Machine Learning Model Comparison for WiFi CSI Activity Classification

Samuel Mclean

September 19, 2024

1 Introduction

Machine learning has proven to be a transformative tool in the analysis of WiFi Channel State Information (CSI), enabling more advanced and accurate classification of human activities based on wireless signals. By capturing the detailed characteristics of signal fluctuations caused by movements and environmental changes, machine learning models can be trained to recognize and differentiate various activities within indoor environments.

In the "Guardian Monitor" project, machine learning techniques are employed to classify patient activities based on CSI data, allowing for the development of non-intrusive, real-time monitoring systems. Different models, such as decision trees, support vector machines (SVMs), and deep learning architectures, can be trained on labeled datasets to predict patient actions like walking, sitting, or falling. The project aims to evaluate and compare the performance of these models in terms of accuracy, computational efficiency, and their ability to generalize across different environments and patients.

This approach to activity classification using WiFi CSI and machine learning offers a powerful solution for healthcare settings, where continuous and passive monitoring can greatly improve patient care and safety.

2 Machine Learning Algorithms Research

2.1 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges them for a more accurate and stable prediction.

Pros:

- Handles large datasets and high dimensionality.
- Reduces overfitting compared to individual decision trees.
- Can handle both classification and regression problems.

Cons:

- Computationally expensive, especially with large forests.
- Less interpretable than single decision trees.

Time Complexity:

- Training: $O(n \cdot m \cdot \log(n))$, where n is the number of samples and m is the number of trees.
- Testing: $O(m \cdot \log(n))$.

2.2 Support Vector Machine (SVM)

SVM constructs hyperplanes in a high-dimensional space to classify data points.

Pros:

- Effective in high-dimensional spaces.
- Works well when there is a clear margin of separation.

Cons:

- Not suitable for large datasets due to high training time.
- Performance depends heavily on the choice of kernel.

Time Complexity:

- Training: $O(n^2 \cdot d)$, where n is the number of samples and d is the dimensionality of the data.
- Testing: $O(n \cdot d)$.

2.3 Bidirectional Long Short-Term Memory (biLSTM)

biLSTM is a recurrent neural network (RNN) architecture that can capture long-term dependencies in sequential data by processing the input in both forward and backward directions.

Pros:

- Effective for sequential data like time series or WiFi CSI signals.
- Can capture long-range dependencies in data.

Cons:

- High computational cost and memory requirements.
- Training can be slow and require a large dataset.

Time Complexity:

- Training: $O(n \cdot T \cdot d^2)$, where n is the number of samples, T is the sequence length, and d is the dimensionality of the data.
- Testing: $O(T \cdot d^2)$.

2.4 Long Short-Term Memory (LSTM)

LSTM is another type of RNN designed to remember long-term dependencies and mitigate the vanishing gradient problem.

Pros:

- Excellent for capturing temporal dependencies.
- Useful for sequential data processing and time-series classification.

Cons:

- Requires more training data than traditional machine learning algorithms.
- Computationally intensive and prone to overfitting with small datasets.

Time Complexity:

- Training: $O(n \cdot T \cdot d^2)$.
- Testing: $O(T \cdot d^2)$.

2.5 Convolutional Neural Network (CNN)

CNNs are primarily used for spatial data, such as images, but can also be applied to CSI data for feature extraction.

Pros:

- Efficient at feature extraction from spatial data.
- Performs well with complex datasets and large input sizes.

Cons:

- Requires a large amount of training data.
- Computationally expensive to train.

Time Complexity:

- Training: $O(n \cdot d^2 \cdot f)$, where n is the number of samples, d is the dimensionality of the input, and f is the number of filters.
- Testing: $O(d^2 \cdot f)$.

3 Summary

Each of these machine learning algorithms offers unique advantages and drawbacks for classifying activities based on WiFi CSI data. For instance, Random Forests provide a balance between accuracy and interpretability but can become computationally expensive. SVMs perform well with smaller datasets but struggle with scalability. LSTM and biLSTM are well-suited for sequential data

but require more computational resources. CNNs, while powerful in extracting features, also demand large datasets and high computational power.

Below is a table summarizing the training and testing time complexities of each algorithm:

Algorithm	**Training Time Complexity**	**Testing Time Complexity**
Random Forest	$O(n \cdot m \cdot \log(n))$	$O(m \cdot \log(n))$
SVM	$O(n^2 \cdot d)$	$O(n \cdot d)$
biLSTM	$O(n \cdot T \cdot d^2)$	$O(T \cdot d^2)$
LSTM	$O(n \cdot T \cdot d^2)$	$O(T \cdot d^2)$
CNN	$O(n \cdot d^2 \cdot f)$	$O(d^2 \cdot f)$

Table 1: Training and Testing Time Complexities of Machine Learning Algorithms

3.1 Hypothetically most suitible algorithm for WiFi CSI

Overall, for our specific task, **BiLSTM**, **LSTM** and **CNN** seem promising. **BiLSTM** is likely to perform well if sequential data is important, otherwise the slow computation may not be worth the effort. The **LSTM** is also able to capture sequential data but as it does not do a forward and backward pass so it will compute faster but may be able to capture the features as well. Finally **CNN** are very good at feature extraction, which may be critical for sifting through the noisy data. The other models are less likely to perform as well in this context. **SVM** as it typically performs worse on noisy data like our datasets. **RF** also typically performs better on more structured data.

3.2 Hypothetically Most Suitable Algorithm for WiFi CSI

Overall, for our specific task, **biLSTM**, **LSTM**, and **CNN** seem promising. **biLSTM** is likely to perform well if sequential data is important, as it captures temporal dependencies in both forward and backward directions. However, the computational cost and longer training time may not justify the benefits if the dataset lacks strong sequential patterns. The **LSTM** can also handle sequential data, but with only a single pass (forward), which makes it faster to compute than biLSTM. Despite its faster processing, LSTM might not capture certain features as effectively as biLSTM when working with complex, noisy WiFi CSI data.

CNNs excel at feature extraction, making them ideal for identifying patterns in noisy data, which is often the case with WiFi CSI signals. CNNs' ability to focus on local features within the data makes them suitable when signal quality varies, or when specific features need to be extracted from the signal. However, CNNs require a large dataset to train effectively and are computationally expensive.

SVM and **Random Forest (RF)** are less likely to perform as well in this context. SVM struggles with noisy data, which is common in WiFi CSI, as the

algorithm is sensitive to outliers and signal distortions. Similarly, RF tends to perform better on structured datasets and may struggle to generalize well on WiFi CSI data, which can have high variability and noise.

3.3 Training Complexity and Sample Requirements

Training complexity and the amount of data required are critical factors when choosing models for WiFi CSI activity prediction.

biLSTM and **LSTM** models are computationally intensive due to the need to capture long-term dependencies in sequential data. They require large datasets to avoid overfitting and perform well, but their training times are significantly longer due to backpropagation through time (BPTT).

CNNs also demand large datasets, as their strength lies in extracting features from noisy data. While CNNs are computationally expensive to train, they are generally faster during inference compared to biLSTM and LSTM models.

In contrast, **SVM** and **Random Forest (RF)** are easier and faster to train, requiring fewer data samples. However, they are less suited to handling noisy, unstructured data like WiFi CSI. SVM tends to struggle with large, noisy datasets, while RF performs better with structured data but is less effective in dynamic environments.

In summary, deep learning models (biLSTM, LSTM, CNN) are more suitable for WiFi CSI due to their ability to handle sequential and spatial data but come with higher training costs and data requirements. SVM and RF offer faster training but are less effective in this context.

Algorithm	Training Complexity	Amount of Data Required
biLSTM	High	Large
LSTM	Moderate-High	Large
CNN	High	Large
SVM	Moderate	Small-Medium
Random Forest	Moderate	Small-Medium

Table 2: Training Complexity and Data Requirements of Machine Learning Algorithms

4 Practical Performance of Algorithms

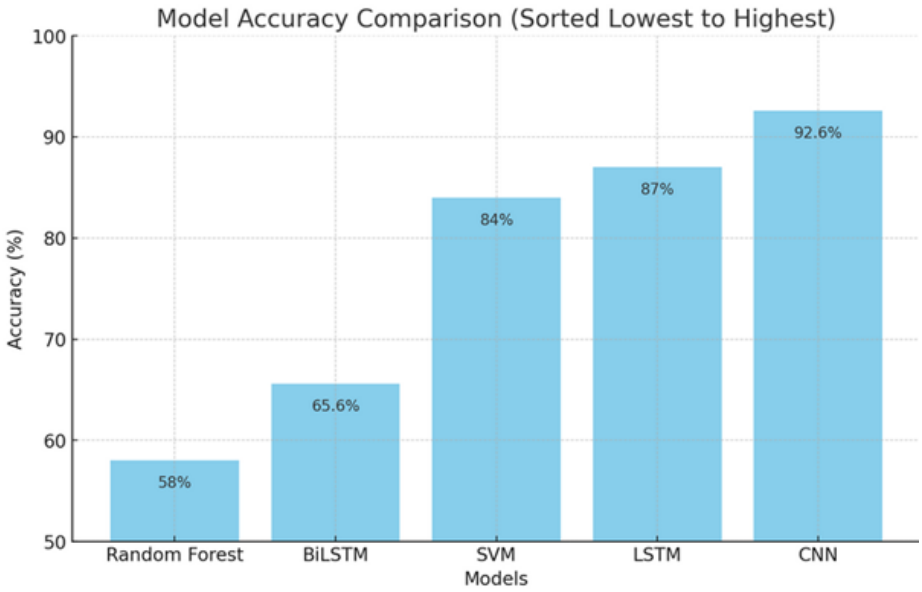
The practical performance results show a clear distinction between the machine learning algorithms in terms of accuracy. **Random Forest (RF)** performed the lowest with 58%, likely due to its inability to handle the noisy, unstructured nature of WiFi CSI data. Similarly, while **biLSTM** performed slightly better with 66%, its higher computational demands might have limited its capacity to generalize well in this case.

SVM achieved a higher accuracy of 84%, performing better than RF and biLSTM, possibly because of its ability to find optimal margins for classification.

However, its susceptibility to noise might explain why it didn't outperform the deep learning models.

LSTM and **CNN** provided the best results, with 87% and 92.6% accuracy, respectively. This is likely due to their ability to capture temporal and spatial features effectively. **CNN's** high performance can be attributed to its strength in feature extraction from complex data, making it particularly well-suited for dealing with the variability in WiFi CSI signals.

In summary, deep learning models (LSTM and CNN) showed superior performance due to their ability to handle the complexity and noise in WiFi CSI data, while simpler models like RF struggled with the unstructured nature of the signals.



5 Discussion

While models like **biLSTM**, **SVM**, and **Random Forest (RF)** did not achieve the highest accuracies in this task, they still offer certain advantages that can make them useful depending on the specific context.

biLSTM may not have performed as well here, but it remains valuable for tasks where bidirectional temporal dependencies are critical. In scenarios where capturing both past and future context from the WiFi CSI data is important, biLSTM could offer more detailed insights compared to simpler models.

SVM, though sensitive to noise, excels in smaller datasets and high-dimensional spaces. In cases where the data is cleaner or less noisy, SVM could offer strong

performance, especially when computational resources are limited, as it is less resource-intensive than deep learning models.

Random Forest also has its strengths. It is highly interpretable, which can be advantageous when explainability is important. RF is also fast to train and performs well on structured data, so in situations where the WiFi CSI signals are less variable or noisy, it could still be an effective choice. Additionally, its ability to handle both classification and regression tasks makes it versatile in different applications.

In summary, although these models did not perform as well in this specific experiment, their unique strengths—such as interpretability, speed, and handling of certain data structures—may make them suitable for other scenarios or different types of WiFi CSI data.

5.0.1 Low Reliability of Results

The results presented here were derived from individual efforts in optimizing hyperparameters and preprocessing data for each algorithm. This likely introduced inconsistencies, as some individuals may have been more successful in fine-tuning their respective algorithms. These variations could have impacted the performance outcomes. To ensure more reliable and consistent results, it would be beneficial to reassess the algorithms in a more controlled and standardized experiment. This reassessment should also include comparisons of training and testing times to evaluate not only accuracy but also the computational efficiency of each model.

6 Conclusion

Based on both current research and the practical results of this experiment, **CNN** emerged as the most effective model for WiFi CSI activity classification. Its ability to extract spatial features from noisy data makes it well-suited to this task, providing the highest accuracy among the tested models. However, further research is warranted to more comprehensively evaluate the suitability of different algorithms. Future studies should focus on not only accuracy but also the computational resources required by each model, such as training and testing times. Additionally, conducting a more controlled and standardized assessment will help ensure consistency across experiments, allowing for a clearer comparison of each algorithm's performance.