model. The feature extraction strategy in [2] makes use of a long short-term memory (LSTM) network and a softmax classifier since LSTMs are good at repeated jobs. Due to the one-dimensional nature of the dataset, RNN-LSTM has also been investigated for this application. This architecture is comparable to RNN, which has shown the greatest promise in natural language processing (NLP) applications [6]. To further enhance performance, HAR has also been used in [7] using hybrid models of CNN and RNN-LSTM. Because of their superior ability to infer time-order correlations, RNNs and LSTMs work together in a hybrid model to better identify short-range activities.inter-sensor measurements. Due to their superior feature learning in recursive patterns, CNNs are better able to infer repeated, long-term behaviors. Using deep belief networks (DBNs) via restricted Boltzmann Machines (RBMs) to prevent the model from overfitting and keep training time consumption low is an intriguing method to HAR [8]. As an alternative, [9] outlines a hybrid method for sequential human activity identification. As a feedforward network with pre-selected hidden layers and nodes that don't need adjusting, an extreme learning machine (ELM) is investigated. A combined CNN-LSTM architecture incorporates the ELM as a feature extractor and uses it as a classifier. Effective sequential activity recognition is achieved by combining convolutional procedures with an LSTM for recurrent units. In reference 10, we meet yet another LSTM-CNN design for HAR. This design incorporates convolutional layers after a two-layer long short-term memory (LSTM) that processes raw data acquired by mobile sensors. In order to further reduce the number of parameters in the model, a global average pooling layer (GAP) is used. After the GAP layer, a batch normalization layer (BN) is added to further accelerate convergence. Several HAR datasets show that the model performs well in evaluations. Many deep learning algorithms are associated with increasing computational complexity and resource overheads; [11] presents a method that accounts for this. This method makes HAR easier to handle and more efficient on mobile devices that don't have a lot of processing power. Using temporal convolutions on the spectrogram domain of the data, the technique described in the study extracts features from various portions of the raw signal. The result is that the learnt traits are resistant to changes in a wide range of parameters. Many variables determine whether current approaches for HAR are effective. Traditional deep learning methods suffer greatly from overfitting and lengthy model training times. Also, when compared to HAR models made with devices like mobile phones in mind, which have lesser computational complexity, these models cause problems. In order to predict actions on the 1D HAR dataset, we provide a method that involves building three convolutional neural networks (CNNs) with 1D convolutional layers. One spatial or temporal dimension is all it takes to build a convolution kernel with the layer input in the 1D convolution layer. After that, it takes an average of the three models and uses it to create an ensemble model. As a multiple classifier system, ensemble learning [12] boosts model performance and outperforms individual models. We continue by generating further pairs of models, such as first and second, second and third, and first and third. Overall, the model's performance has been enhanced via ensemble learning. Here is the breakdown of the remaining sections of the paper: Section II provides information on the dataset, Section III explains the process, and Section IV shows and talks about the results findings, and the article is wrapped up in Section V.

II. DATASET

Human Activity Recognition ensemble learning was applied using a dataset supplied by Wireless Sensor Data Mining (WISDM) Lab [13]. All all, 36 people, each armed with a smartphone, contributed to this dataset. The integrated tri-axial accelerometer (x-, y-, and z-axis) is the data-collecting sensor. Twenty samples per second is the sampling frequency. Six categories representing common everyday tasks make up the dataset. Here is the breakdown of the classes:

- Walking: 424,400 examples, 38.6%,
- Jogging: 342,177 examples, 31.2%,
- Upstairs: 122,869 examples, 11.2%,
- Downstairs: 100,427 examples, 9.1%,
- Sitting: 59,939 examples, 5.5%,
- Standing: 48,395 examples, 4.4%

This class distribution is plotted as shown in Figure 1.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

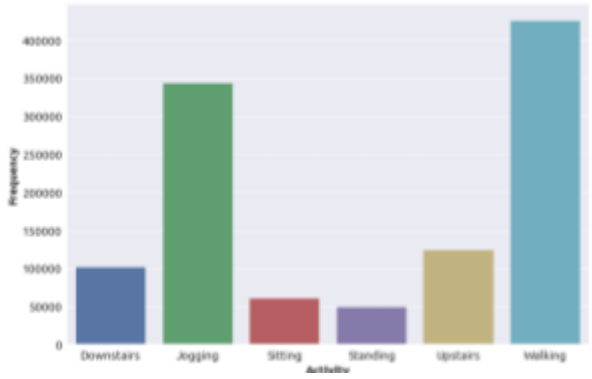

Fig. 1. Class distribution for WISDM dataset

As a first step, we normalize the data by pre-processing it to have a zero-mean axis. One activity's three-axis graph is, as seen in Figure 2.

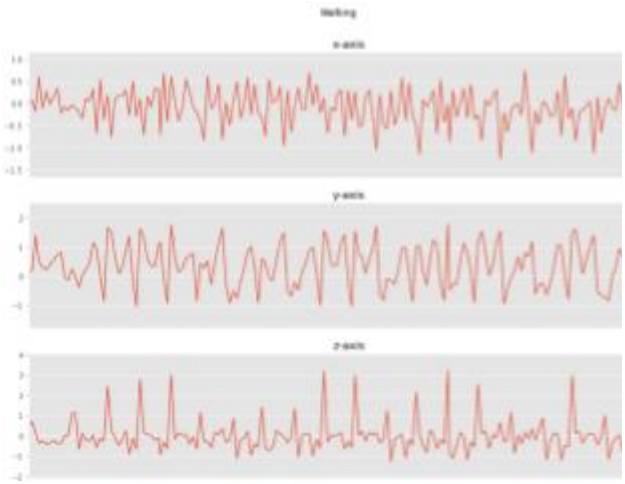

Fig. 2. Tri-axial data for a single class (walking)

Following the pre-processing, we modify the data by segmenting and shaping it to fit the 1D convolutional layers of the CNNs that are currently in use. We begin by creating indexes of a fixed-size window and then advance by half of the size of the window in order to split the data. Our final dataset will have dimensions like [total segments, input width and input channel] (in this example, [24403, 90, 3]) after we build segments of fixed size and add the windowed data.

III. METHODOLOGY

Our work in this study includes the development of three separate CNN-based models, the basis for which we built a number of ensemble learning models. In ensemble learning, several models or learners are taught to address a single issue. This paradigm's strength is in its generalizability. It may enhance performance output by enhancing the learning impact of poorer learners. The models' features and an overview are detailed in the following sections. First Model (A) The ConvPool-CNN-C architecture described in [14] served as an inspiration for the first model. It uses a typical pattern of max pooling and convolution layers back-to-back. In this model, the convolutional layers were activated using ReLU functions, and the max pooling and 2D convolutions were converted into 1D operations. To further decrease the propensity to overfit the training data, an extra feature is added: dropout regularization is used after each convolution filter. When compared to only using convolution and max pooling layers, this approach produces a better response. This model also makes use of fully linked layers, as opposed to ConvPool-CNN-C's global average pooling. The model's last fully connected layer uses a tanh

activation to normalize, and then, in the last dense layer, a softmax function to optimize for the current classification situation. The model's architecture is shown in Figure 3. B. Adjacent Model A variant of the ALL-CNN-C, which is referenced in [14], is used as the second model. The term, ALL-CNN, comes from the fact that convolutional layers are used instead of max pooling layers, which is the fundamental difference between this model and the first. Instead of using max pooling layers, this model made use of many 1D convolutional layers activated by ReLU with stride 2. Model 2's construction is seen in Figure 4. The Third Model (C) The 'network in a network' model from [15] served as an inspiration for the third model. This model minimizes training time and optimization parameters by using a succession of convolution layers with 1x1 kernels. Additionally, dropout regularization was used to enhance precision. Figure 5 displays the model's architecture.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

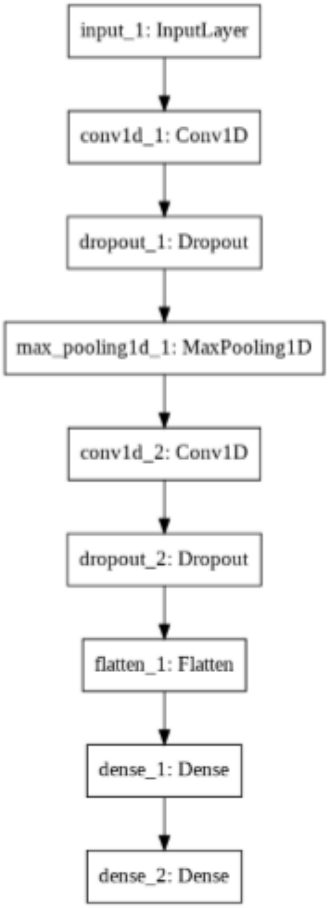

Fig. 3. Model 1, inspired from the ConvPool-CNN-C architecture. It utilizes alternating 1D convolution and max pooling layers with a fully connected layer.

D. A combination of three models Next, we combined the three models into one ensemble. At the end, it takes the three models' outputs and averages them to get a mean output. The input layer is the same for all three models. While there are other ways to construct an ensemble model, we opted for a stacking ensemble pattern that averages the results from all of the convolutional neural network (CNN) architectures used. No training is necessary for the ensemble model; all training is done for the individual models. The three convolutional neural network ensemble model is shown in Figure 6. The three convolutional neural network (CNN) models were not only combined into an ensemble, but they were also coupled to create pairwise ensemble models.

IV. RESULTS AND DISCUSSION

Using the (WISDM) dataset, we trained and evaluated our models and ensembles [13]. Training accuracy and loss for all three CNN models are shown in Figures 7 and 8, respectively. After about 5 epochs, the accuracies surpassed 90% and, by the end of the 50 epochs, they were approaching 99%. In contrast to the training process, the validation accuracies ranged from 70% to 80%, with a greater amount of variance noted. The ensemble model achieved much greater training accuracies and improved overall performance.

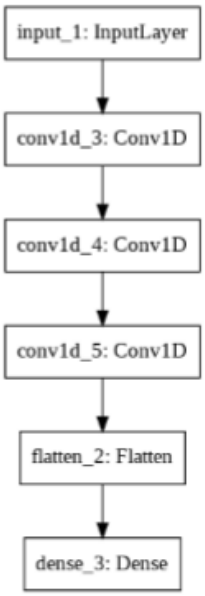

Fig. 4. Model 2, adopted from the ALL-CNN-C architecture. This model utilizes all convolution layers and does not use any max pooling layers.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

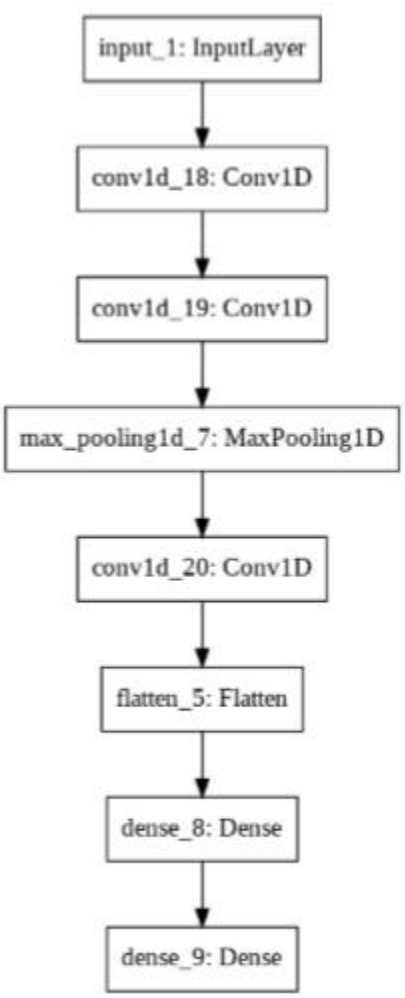

Fig. 5. Model 3, partially inspired by the Network in a Network (NIN) model. It utilizes a series of 1D convolution layers, a max pooling layer and fully connected layers.

preciseness of assessment. The assessment accuracies of the CNN models compared to the ensemble models are summarized in Table 1. Table 1 shows that, compared to other models, ensemble models often result in more accurate evaluations.

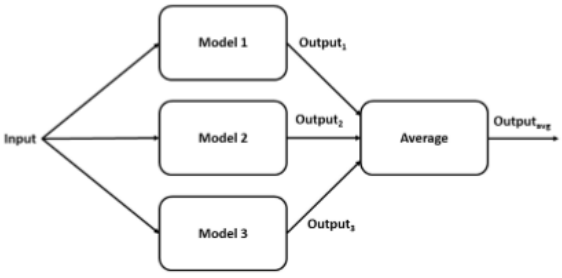

Fig. 6. Ensemble Model

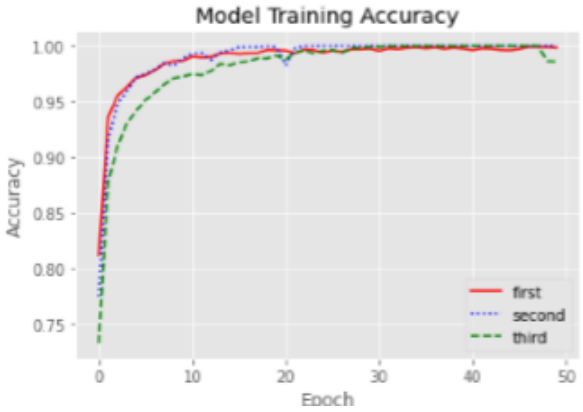

Fig. 7. Training Accuracies for Individual Models

specifically designed convolutional neural network (CNN) models. Take note of the outcomes when the third model is part of the ensemble and its accuracy is raised (approx. 93%) as opposed to the third model's 90% accuracy when used alone. In the literature, many non-ensemble designs were found, each producing its own set of gains and disadvantages. After 200 iterations, the LSTM-RNN HAR model described in [6] achieved an accuracy of 90%. A maximum recognition rate of 92.5% is achieved using the straightforward CNN-based method described in [5]. An 82% success rate was achieved using the multiclass SVM method that was discussed in reference [8]. Achieving an accuracy of 93.75% was the convnet plus multilayer perceptron based on inverted pyramid design described in [16]. The accuracy of our paired ensemble model, Paired Ensemble (1,2), was over 94% when compared to various methods. Reduced evaluation losses were a common result of the constructed ensemble models.

TABLE I MODEL ACCURACIES

Model	Accuracy (%)
First	93.08
Second	93.55
Third	90.93
3-Model Ensemble	93.66
Paired Ensemble (1,2)	94.01
Paired Ensemble (2,3)	93.04
Paired Ensemble (1,3)	93.08

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

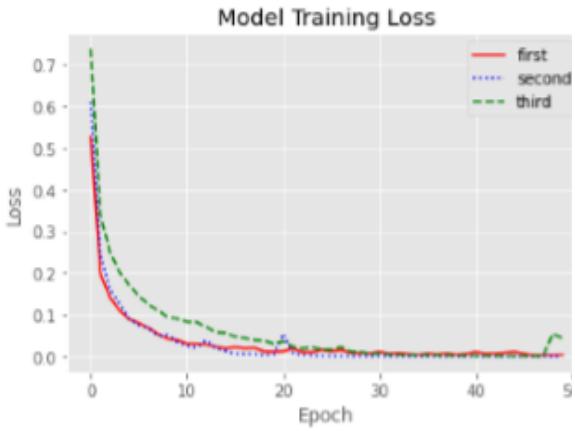

Fig. 8. Training Losses for Individual Models as compared to the individual deep learning models developed.

V. CONCLUSION

We showcased three convolutional neural network (CNN) models and their ensembles on the WISDM dataset of HAR. The results showed that compared to the individual models, the ensemble model performed better. Compared to the approaches described in the literature, one ensemble model outperformed the others. We see a class imbalance in the dataset we used; for example, 38% of the samples are from the walking class, whereas practically 5% are from the sitting and standing classes. If we can eliminate the class imbalance from the dataset in the future, the findings could be much better. A possible area for further investigation may be to do weighted ensemble learning such that the best performing model has the greatest influence in the ensemble. Currently, an ensemble is produced by averaging the three models. In addition, we may go into the realm of hybrid model ensemble learning, which entails combining CNN and RNN models.

REFERENCES

[1] T. Pl¨otz, N. Y. Hammerla, and P. L. Olivier, “Feature learning for activity recognition in ubiquitous computing,” in 22nd International Joint Conference on Artificial Intelligence, 2011.

[2] Y. Chen, K. Zhong, J. Zhang, Q. Sun, and X. Zhao, “LSTM networks for mobile human activity recognition,” in International Conference on Artificial Intelligence: Technologies and Applications (ICAITA 2016), 2016.

[3] L. K. Hansen and P. Salamon, “Neural network ensembles,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993–1001, 1990.

[4] J. 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 24th International Joint Conference on Artificial Intelligence, 2015.

[5] M. Panwar, S. R. Dyuthi, K. C. Prakash, D. Biswas, A. Acharyya, K. Maharatna, A. Gautam, and G. R. Naik, “CNN based approach for activity recognition using a wrist-worn accelerometer,” in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017, pp. 2438–2441.

[6] S. W. Pienaar and R. Malekian, “Human activity recognition using LSTM-RNN deep neural network architecture,” in 2019 IEEE 2nd Wireless Africa Conference (WAC). IEEE, 2019, pp. 1–5.

[7]J.Wang,Y.Chen,S.Hao,X.Peng,andL.Hu,“Deeplearningforsensorbased activity recognition: A survey,” Pattern Recognition Letters, vol. 119, pp. 3–11, 2019.

[8] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, “A robust human activity recognition system using smartphone sensors and deep learning,” Future Generation Computer Systems, vol. 81, pp. 307–313, 2018.

[9] J. Sun, Y. Fu, S. Li, J. He, C. Xu, and L. Tan, “Sequential human activity recognition based on deep convolutional network and extreme learning machine using wearable sensors,” Journal of Sensors, vol. 2018, pp. 1–10, 09 2018.

[10] K. Xia, J. Huang, and H. Wang, “LSTM-CNN Architecture for Human Activity Recognition,” IEEE Access, vol. 8, pp. 56855–56866, 2020.

[11] D. Ravi, C. Wong, B. Lo, and G. Yang, “Deep learning for human activity recognition: A resource efficient implementation on low-power devices,” in 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2016, pp. 71–76.

[12]M.Wo´zniak,M.Gra˜na,andE.Corchado,“Asurvey ofmultipleclassifier systems as hybrid systems,” Information Fusion, vol. 16, pp. 3–17, 2014.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

[13] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, vol. 12, no. 2, pp. 74–82, 2011.

[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," arXiv preprint arXiv:1412.6806, 2014.

[15] M. Lin, Q. Chen, and S. Yan, "Network in Network," arXiv preprint arXiv:1312.4400, 2013.