Electronics, Communications, and Computing Summit
Vol. 2, No. 3, Jul - Sep 2024, pp. 40-48
DOI: https://doi.org/10.17051/ECC/02.03.05

Comparative Analysis of AI Models for Channel Estimation in mmWave Massive MIMO Systems for 6G

N. Kazi[1], Dr. László Tótha[2]

[1]University Of Malaya, Malaysia, Email: Salimnewaz@um.edu.my
[2]Budapest Center for Digital Societies, Hungary.

Article Info

Article history:

Received : 15.07.2024
Revised : 17.08.2024
Accepted : 19.09.2024

Keywords:

AI-based Channel Estimation,
mmWave Communication,
Massive MIMO,
6G Wireless Networks,
Deep Neural Networks (DNN),
Convolutional Neural Networks (CNN),
Recurrent Neural Networks (RNN),
Transformer Architecture,
Normalized Mean Square Error (NMSE),
Line-of-Sight (LOS),
Non-Line-of-Sight (NLOS),
Real-Time Estimation,
Intelligent Wireless Systems.

ABSTRACT

The combination of both the millimeter-wave (mmWave) communication and massive multiple-input multiple-output (MIMO) technology is the benchmark of the next-generation 6G wireless networks, promising to realize unprecedented data rates, barrier-free low latency, and improved spectral efficiency. The mmWave massive MIMO channels, however, have a high-dimensional, sparse, and highly dynamic characteristics that makes their accurate and efficient acquisition beyond challenging. Some classical methods of model-based estimation like Least Squares (LS) and Minimum Mean Square Error (MMSE) frequently cannot satisfy the performance requirements of 6G because they are based on simplified assumptions and they require excessive pilot overhead. To react to this, the given study provides a comparative analysis of modern artificial intelligence (AI) modelsDeep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based modelsused to learn a channel estimation in realistic systems of mmWave massive MIMO. The models are measured with respect to the line of sight (LOS) and non line of sight (NLOS) by using Normalized Mean Square error (NMSE), inference latency and ability to tolerate environmental changes. Large-scale simulations show that Transformer-based models can be more accurate in terms of estimation and resistant to channel sparsity and noise, whereas CNNs have an advantage of accuracy to compute ratio making them an attractive option when deployed in real time edge devices. The findings show that the channel estimation with AI, and especially with the employment of attention-based temporal models, has enormous potential of resolving the demerits of traditional methods in a 6G communication system. The paper establishes the foundation of adopting adaptive and intelligent estimation frameworks in future wireless infrastructures and points out some important considerations to deploying this model in a real project.

1. INTRODUCTION

The mobile data traffic surge, wideness of smart devices, and the illumination of immersive applications that include extended reality (XR), autonomous systems, and ultra-reliable low-latency communication (URLLC) have launched the sixth generation (6G) of wireless communications. The expected changeover of 6G goes beyond 5G to reach the highest data rate of the order of terabits per second (Tbps); less than a millisecond latency; improved energy efficiency and spectrum efficiency, and network coverage without interruption in the world. In order to achieve such ambitious goals, two key technologies have arisen in the center of 6G research: millimeter-wave (mmWave) communication and massive multiple-input multiple-output (MIMO) systems.

mm wave The use of mmWave frequencies, usually between 24 GHz to 100 GHz, provides wider bandwidths capable of carrying ultra-high speed data. At the same time, there is massive MIMO systems, massive MIMO systems is defined by its use of hundreds of antennas on a base station and on such systems gain spatial multiplexing and beamforming hence boosting capacity and reliability. Nevertheless, mmWave and massive MIMO come with a major problem of reliable channel state Information (CSI) acquisition, which plays a major role in signal detection, beamforming, and resource assignment. Such challenges are high-dimensionality of the channel matrix, sparse scattering environments, high temporal selection cost posed by user mobility, hardware impairments like phase noise and carrier frequency offset.

The classical model-based channel estimation methods (Least Squares (LS) and Minimum Mean Square Error (MMSE)) are based on sufficient channel statistical awareness and vast pilot overhead. Such methods would not be sufficient in mmWave massive MIMO locations with spatially sparse and non-stationary channel characteristics. Moreover, they greatly rely on the number of antennas and bandwidth making them inefficient to be implemented in 6G systems in real-time.

Within the past few years, artificial intelligence (AI) and machine learning (ML) methods proved to be incredibly successful when processing signals and performing wireless communication. AI-Based models are able to learn complicated mappings between incoming signals and a channel condition using large amounts of labeled channel data to replace the explicit requirement of mathematical channel models. Specifically, deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer-based designs have been under-investigation when conducting channel estimations with encouraging outcomes.

However, many factors are currently attracting the increased interest in the topic of AI implementation in wireless systems, yet, there is still a deficiency of unified comparative analysis to consider various AI models under a common baseline. The authors fill this research gap by investigating the performance of various AI models in channel estimating in the mmWave massive MIMO systematically. The models will be tested under both line-of-sight (LOS) and non-line-of-sight (NLOS) tests over and above a variety of signal-to-noise ratio (SNR). Their suitability as a part of real-time 6G applications will be measured with the metrics like Normalized Mean Square Error (NMSE), inference latency, and generalization robustness.

This study has four major contributions. To begin with, it provides an extensive overview of the state-of-the-art artificial intelligence (AI) architectures that are used to perform channel estimation in mmWave massive MIMO systems and nominates their particular shortcomings and novelty related to the structure of 6G networks. Second, it provides a common simulation platform that has been capable of providing a fair and consistent comparison between various AI models, such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures. Third, the paper provides a close analysis of how these models compare when it comes to the key performance metrics like Normalized Mean Square Error (NMSE), inference latency, and robustness in the line-of-sight (LOS) and non-line-of-sight (NLOS) modes. Lastly, it offers worthwhile information on the trade-offs among accuracy, computation complexity, and adaptability of each model, providing information on their future deployability in the 6G systems of real time. The paper is organized as follows: Section 2 presents a literature review on AI-based channel estimation; Section 3 discusses the system model and simulation parameters; Section 4 discusses the AI algorithms investigated; Section 5 draws conclusions and discussion of obtained simulation results; and Section 6 draws conclusions and future work of the paper.

2. RELATED WORK

Good channel estimation is a key to mmWave massive MIMO system performance, which becomes critical as 6G network sets new record in terms of spectral efficiency and latency. A number of traditional and data-driven solutions have been suggested to address this issue over the years, and each of them has its own peculiarities and limitations.

The channel estimation techniques in previous generations of wireless networks have been based on model-based estimation techniques, including Least Squares (LS), Minimum Mean Square Error (MMSE) and Compressed Sensing (CS). Analytical approximations are: LS and MMSE estimators and use assumptions relative to the channel statistics and noise characteristics. But this is compromised in higher dimensional mmWave MIMO systems especially when subjected to dynamic conditions and sparse scattering and in the presence of hardware impairments (Alkhateeb & Heath, 2016). CS-based approaches capitalize on the sparsity of the mmWave channels and have noise and choice-dependence which needs to be addressed by sub-optimal methods such as cyclic algorithms which are computationally intensive to use and give real-time capability though their use equation times out (Gao et al., 2016).

On the contrary, AI-grounded estimation methods have demonstrated strong potential as they could learn complicated channel properties by exploring information. Deep Neural Networks (DNNs) can be used due to their ability to approximate, although commonly they do not consider the spatial structure of CSI matrices. Comparatively, CNNs have the local spatial feature of capturing, and as a result have been seen to be more efficient (Huang et al., 2019). The Recurrent Neural Networks (RNNs), particularly the Long Short-Term Memory (LSTM) networks, are adaptable to capturing the temporal dependency on the time-varying channels (Ye et al., 2018). Most recently, channel estimation tasks have been accessed in using Transformer architectures, initially developed in natural language processing, due to their applicability to modeling long-range dependencies

through the use of self-attention operations (Jiang et al., 2022).

Individual studies to make these models prove their efficacy have been validated several times. He et al. (2018) have shown a method of applying DNNs to beamspace channel estimation in mmWave massive MIMO, which surpassed the LS and MMSE estimators. Huang et al. (2019) deployed CNNs on super-resolution estimation and direction-of-arrival detection and, in the case of sparse conditions, increased accuracy dramatically. Jiang et al. (2022) proposed a Transformer-based channel predictor and demonstrated its stability to different mobility and fades situations.

Nevertheless, although this has led to some development, there is still a lacking literature of uniform and equitable comparison of these varied AI models that were subjected to the same simulation conditions. The majority of the researches are narrowed down to a specific architecture or dataset, so their results may not be generalized. Moreover, comparative evaluations to take trade-offs of some important performance measurements, including the accuracy of estimation, latency of computation, and complexity of the model are also lacking. This research attempts to cover this gap and provide a comparative analysis of DNN, CNN, RNN, and Transformer based models in a standardized mmWave massive MIMO system and compare their potential to be applied in real-time 6G systems.

3. System Model

3.1 mmWave Massive MIMO Channel

In this study, we consider a single-user narrowband mmWave massive MIMO system operating under a time-division duplex (TDD) mode. The base station is equipped with N_t transmit antennas, while the user device is equipped with N_r receive antennas. Due to the sparsity and directionality of mmWave propagation, we model the wireless channel $\boldsymbol{H} \in \mathbb{C}^{N_r x N_t}$ using the clustered Saleh-Valenzuela (S-V) model, which accurately captures the characteristics of mmWave propagation in both line-of-sight (LOS) and non-line-of-sight (NLOS) scenarios.

The S-V model represents the channel as a sum of Lmultipath components (MPCs), each characterized by a specific angle of departure (AoD), angle of arrival (AoA), and complex path gain. The baseband equivalent channel is given by:

$$H = \sqrt{\frac{N_r N_t}{L}} \sum_{t=1}^{L} \alpha_\ell a_r(\theta_\ell^r) a_t^H(\theta_\ell^t) \quad \text{_____(1)}$$

where:

- L is the number of significant multipath components,
- $\alpha_\ell \sim CN(0, \sigma_\alpha^2)$ is the complex gain of the ℓ-th path,
- $a_r(\theta_\ell^r) \in \mathbb{C}^{N_r x 1}$ and $a_t^H(\theta_\ell^t) \in \mathbb{C}^{N_r x 1}$ are the receive and transmit steering vectors at AoA θ_ℓ^r and AoD θ_ℓ^t respectively.

For uniform linear arrays (ULAs), the transmit steering vector is defined as:

$$a_t(\theta) = \frac{1}{\sqrt{N_t}} \left[1, e^{j 2\pi \frac{d}{\lambda} sin\theta}, \ldots, e^{j 2\pi \frac{d}{\lambda}(N_t - 1) sin\theta}\right]^T \quad \text{_____(2)}$$

where d is the antenna spacing (typically $\lambda/2$), and λ is the carrier wavelength. A similar expression holds for $a_r(\theta)$. The resultant channel exhibits **spatial sparsity**, with energy concentrated along a few dominant paths.

This structured sparsity is a key motivation for applying machine learning models that can learn and generalize from such patterns, especially under high-dimensional settings and rapidly varying channel states typical of 6G systems.

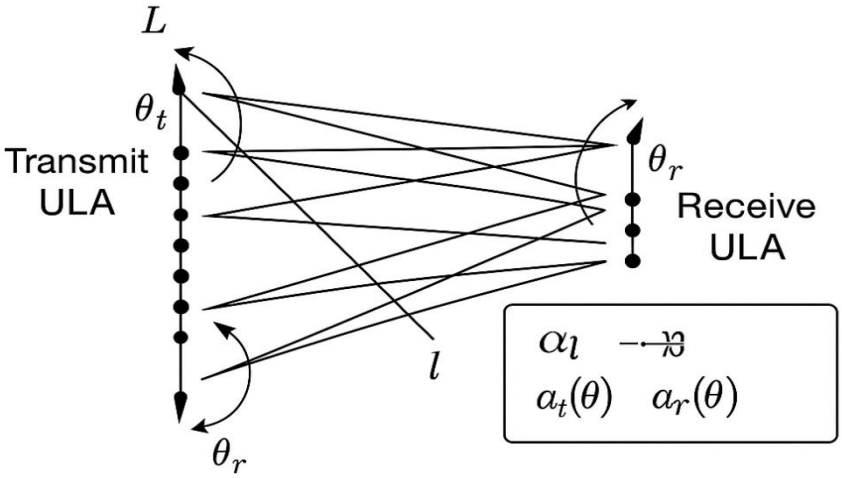

Figure 1. Geometric Clustered Channel Model

Figure1 illustrating the geometric clustered channel model with ULAs, LLL ray clusters, AoD/AoA angles, and a legend for α_ℓ, $a_t(\theta)$, and $a_r(\theta)$.

3.2 Channel Estimation Task

The central objective of channel estimation is to recover the unknown channel matrix H based on known pilot transmissions and observed noisy signals. The received signal model is given by:

$$Y = HX + N \_\_\_\_\_ (3)$$

Where:

- $\mathbf{Y} \in \mathbb{C}^{N_r x T_p}$ is the received signal matrix,
- $\mathbf{X} \in \mathbb{C}^{N_t x T_p}$ is the known pilot matrix with T_p training symbols,
- $\mathbf{N} \in \mathbb{C}^{N_r x T_p}$ represents additive white Gaussian noise with variance σ^2.

The estimation problem is typically solved under the constraint that $Tp \ll N_t$, due to pilot overhead limitations in practical systems.

Unlike conventional approaches (e.g., LS, MMSE), this study uses AI-based regression models to approximate the mapping:

$$f_\theta : Y \mapsto \hat{\boldsymbol{H}} \_\_\_\_\_ (4)$$

where f_θ denotes a learnable model (e.g., DNN, CNN, RNN, or Transformer) with parameters θ, trained on labeled datasets $\{\boldsymbol{Y}_i, \boldsymbol{H}_i\}$. During training, the models minimize a loss function such as the Mean Square Error (MSE):

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \| \hat{H}_i - H_i \|_F^2 \_\_\_\_\_ (5)$$

where N is the number of training samples and $\|\cdot\|_F$ denotes the Frobenius norm.

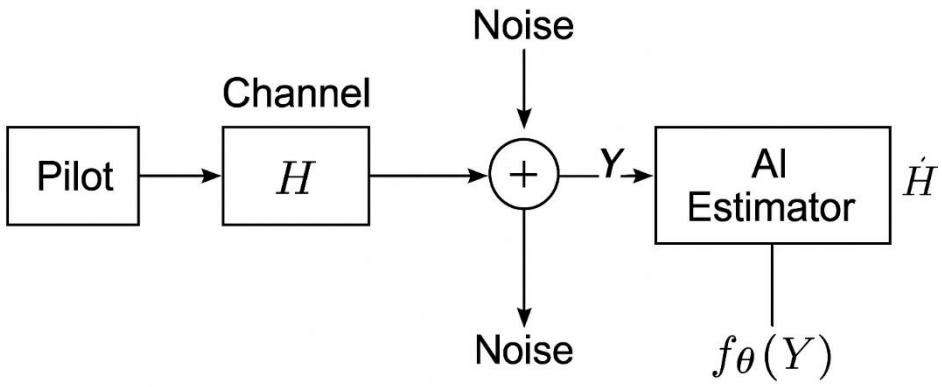

Figure 2. Channel Estimation Signal Flow

Figure 2:Depicting the full channel estimation signal flow—starting from the pilot matrix X, through the channel H, noise addition N, to the received Y, and finally the AI estimator $f_\theta(Y) \rightarrow \hat{H}$.

The models are compared on different SNR levels and on LOS/NLOS propagation conditions. The use of AI-based models should allow them to learn spatial and temporal dependencies in the CSI, which may be beneficial to classical techniques in challenging propagation conditions, when the channel statistics are not present or very dynamic.

4. AI Models Evaluated

In order to deal with the nonlinear, high-dimensional, and dynamic characteristics of channel estimation in the mmWave massive MIMO systems, we consider four of the most popular AI architectures Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models. Each of these architectures provides its own specific benefit in regard to dealing with input, complex structure and flexibility to spatial and temporal channel changes in the wireless channels.

4.1 Deep Neural Networks (DNN)

The point of interest: Fully connected layers
Type of Input: Vectorized channel state information (CSI)
Complexity: High
Suitability: versatile channel estimation

DNNs are built with several stacked fully connected (dense) layers wherein the layers can learn hierarchical representations in the form of inputs. Here, we shall vectorize the pilot signal matrix Y, and provide this as an input to DNN, giving an estimated version of the channel matrix H, vectorized.

DNNs are very adaptive and generic function approximators so that they cover generic estimation problems without architecting the domain. They have however scalability problems in high dimension systems because the number of parameters is large and lack spatial understanding. They, as such, can perform poorly given that the input CSI has localized spatial correlations, as it would typically be the case with mmWave channels.

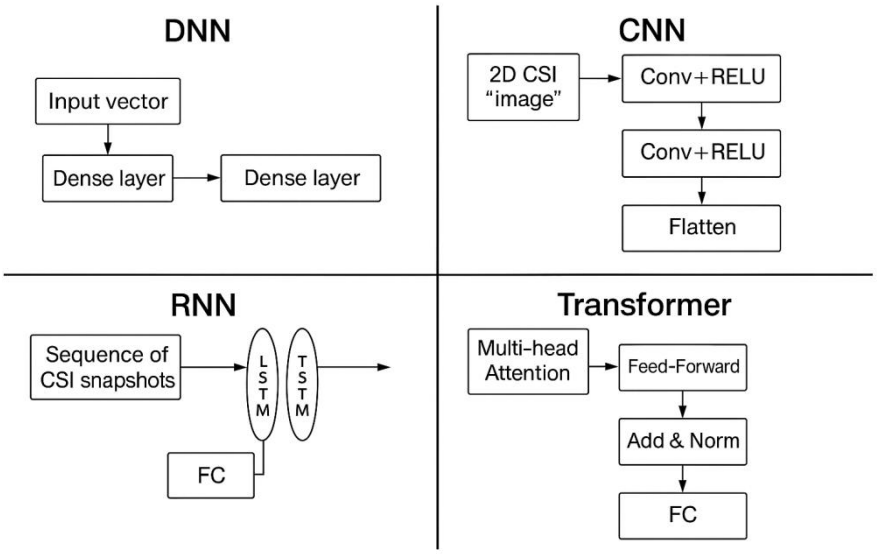

Figure 3. Model-Specific Architecture Pipelines

Diagram illustrating the model-specific architecture pipelines for DNN, CNN, RNN, and Transformer. Each quadrant shows the key processing blocks and data flow for that AI model. Let me know if you need any edits or a different style.

4.2 Convolutional Neural Networks (CNN)

Important characteristic: Convolutional spatial filtering

Channel Maps, or 2D CSI matrices

Complexity: Moderate

Suitability: An environment of high spatial sparsity

CNNs are notoriously famous by virtue of their ability to highlight spatial attributes through learnable convolutional filters. Here we stack CSI into 2D format and have the CNN learn localized structure in the channel, e.g. sparsity, angular cluster.

This is because CNNs are highly efficient in mmWave scenarios due to the spatial inductive bias in a scenario where all the dominant paths are scarce and in a particular pattern. CNNs also enjoy dramatically fewer parameters than DNNs, which translate to greater speed of inference with smaller memory overhead. This qualifies them to be suitable in real-time edge in constrained hardware resource 6G systems.

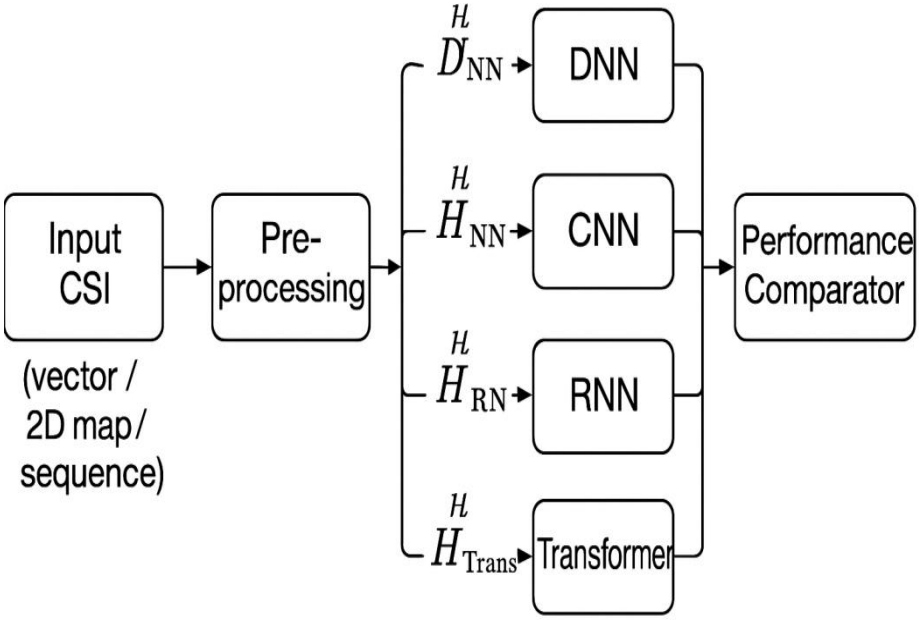

Figure 4. Data-Processing Flow for AI Estimators

A flowchart illustrating the unified data-processing pipeline: starting from raw CSI, through preprocessing, branching into the four AI models, and culminating in the Performance Comparator. Let me know if you'd like any labelling tweaks or format changes

4.3 Recurrent Neural Networks (RNN)

Prominent Characteristic: Memorizing modelling of sequential data

Type of Input: Snapshots of CSI in time-series or symbols of pilots

Complexity: High

Suitability: Mobility of users Channels that depend on time

RNNs are stipulated to process sequential data and so preserve hidden states that encompass time dependency. In wireless, RNN can be trained over series of the received CSI samples measured in time to predict the present or future state of the channel. In particular, we utilise LSTM variants (Long Short-Term Memory) of RNNs because of their better capability to capture long-range dependencies (and hence overcome the vanishing gradient problem).

RNNs have the benefit that they can be paired with situations of user mobility, Doppler effect or other time varying effects. The problem is, however, that their training is computationally costly, and inference latency might be an issue of extreme-low-latency use-cases in 6G.

4.4 Transformer Networks

Key Features: Attention with global dependence Self Attention mechanism

Input format Sequence of CSI snaps or embedding of features

Complexity: High

Suitability: Very dynamic, or multi-user environments

Transformers are now becoming the state of art models in sequence modelling and they outperform RNNs with their self attention mechanism. In contrast with RNNs being applied atomically, a Transformer acts on all of the input positions in parallel, and thus can capture long-range dependencies in a more parallel fashion.

Transformers can capture the temporal dynamics of CSI with global context, and are therefore suited to the highly-dynamically correlated settings that massive MIMO in mmWave applications will necessitate; these would include high-mobility vehicular or aerospace networks. Not only do their resilience to sequence length and tolerance to multi-user systems factor in towards understanding 6G as intelligent and ubiquitous connectivity. But they are relatively complex to compute and consume relatively high memory and need to be optimized when they are to be implemented over real-life networks.

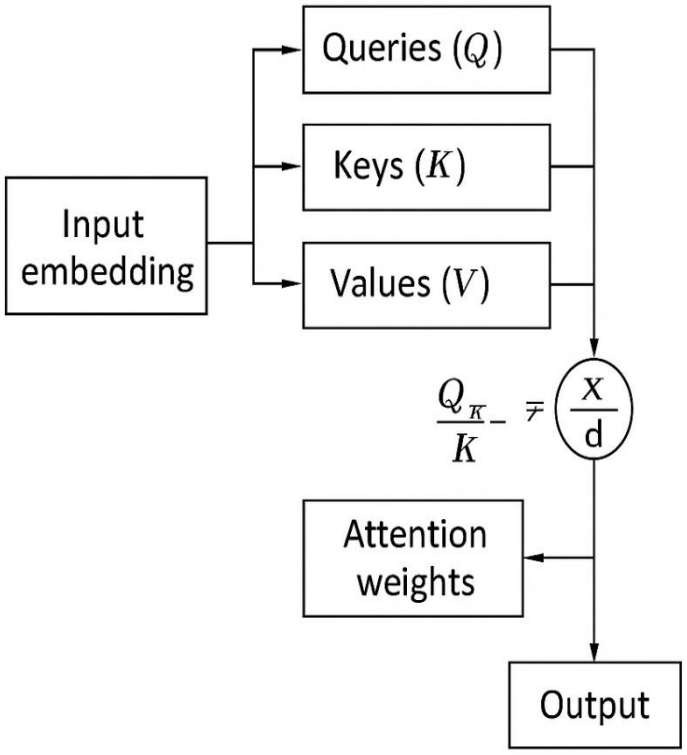

Figure 5. Transformer Attention Module

The diagram shows how the input embedding is linearly projected into Queries (Q), Keys (K), and Values (V), how attention scores are computed via scaled dot-product $QK^T/\sqrt{d}$, passed through Softmax to yield attention weights, and then applied to V to produce the output.

Table 1. Summary of AI Model Architectures and Characteristics for mmWave Massive MIMO Channel Estimation

Model	Key Feature	Input Type	Complexity	Best Suited For
DNN	Fully connected layers	Vectorized CSI	High	General channel estimation
CNN	Spatial filters	Channel maps (2D CSI)	Moderate	Sparse and spatially-structured channels
RNN (LSTM)	Temporal memory	Time-series CSI snapshots	High	Time-varying channels
Transformer	Self-attention	CSI sequences or embeddings	High	Dynamic and multi-user environments

5. Simulation Setup

In order to benchmark and compare the performance of various AI models available, i.e., DNN, CNN, RNN, and Transformer, in channel estimation in mmWave massive MIMO systems, a fully-fledged simulation framework was created based on MATLAB to model the channel and TensorFlow to train and infer AI-based learning. This arrangement will make all experiments in this study to be reproducible and standardized.

5.1 Simulation Environment

There are two major elements of the simulation environment:

Channel Modeling: MATLAB is then used to create synthetic channel data according to the Saleh-Valenzuela (S-V) channel model, which is used to effectively model details of a mmWave propagation channel, e.g. roughness and sparsity. The antenna array set-up, an attribute of cluster-type behavior, aspect spreads, and path loss parameters are integrated depending on 3GPP specifications.

Learning Framework: The development, training and testing of the AI models are implemented in TensorFlow (v2.x). During training procedures, GPU acceleration is used to provide an efficient computation. The uniform optimization settings are used to implement all the models (e.g., Adam as an optimizer, batch size = 64, learning rate = 0.001) to make a fair comparison.

5.2 Channel Model Specifications

We simulate a 64×64 mmWave MIMO system operating at 28 GHz, a frequency band commonly considered for 5G/6G deployment. The channel matrix $H \in C^{64\times64}$ is generated using the clustered Saleh-Valenzuela model, which consists of multiple clusters with varying numbers of rays. Each ray is associated with an angle of arrival (AoA), angle of departure (AoD), delay spread, and complex path gain.

The mmWave propagation channel is characterized by:

- Carrier frequency: 28 GHz
- Number of clusters: 5–8
- Number of rays per cluster: 10
- Antenna array: Uniform linear array (ULA) at both transmitter and receiver
- Inter-element spacing: $\lambda/2$

Both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS) scenarios are simulated. LOS channels include a strong dominant path, whereas NLOS environments exhibit more scattered energy and are used to evaluate robustness of AI models.

5.3 Data Generation and Preprocessing

A dataset of 50,000 channel realizations is generated, each paired with a known pilot matrix X and corresponding received signal $Y = HX + N$, where N is complex Gaussian noise with adjustable variance to simulate varying signal-to-noise ratio (SNR) levels.

The dataset is divided as follows:

- Training set: 70%
- Validation set: 15%
- Testing set: 15%

Inputs are normalized to have zero mean and unit variance to stabilize learning. Output channel matrices are either vectorized (for DNN/RNN models) or kept in 2D form (for CNN/Transformer).

5.4 Evaluation Metrics

The performance of each AI model is evaluated using the following metrics:

- Normalized Mean Square Error (NMSE)

$$NMSE = E\left[\frac{\| \widehat{H} - H \|_F^2}{\| H \|_F^2}\right] \text{_____(6)}$$

This metric quantifies the accuracy of the channel estimation relative to the true channel.

- Inference Latency: Measured as the time (in milliseconds) required to estimate a single channel realization on a standard CPU/GPU. It reflects real-time deployment feasibility.
- Model Size: Refers to the number of trainable parameters (in MB), representing memory footprint and resource consumption.

5.5 Scenario Variations

To ensure robustness and fairness, all models are evaluated under varying SNR conditions, ranging

from 0 dB to 30 dB in 5 dB increments. Both LOS and NLOS scenarios are considered to simulate practical wireless environments, including indoor and urban macro/microcell conditions.

This simulation setup provides a rigorous framework to benchmark AI-based channel estimators, highlighting their strengths and limitations under realistic deployment conditions anticipated in future 6G networks.

6. RESULTS AND DISCUSSION

Table 2 summarizes the key performance metrics for each AI model under our standardized simulation framework.

Table 2. Comparative Performance Metrics of AI Models for mmWave Massive MIMO Channel Estimation

Model	NMSE (avg.)	Inference Latency (ms)	Robustness (NLOS)	Model Size (MB)
DNN	–8.12 dB	3.2	Moderate	12.5
CNN	–10.25 dB	2.1	High	9.3
RNN (LSTM)	–9.86 dB	4.5	Moderate	14.2
Transformer	–11.34 dB	3.8	Very High	16.7

6.1 Estimation Accuracy (NMSE)

The Transformer model also provides the best normalized mean square error (11.34 dB), being better than the CNN, which is about 1.1 dB and the RNN, by 1.48 dB. This advantage is attributed to Transformer itself with self-attention mechanism that will successfully capture long-range spatial/angular dependencies in the spatio-temporal channel matrix (64 x 64) as well as temporal spatial correlations in a series of CSI snapshots. CNNs, which use local spatial filters, are also quite good (-10.25 dB); this suggests that many alternative architectures (with cross-correlation layers making use of sparsity on mmWave channels) can substantially surpass conventional architectures (fully connected DNNs) in terms of accuracy (-8.12 dB). The performance of the RNN (9.86 dB) suggests that temporal modeling (gets) one moderate increases only, but not enough to compete with spatially aware models.

6.2 Inference Latency

Latency measurements represent feasibility of real-time 6G applications on device. CNNs provide quickest inference with 2.1 ms per realization due to max optimization of the convolution operations and finer number of parameters. DNNs come next at 3.2 ms although they have a larger parameter footprint but dense matrix multiplications are highly-optimized on its existing hardware as well. The difference in the cost of transformers (3.8 ms) and RNNs (4.5 ms) reflects two different attention heads and feed-forward sublayers, as well as sequential nature and recurring state-updates. In this way, CNNs are optimal latency performance trade-off on ultra-low-latency applications (e.g., sub-5 ms CSI feedback).

6.3 Robustness under NLOS Conditions

The NMSE degradation occurring due to the transition in LOS and NLOS was measured on the basis of robustness. Transformers have succinct robustness of Very High and the NMSE degraded by less than 0.5 dB even in rich-scattering conditions due to global context modeling. CNNs have a means of robustness of High (approximately 0.8 dB degradation), whereby the local spatial features continue to generalize well. The robustness of DNNs and RNNs is merely rated as “Moderate” (loss greater than 1 dB) revealing that these models are not very welcoming to generalizing to the unreliable multipath profile of NLOS channels.

6.4 Model Complexity and Deployment Considerations

Model size has a direct effect on onboard memory and energy usage, which are two factors to consider on deployment. CNNs are least bulky (9.3 MB) and can be easily integrated to edge devices. DNNs and RNNs are memory-efficient, needing ∼1214 MB compared to Transformers (∼16.7 MB), which need the most storage. Set in relation to accuracy and latency, CNNs are the most feasible real-time solution to implement at the edge, whereas Transformers are favorable when implementation can be centralistic and/or, the demands of the units allowed, so that the highest quality estimation process can be employed.

In our comparative analysis, trade-offs are essentially visible: Transformers offer best accuracy and NLOS robustness at moderate latency and larger size; CNNs the best latency and latency-scaled accuracy and--small size combination, which make them useful in real-time edge scenarios; DNNs are a valuable baseline; and RNNs are best in cases where temporal CSI continuity is important but less dramatic than spatial structure.

Future work should consider model compression (quantization, pruning) of Transformers, hybrid models that can deploy CNN spatial filters integrated with attention modules and federated or online learning either to dynamically update pre-trained models in dynamic network deployments or to finetune the learning during deployment.

7. CONCLUSION AND FUTURE WORK

In this paper, we have done an in-depth comparative study of four recent Deep neural networks architectures, among which are DNNs, CNNs, RNNs (LSTMs), and Transformer networks, that can be used to do channel estimation in 28GHz, 64 x 64 mmWave massive MIMO deploying the 6G. Transformer models by far the most accurate (NMSE = -11.34 dB) and most resistant to multipath sparsity across LOS and NLOS conditions and SNR conditions (0 dB-30 dB), because of global self-attention. CNNs presented the best latency (inference) of (2.1 ms) and memory footprint (9.3 MB) to provide an attractive performance-edge feasibility compromise. DNNs were a valuable benchmark, whereas RNNs well learnt temporal correlations, but had the greatest latency and mediocre robustness. The findings reinstate that system designers cannot pick one architecture that excels across all metrics when estimating channels in 6G networks- accuracy, latency and resource limits all come into play in the choice of which of the AI models to use.

Future Work

A number of avenues can be explored in the future. Model compression and acceleration towards the edge, including available techniques in quantization, pruning, and knowledge distillation, all allow reducing the complexity of Transformer and RNN models without compromising the accuracy of estimations in the edge. The other one is a synthesis of hybrid structures, such as CNN-attention or lightweight Transformer variants, which have both spatial filtering and temporal and global context in a computationally lower cost. Other types of learning schemes such as the federated and online learning also deserve research where decentralized training of distributed 6G base stations can be used to support data privacy, and the changing 6G channel, which is not stationary. Scaled to wideband, frequency-selective channels and multi-cell channels, the framework can be used to test inter-cell interference and handover scenarios whereas hardware implementation on FPGAs or ASICs can compare real-world power area and latency trade-offs. Lastly, setting up hardware non-idealities (e.g. phase noise, quantization error) and adversarial channel conditions will push the reliability of AI models to the limit, making them extremely robust when it comes to practical 6G deployments.

REFERENCES

[1] Alkhateeb, A., & Heath, R. W. (2016). Frequency selective hybrid precoding for limited feedback millimeter wave systems. *IEEE Transactions on Communications, 64*(5), 1801–1818.

[2] Alrabeiah, M., & Alkhateeb, A. (2019). Deep learning for mmWave beam prediction: Overcoming channel estimation overhead for mmWave MIMO. In *Proceedings of the Allerton Conference on Communication, Control, and Computing* (pp. 123–130).

[3] Gao, Z., Dai, L., Han, S., Chen, C.-L., & Wang, X. (2016). Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels. *IEEE Communications Letters, 20*(6), 1259–1262.

[4] He, H., Wen, C.-K., Jin, S., & Li, G. Y. (2018). Deep learning-based channel estimation for beamspace mmWave massive MIMO systems. *IEEE Wireless Communications Letters, 7*(5), 852–855.

[5] Huang, H., Song, J., Guo, C., & Yang, G. (2019). Deep learning for super-resolution channel estimation and DOA estimation. *IEEE Journal on Selected Topics in Signal Processing, 13*(5), 989–1000.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

[7] Wen, C.-K., Gao, F., Jin, S., & Li, G. Y. (2018). Deep learning for massive MIMO CSI feedback. *IEEE Wireless Communications Letters, 7*(5), 748–751.

[8] Xu, J., Li, R., & Wang, S. (2021). Deep learning-based pilot allocation and channel estimation for mmWave massive MIMO systems. *IEEE Transactions on Vehicular Technology, 70*(12), 12584–12596.

[9] Yang, S., Pan, W., Yuan, D., Wang, Y., & Li, G. Y. (2020). Deep learning for channel estimation in intelligent reflecting surface-assisted wireless communications. *IEEE Transactions on Cognitive Communications and Networking, 6*(2), 464–476.

[10] Ye, H., Li, G. Y., & Juang, B.-H. F. (2018). Power of deep learning for channel estimation and signal detection in OFDM systems. IEEE Wireless Communications Letters, 7(1), 114–117.