# A Deep Learning Based Hybrid Precoding Scheme with Limited Feedback Approach for Improved Compression and Minimized Reconstruction Error in Massive MIMO

Shruthi N. [1],* and K. Ramesha [2]

[1] Department of Electronics and Telecommunication Engineering, Bangalore Institute of Technology,
Affiliated to Visvesvaraya Technological University, Bangalore, India
[2] Department of Electronics and Instrumentation Engineering, Dr. Ambedkar Institute of Technology,
Affiliated to Visvesvaraya Technological University, Bangalore, India
Email: shruthin@bit-bangalore.edu.in (S.N.); kramesha13@gmail.com (K.R.)
*Corresponding author

*Abstract*—**The technological advancements and demand of high speed communication has led to evolvement of Multiple-Input and Multiple-Output (MIMO) and massive MIMO (mMIMO) communication systems. However, the increased number of antennas lead to an increase in computational complexity and implementation cost. Moreover, achieving the performance to meet the communication demand also remains a challenging task. The current researches have reported that the precoding scheme can help to minimize the computational complexity and increase the performance of mMIMO system. Hybrid precoding schemes have gained huge attention due to their significant nature to improve the overall efficiency of the system but the traditional schemes usually focus on optimization or greedy mechanism which suffer from the complexity issues and provide the sub-optimal performance. Moreover, the performance of these systems is directly affected by the quality of channel data. Therefore, we present a Deep Learning (DL) based approach using Deep Neural Network (DNN) model which uses limited feedback mechanism to handle the compression and reconstruction error. It aims to minimize the reconstruction error by providing the transmitter with sufficient information about the Channel State Information at the Receiver (CSIR) despite using a reduced amount of feedback compared to full feedback systems. This scheme uses encoder and decoder based module for limited feedback modelling. In order to prove the robustness of proposed DL based approach, we have presented extensive experimental analysis where the proposed DL based mechanism achieves average performance as 16.85 bits/s/Hz, 12.45 bits/s/Hz, and 8.028 bits/s/Hz in terms of achievable rate, spectral efficiency and average sum rate respectively. In contrast to this, the existing Simultaneous Orthogonal Matching Pursuit (SOMP) achieves the average sum rate as 6.042 bits/s/Hz.**

*Keywords*—**deep learning, limited feedback, precoding, compression, massive Multiple-Input and Multiple-Output (mMIMO)**

## I. Introduction

Recently, the demand for wireless communication has increased drastically due to its enormous use in a wide range of real-time applications such as health monitoring, underwater exploration, environmental monitoring, cellular communication, etc. [1]. In this domain of wireless communication, cellular and mobile communication traffic has upsurged because of the increased popularity of mobile devices. Moreover, the research community has observed a 1000-fold increase in data traffic in the year 2020, and it is expected to increase by over 10000-fold by the year 2030 [2].

The increased demand for cellular communication urges high-speed Internet connectivity in various domains such as smart cities, self-driving cars, infotainment applications, etc. Therefore, supporting the incessant growth in data traffic and facilitating guaranteed ubiquitous communication have become important aspects of the current cellular communication systems. Moreover, Ericsson presented a mobile data traffic forecast report which suggests that by 2028 all mobile data traffic will come from 5G. The global monthly average usage per smartphone is expected to be 19 GB in 2023 and this average usage will reach up to 46 GB by the end of 2028 [3]. The 5G data traffic is expected to reach 69% by the end of 2028. According to the current demand and communication, high data rate, communication capacity, spectral efficiency, and energy efficiency are the most desired characteristics. Bhairanatti *et al.* [4] presented an extensive literature review to report the current progress and challenges faced in this domain. The traditional communication standards such as 2G, 3G, and 4G are not efficient to meet the current demand for communication quality. Recently, the use of Multiple-Input Multiple-Output (MIMO) technology has gained substantial

Manuscript received December 26, 2023; revised March 26, 2024; accepted April 9, 2024; published January 15, 2025.

doi: 10.12720/jait.16.1.91-100

consideration because it helps to achieve increased throughput and improves spectral efficiency [5].

MIMO refers to a communication standard that uses manifold antennas both at transmitter and receiver sides. This multiple antenna setup is used to improve the data rate and data transmission quality [6]. MIMO systems adopt spatial multiplexing and spatial diversity techniques to transmit independent and separately encoded data by using the same time period and frequency resources [7]. The MIMO systems are categorized into two classes as single-user MIMO (SU-MIMO) with one user and this mechanism improves the data rate for one user only. On the other hand, multi-user MIMO (MU-MIMO) systems [8] are such where data streams are assigned to various users and it facilitates the spatial multiplexing to enhance the overall communication. Fig. 1 depicts a sample representation of SU and MU MIMO systems.

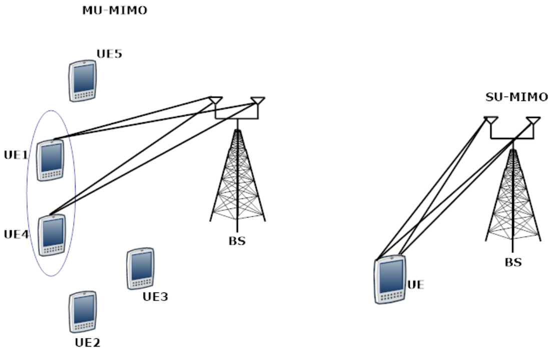

Fig. 1. MU and SU MIMO system.

The expansion in antenna size is termed as massive MIMO [9]. Similarly, the millimetre Wave (mmWave) communication standards also provide strength to cellular communication systems by addressing the faster data rate-related issues for cellular networks. The mmWave communication has resulted in a ten-fold increase in the carrier frequency of wireless communication systems [10]. The large antenna array has several advantages but the real-time deployment of fully digital massive MIMO poses challenges such as excessive power consumption. Similarly, the mmWave standards face more path loss than the traditional sub-30GHz communication [11].

However, the ever-increasing demand for high throughput with limited resources has led to adopt the mMIMO systems which use a large number of antennas at the base station. Fortunately, these issues of mmWave communication can be mitigated by using the mMIMO antenna arrays [12]. Therefore, the mMIMO systems are considered as a promising solution to increase capacity, spectral and energy efficiency [13]. The mMIMO systems include three key features such as spatial multiplexing [14] which is useful in serving multiple users simultaneously, precoding which helps to minimize the interference [15], and, beamforming which helps to concentrate the energy of signal towards the receiver to ensure the efficient directional transmission [16]. Michael *et al.* [17] introduced a novel approach for antenna arrangement and beam pattern design to mitigate the interface issues while maintaining the MIMO radar's performance. Similar to this, Sameera *et al.* [18] focused on designing a detector for MIMO radar network with two transmitters and three receivers. This model performs detection for given bi-static pair of networks by employing the Linear Frequency Modulated (LFM) signal. Despite potential applications of mMIMO, this technology still faces several challenges which need to be addressed such as the mMIMO systems operate at extremely high frequencies due to which the Doppler shift increases linearly and as a result channels face frequent variations [19]. These systems are employed in Time-Division Duplex (TDD) and Frequency Division Duplex (FDD) [20, 21] modes, however, TDD is preferable because it supports multiple antenna placement at BS to increase the system capacity. In TDD mode the time required for uplink transmission and downlink CSI feedback may exceed the coherence time constraints of the communicating channel [22]. Moreover, hardware complexity, implementation cost, and power consumption are the well-researched challenges of mMIMO systems.

However, the increased number of antennae at BS in the mMIMO system leads to increase in the complexity of the system. Therefore, reducing the complexity is one of the important aspects of uplink and downlink communication systems. Precoding is considered as a promising technique to reduce the computational complexity in mMIMO systems [23]. It is a method that includes the encoding of signals to ensure reliable downlink transmission. Generally, precoding techniques are classified into three main categories as linear, non-linear, and hybrid precoding [24]. However, it is a difficult process to equip more antennas with a dedicated Radio Frequency (RF) chain due to limited physical space, power consumption, and close placement of antennas. In order to address these issues of implementation cost and complexity, researchers have suggested adopting the phase-shifter-based two-stage structure which is called hybrid precoding [25]. In this work, we focus on introducing a DL based precoding scheme to overcome the complexity issues. Some of the challenges faced in existing precoding methods are as follows:

- **Complexity**: These methods require significant computational resources therefore it becomes crucial for real-time application with stringent latency criteria.
- **Channel State Information:** Most precoding methods rely on accurate CSI at the Transmitter (CSIT) to optimize the transmission. However, obtaining accurate CSIT often involves feedback from the receiver, which introduces overhead and latency. Moreover, imperfect CSI due to channel estimation errors can degrade precoding performance.
- **Channel conditions:** Varying channel conditions in real-time scenario affects the precoding process.
- **Energy efficiency:** Energy consumption is considered as critical factor in wireless communications. Energy-efficient precoding algorithms that minimize transmit power while maintaining communication quality are desirable.

Rest of the article is structured in following sections: Section II presents a brief analysis on literature review where linear, nonlinear and hybrid precoding are discussed, Section III presents the proposed DL based solution for hybrid precoding, Section IV presents the experimental analysis, and finally Section V presents the concluding remarks.

## II. Literature Review

This section presents a brief literature review of existing techniques in this domain of massive MIMO systems. The previous section has described the challenges faced in massive MIMO systems and the role of precoding to increase the number of antennas at BS. Several types of research have been carried out in this domain of cellular communication due to the advancement of precoding mechanisms in MIMO systems.

### *A. Linear & Non-Linear Precoding*

Precoding is a technique to exploit the Channel-State Information at the Transmitter (CSIT) by analyzing the signal before transmission. Linear precoder is operated as a multimode beamformer which helps to match the input signal on one side to the channel on the other side. In order to do this, the signal is divided into orthogonal and spatial Eigen beams. Further, the high power is allocated where channel is strong and low power is allocated where channel is weak [26].

Interference cancellation becomes an important task to achieve high throughput performance for downlink cellular communication systems. Therefore, Wei *et al.* [27] introduced a linear precoding method that uses channel side information with the help of cache files to model the cache-aided mMIMO system. This system increases the degree of freedom for precoder design and it frees power that can be beneficial to the users requesting non-cached files.

According to Ref. [28], the traditional methods steer the distortions towards the users therefore authors considered the nonlinear nature of Power Amplifiers (PAs) to design the linear precoders to mitigate the distortion completely. However, it affects the array gain, therefore a precoder optimization algorithm is also presented which considers array gain, distortion, interference and noise. Specifically, an iterative mechanism is introduced to obtain the precoding matrix which is used to minimize the consumed power and improve the achievable sum rate.

Liu *et al.* [29] reported that the traditional Zero-Forcing (ZF) precoding method can achieve the near-optimal sum-rate performance for downlink communication in mMIMO systems. However, computational complexity with matrix inversion remains a challenging issue in this. To overcome the drawbacks of ZF precoding, authors introduced Weighted Two-Stage (WTS) precoding which converts the complex matrix into two-half iteration stages. Further, the speed and convergence of these models is improved with the help of weighted coefficients.

Zhang *et al.* [30] reported that resource allocation is a challenging task in downlink mMIMO systems therefore authors introduced a joint Proportional-Fair (PF) resource allocation approach which considers user selection, power optimization, linear precoding, modulation, and coding scheme in single-cell mMIMO systems.

Ha *et al.* [31] focused on satellite communication with mMIMO systems. This approach presents a joint approach with Linear Precoding (LP) and codebook-based beamforming mechanism. This approach is also based on the codebook method which are constructed with the help of Discrete Fourier Transform (DFT) which helps to maximize the achievable throughput. Later, LP and DFT-based beamforming methods use a weighted minimum mean square error transformation, duality, and Hungarian algorithm to address the challenges of non-linear programming.

As discussed in [29], the matrix inversion affects the performance of the downlink mMIMO system. Therefore, Wang *et al.* [32] suggested linear precoding for faster convergence with reduced complexity and global convergence. The authors introduced a randomized iterative precoding approach to mitigate the approximation error. Later, the conditional sampling paradigm is also introduced which helps to increase the convergence and efficiency of randomized iterations.

Other widely known LP methods include Successively-Regularized Zero Forcing [33], deep learning based zero-forcing [34], maximum ratio transmission [35], truncated polynomial expansion [36], and regularized zero-forcing [37] etc.

Similarly, non-linear precoding techniques are also widely adopted in mMIMO systems. These schemes are implemented when the channel state information is known to the transmitter side. Some of the widely known non-linear precoding techniques include dirty-paper coding [38], vector-perturbation coding [39] and Tomlinson-Harashima (TH) coding [40], etc.

### *B. Hybrid Precoding*

This section presents a brief discussion about hybrid precoding because of its importance in reducing hardware complexity and minimizing energy consumption.

Kabalci *et al.* [41] presented an optimal hybrid precoding method that uses Iterative Geometric Mean Decomposition (IGMD) to achieve optimal performance. However, this method fails to exploit spatial information and suffers from the computational complexity issue. Therefore, Huang *et al.* [2] focused on the importance of the deep learning-based approach and presented a deep learning-based hybrid precoding scheme that helps to obtain the optimized decoder by mapping the deep neural network.

According to Ref. [42], channel estimation and hybrid precoding are promising techniques for mMIMO system. Therefore, the authors introduced DL based compressed sensing approach for channel estimation. The DL model is trained using offline environments to predict the beamspace channel amplitude. Later, DL quantized phased hybrid precoder is designed for channel estimation. This model is trained offline by considering the phase quantization and phase quantization approximation is replaced with ideal phase quantization to obtain the hybrid precoding.

Elbir and Papazafeiropoulos [43] suggested adopting the hybrid precoding to improve the sum rate performance of mmWave MIMO system because the existing schemes are based on greedy and optimization-based approaches. The traditional methods provide sub-optimal performance therefore authors introduced deep learning-based approach that accepts input as an imperfect channel matrix. The complete process is divided into two stages: in the first stage exhaustive search approach is developed to select the analog precoder with the help of a predefined codebook.

Ravikumar *et al.* [44] discussed the need of RF chain in mMIMO system. Therefore, authors focused on development of a cost-efficient solution for hybrid precoding. This method facilitates collection of short dimensional precoding data from high dimensional beam-former in the digital domain. Thus, the main aim of this approach is to develop a combined channel estimation and hybrid precoding method for mmWave communication system. The channel estimation is carried out with the help of adaptive deep convolution neural network which performs channel estimation and reconstruction. Further, Forest-Tunicate Swarm Algorithm (F-TSA) is also employed to enhance the efficiency of convolution neural network.

Liu *et al.* [45] focused on deep learning-based approach for channel estimation and proposed a spatial-frequency ECAUNet++ (SF-ECAUNet++) approach. This mechanism performs compression and reconstruction by using correlation between spatial and subcarrier frequency. Moreover, it also includes attention neural network to predict the channel.

Rajarajeswarie *et al.* [46] reported that the precoder play vital role in mMIMO systems by reducing the complexity. The traditional precoder systems require multiple radio frequency chains which need to be reduced to improve the overall performance. Thus, authors introduced hybrid precoding mechanism by using deep learning for designing the hybrid precoder. This helps to solve the non-convex problem. The deep learning training is based on the Uniform Channel Decomposition (UCD) and Generalized Triangular Decomposition methods.

Ismail *et al.* [47] proposed a deep learning approach to design the hybrid precoder. This mechanism is based on the Parametric Rectified Linear Unit (PReLU) activation function which is helpful in increasing the accuracy with reduced cost.

The literature review on precoding techniques in massive MIMO systems highlights both linear and non-linear precoding methods, emphasizing their role in exploiting Channel-State Information at the Transmitter (CSIT) to enhance system performance. Various approaches, including linear precoding with cache-aided systems and non-linear precoding to mitigate power amplifier distortion, are discussed. Additionally, hybrid precoding methods are explored for reducing hardware complexity and energy consumption. However, many existing approaches face challenges such as computational complexity and suboptimal performance. The review suggests further investigation into deep learning-based precoding methods to address these challenges effectively.

## III. Proposed Model

This section presents the proposed solution to design the optimal precoder for downlink mMIMO communication system. As discussed before hybrid precoding is a favorable mechanism that helps to reduce the number of RF chains to achieve the improved performance. Several techniques have been presented to accomplish the hybrid precoding but acquiring the accurate downlink Channel State Information (CSI) is an important aspect for BS. Therefore, the User Equipment (UE) role comes into picture where these UEs estimate the CSI of downlink and report the CSI information to BS with the help of feedback. Some of the researchers have adopted joint optimization of CSI feedback and hybrid precoding. However, these methods suffer from the issue of additional implementation cost and increased overhead for large number of users and antennas [48]. Previous section has described the several hybrid precoding techniques where CSI feedback and hybrid precoding are realized in separate modules and therefore a combined module is much needed to explore the capabilities of the system. In this work, we adopt the deep learning based framework to design the optimal precoder with CSI feedback.

In this work, we consider a single-cell massive MIMO downlink model with limited feedback where a base station BS is equipped with $M$ antennas. These antennas serve K number of single-antenna users simultaneously. Let us consider that a precoding matrix is denoted by $\boldsymbol{F}$ as $\boldsymbol{F}(\triangleq [f_1, \dots f_K]) \in \mathbb{C}^{M\times K}$ and $\boldsymbol{d}(\triangleq [d_1, .. d_K]) \in \mathbb{C}^{K\times 1}$ denotes the data symbol vector for all user equipments (UEs) $k \in \{1, \dots, K\}$. The BS is equipped with $N_T$ number of transmit antennas along with $N_{RF}$ chains where $K \leq N_{RF}$. Based on these assumptions, the transmit signal vector at BS has been denoted. In this communication setup, the hybrid precoding is obtained via digital baseband precoder denoted by $\boldsymbol{D}$ which has dimensions as $K \times K$ and analog precoder denoted by $\boldsymbol{F}$ which has dimensions as $N_t \times K$ . The input signal is processed through the digital baseband precoder where it adjusts amplitude and phase of the signal whereas the analog baseband precoder adjusts only the phase of input signal. This realization of precoders presents a constraint that modulus of every element in $\boldsymbol{F}$ is a constant as $\left|[\boldsymbol{F}]_{i,j}\right| = \frac{1}{\sqrt{N_t}}$ where $[\boldsymbol{F}]_{i,j}$ is the $(i,j)^{th}$ element of $\boldsymbol{F}$. We consider a flat fading channel, thus, the received signal by $k^{th}$ user can be denoted as:

$$y_k = \boldsymbol{h}_k^H \boldsymbol{F} \boldsymbol{d}_k s_k + \sum_{j\neq k} \boldsymbol{h}_k^H \boldsymbol{F} \boldsymbol{d}_j s_j + n_k \tag{1}$$

where $\boldsymbol{h}_k^H$ represents the channel between BS and $k^{th}$ user, $\boldsymbol{d_j}$ is the $j^{th}$ column of digital precoding $D$ and $n_k$ is the additive white Gaussian noise which has unit variance. Further, the signal vector for all $K$ users is denoted as $s = [s_1, \dots s_k]^T \in \mathbb{C}^{K\times 1}$. As discussed before that energy consumption is also considered as challenging issue. Therefore, we consider the transmit power constraint and normalize the energy of $F$ and $D$ precoder such as that it satisfies $\|FD\|_F^2 = K$. Based on this, the sum-rate of the system can be computed as follows:

$$R = \sum_{k=1}^{K} \log_2(1 + SINR_k) \tag{2}$$

where Signal-to-interference-plus-noise-ratio ($SINR$) for $k^{th}$ user can be articulated as:

$$SINR_k = \frac{\frac{P}{K}\left|h_k^H F d_k\right|^2}{1+\sum_{j\neq k}\frac{P}{K}\left|h_k^H F d_k\right|^2} \tag{3}$$

Furthermore, the optimization problem in this mMIMO system can be framed as:

$$\{\boldsymbol{F}, \boldsymbol{D}\} = \arg\max_{F,D} R$$
$$\text{s.t. } \left|F_{i,j}\right| = \frac{1}{\sqrt{N_t}} \text{ and } \|\boldsymbol{FD}\|_F^2 = K \tag{4}$$

where $|.|$ is the modulus operation and $\|.\|_F$ denotes the Frobenius norm operation. In this work, we focus on reducing the number of feedback CSI parameters. This can be achieved by compressing the channel matrix. This task is carried out by an encoder in UE and it tries to decode the channel matrix accurately at the BS. The encoder module is given as:

$$\boldsymbol{q}_k = f_{En}(\boldsymbol{h}_k) \tag{5}$$

where $\boldsymbol{q}_k$ denotes the compressed codeword of $h_k$ and $f_{En}$ represents the compression operation of encoder. This operation helps to reduce the feedback parameters from $N$ to smaller value of $M = 2KM_t$.The compression ratio is denoted as $\gamma = \frac{M}{N}$. In the next phase, the channel matrix is recovered at BS with the help of decoder as $f_{De}(.)$. This can be expressed as:

$$\breve{\boldsymbol{h}}_k = f_{De}(\boldsymbol{q}_k) \tag{6}$$

Based on the aforementioned compression and feedback model, the channel matrix at BS can be given as:

$$\breve{\boldsymbol{h}}_k = f_{De}(f_{En}(h_k)) = h_k + \tilde{z} \tag{7}$$

where $\tilde{z}$ denotes the CSI error due to the compression and reconstruction process. The complete setup of digital and analog precoder for downlink massive MIMO system is depicted in below given Fig. 2.

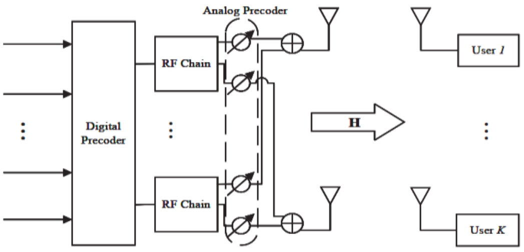

Fig. 2. Digital and analog precoder for downlink massive MIMO system.

## A. Proposed Deep Learning Model for Hybrid Precoding

In massive MIMO systems, full utilization of channel sparsity can help in refining the overall performance of hybrid precoding. In previous segment, we have studied the importance of Deep Learning based framework in mMIMO systems. Therefore, we adopt the deep learning based model to introduce a new deep learning based architecture as precoding framework. The most common deep learning framework, Deep Neural Networks (DNNs), may be compared to a Multiple-Layer Perceptron (MLP). In particular, a DNN has several hidden layers in contrast to a traditional Artificial Neural Network (ANN) to improve its learning and mapping capabilities. Each hidden layer in a DNN has a number of units, and activation functions allow the output to be created depending on the output of these units. For any given argument $p$, the ReLU and Sigmoid can be expressed as $ReLU(p) = \max(0, p)$ and $Sigmoid(p) = \frac{1}{1+e^{-p}}$, respectively. The input and output for this network are denoted as $v$ and $o$ respectively. The mapping operation can be expressed as:

$$\boldsymbol{z} = f(v, w) = f^{(n-1)}\left(f^{(n-2)}\left(\dots f^{1}(v)\right)\right) \tag{8}$$

where $n$ and $w$ represents the number of layers and weights in the network, respectively. The DNN architecture used in this work is presented in Fig. 3.

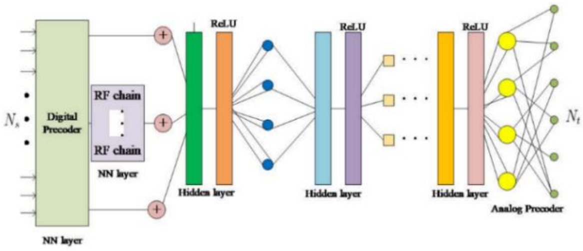

Fig. 3. DNN architecture.

The proposed network architecture consists of input layer, hidden layer and ReLU activation functions. The input layer determines the length of input training sequence. Specifically, the length of input training sequence depends on its dimensions. In this architecture, we have used a fully connected layer with 128 units. These units are utilized to capture the input data features. Later, the two hidden layers are used to process the encoding operation on the data obtained from the FC layer. These two layers have 400 and 256 units respectively to perform the encoding task. In order to consider the distortion in the signal, we use a noise layer which consist of 200 units. This layer is used to distort the original signal by adding the distorted data. The next blocks of hidden layers are used as decoder modules where we use two hidden layers. The first hidden layer comprises of 128 units and the second layer is comprised of 64 units. Finally, the output layer is also applied to obtain the final output signal. Each layer is summarized as follows:

1. **Input Layer**: This layer determines the length of the input training sequence, which is dependent on the dimensions of the input data. It acts as the entry point for data into the neural network.
2. **Fully Connected (FC) Layer**: The FC layer consists of 128 units and is utilized to capture the input data features. Each unit in this layer is connected to every neuron in the previous layer, allowing it to learn complex patterns in the data.

3. **Hidden Layers (Encoding)**:
   - First Hidden Layer: With 400 units, this layer processes the encoding operation on the data obtained from the FC layer. It extracts higher-level features from the input data.
   - Second Hidden Layer: This layer consists of 256 units and further refines the encoded representation of the data, capturing more abstract features.
4. **Noise Layer**: The noise layer consists of 200 units and is used to introduce distortion into the original signal. By adding distorted data, the network learns to handle noise and improve robustness.
5. **Hidden Layers (Decoding)**:
   - First Hidden Layer: Comprised of 128 units, this layer performs decoding operations on the distorted data, aiming to reconstruct the original signal.
   - Second Hidden Layer: With 64 units, this layer further refines the reconstructed signal, helping to recover any lost information due to noise or distortion.
6. **Output Layer**: The output layer is the final layer of the network and is responsible for producing the final output signal.

In order to consider power constraints in output layer, we introduce an activation function as:

$$f(s) = \min(\max(s, 0)\,, N_s) \tag{9}$$

where $N_s$ is the data streams sent by BS to user. In order to map the hybrid coding, we use decomposition method to decompose the mMIMO channel matrix. The channel matrix $H$ is represented as:

$$\begin{gathered} y = WQR^H \\ \boldsymbol{y} = [W_1, W_2]\begin{bmatrix} Q_1 & * \\ 0 & Q_2 \end{bmatrix}\begin{bmatrix} R_1^H \\ R_2^H \end{bmatrix} \end{gathered} \tag{10}$$

here, $W_1 \in \mathbb{C}^{N_r \times N_s}$ and $R_1 \in \mathbb{C}^{N_r \times N_s}$ are regarded as combiner and precoder respectively, $Q_1 \in \mathbb{C}^{N_s \times N_s}$ denotes the upper triangular matrix and $*$ is the arbitrary matrix. In this module, the largest singular values are estimated as $q_{i,i} = (\delta_1, \delta_2, \dots, \delta_N)^{\frac{1}{N_s}} \in \bar{q}, \forall_i$ where $q_{ij}$ denotes the elements of matrix $Q_1$. Therefore, the final received signal is expressed as:

$$\begin{gathered} y = B^H Hx + B^H n \\ W_1^H HR_1 s + W_1^H n \\ Q_1 s + W_1^H n \end{gathered} \tag{11}$$

In order to train the D model, the loss function is expressed as follows:

$$\begin{aligned} loss &= \|R_1 - R_A R_D\|_F \\ &= \sqrt{tr\big((R_1 - R_A R_D)(R_1 - R_A R_D)^H\big)} \\ &= \sqrt{\sum_{i=1}^{\min\{N_t, N_s\}} \delta_i^2 (R_1 - R_A R_D)} \end{aligned} \tag{12}$$

where $\|.\|_F$ is the Frobenius norm, $R_A$ is the analog precoder, and $R_D$ is the digital precoder. Later, we adopt the deep learning framework to construct the autoencoder module which is expressed as follows:

$$R_1 = f(R_A R_D; \Omega) \tag{13}$$

where $f(.)$ Is the mapping relation and $\Omega$ is the dataset samples. The complete deep learning process is as follows:

In order to obtain the structural statistic of massive MIMO model, we adopt deep learning based mapping operation and introduce a training mechanism which uses a certain configuration of deep learning layers. In this process, we initialize the analog and digital precoders as empty matrices and generate the initial random data sequence. The obtained data is used for training the DNN model and precoder matrices $R_A$ and $R_D$ are updated as the training process continues. Furthermore, physical Angle of Arrival (AOA) and Angle of Departure (AOD) are also generated randomly to obtain the bias between $R_1$ and $R_A R_D$ from output layer of DNN with the help of structural features of mMIMO. Therefore, the dataset $\Omega$ with structural features is required. Finally, the loss function is described as follows:

$$\begin{aligned} R_A^{j+1} &= R_A^j + v \\ R_D^{j+1} &= R_D^j + v \end{aligned} \tag{14}$$

here, $v$ denotes the velocity of gradient element, $j$ is gradient $R_A^0$ and $R_D^0$ denotes the randomly generated initial solution for analog and digital precoder. The complete update process can be defined as follows:

$$\begin{aligned} v &= \alpha v - \epsilon g \\ &= \alpha v - \frac{\epsilon 1}{N} \nabla_{R_A, R_D} \sqrt{\sum_{i=1}^{\min\{N_t, N_s\}} \delta_i^2 (R_1 - R_A R_D)} \end{aligned} \tag{15}$$

However, the CSI error due to compression and reconstruction process affects mMIMO performance. To overcome this issue, we have adopted deep learning [2] and introduce a novel deep learning based model for limited feedback. Generally, the downlink mMIMO systems require CSI at the BS to use spatial diversity information. However, the increased overhead degrades the overall spectral efficiency performance. To overcome this issue, we present deep learning based model to compress the channel state matrix. This model uses convolution layers along with the quantization and entropy coding.

### B. Encoder-Decoder Architecture

Fig. 4 depicts the CSI feedback reconstruction where real and imaginary are considered as the channel input as the part of channel matrix. The CSI of user is compressed into a bit stream with the help of local encoder. The encoder model consists of feature encoder model, quantizer block and decoder model. This module helps to extract these key features from the CSI matrix. Further, we use entropy encoder to minimize the feedback amount with the help of arithmetic coding which generates the variable length bit stream. The entropy encoder helps to obtain the lower dimensionality representation of the feature data.

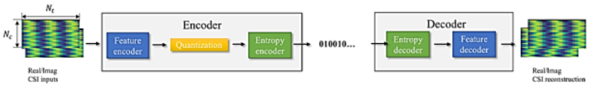

Fig. 4. Proposed architecture for CSI feedback compression.

Below given Figs. 5 and 6 depict the feature encoder and decoder architecture. This architecture uses a convolution layer with 256 kernel where size of each kernel is 9×9. The encoder module is comprised of three convolution layers where first layer uses 9×9 kernel whereas the remaining two layers use kernels of size 5×5. Further, we use downsampling block to reduce the dimensionality followed by a ReLU activation function. Batch normalization is also applied at each layer.

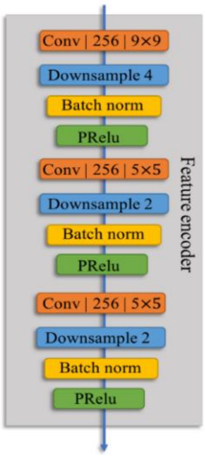

Fig. 5. Feature encoder.

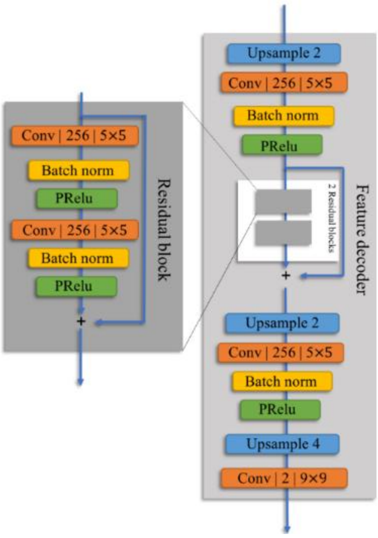

Fig. 6. Feature decoder.

On the other hand, the feature decoder module uses upsampling, convolution, batch normalization and ReLu activation. This module performs the inverse operations as mentioned in encoder module. At BS, the outcome of this entropy decode module is fed into the feature decoder to obtain the channel gain matrix.

Moreover, the decode architecture also consists of residual blocks to skip several layers. This residual block helps to prevent the vanishing gradient problem in the deep learning process. These residual blocks are equipped with the convolutional layer. These layers are activated by applying ReLU activation and they are normalised using the batch norm. Later, quantization and entropy encoding tasks are performed to obtain the final output matrix. The Quantization module helps to quantize each element to the closest integer and the entropy encoder block considers these values as input and converts these quantized bits into streams.

## IV. Result and Discussion

In this section, we present the experimental analysis of proposed approach and compare the achieved performance with traditional precoding techniques in this domain of massive MIMO communication system.

### A. *Dataset Details and Training Process*

In this work, we have used publicly available generic DeepMIMO dataset. The parameters considered for this experiment are presented in Table I.

TABLE I. Simulation Parameters

| Parameter | Value |
|---|---|
| Active Base Station | 4 |
| Active Users | From row R1200 to R1500 |
| BS Antennas Count | $M_x = 1, M_y = 64\,, M_z = 1$ |
| User Antennas Count | $M_x = 1, M_y = 64\,, M_z = 1$ |
| Antenna Spacing (Wavelength) | 0.5 |
| Bandwidth of Antenna | 0.5 GHz |
| OFDM Subcarriers | 1024 |
| Sampling Factor | 1 |

In this experiment, we have considered 4 BSs with the mobile users from row R1200 to R1500. For simplicity, transmitter and receiver modules are considered to employ the 64 antennas. Each antenna has total 3 RF chains. Using the DeepMIMO dataset generator, we first build the channel matrix for each user. Then, a randomly generated noise is added to the current channel matrix. Further, near optimal Gram Schmidt based precoding model is adopted to construct the precoding/combining matrices. The noisy channel and the related RF precoder/combiners codebook indices are then taken into account as a single data point in the dataset. The proposed deep learning based approach, which adopts a Stochastic Gradient Descent (SGD) based loss function, is trained using the produced dataset. The proposed model is implemented by using Keras libraries and it uses Adam optimizer with 0.5 momentum, a batch size of 512 and 0.0005 learning rate.

### B. *Comparative Analysis*

In order to measure the performance of proposed approach, we compute achievable rate for varied transmit power. The obtained performance is presented in below given Fig. 7 where the outcome of Deep Learning based hybrid precoding is compared with the direct precoding. The achievable rate performance is measured for varied total transmit power. The obtained achievable rate performance is compared with the existing approach as mentioned in [49] and upper bound. This experiment shows that the proposed approach is able to achieve near optima solution with reduced overhead as it uses small number of M_t and M_r. The average achievable rate for M_t = 2 and M_r = 2 is obtained as 13.38 and 13.94 by using existing direct precoding and proposed hybrid precoding, similarly, the achievable rate for M_t = 4 and M_r = 4 is obtained as 14.93 and 15.93 by using direct and proposed hybrid precoding, respectively.

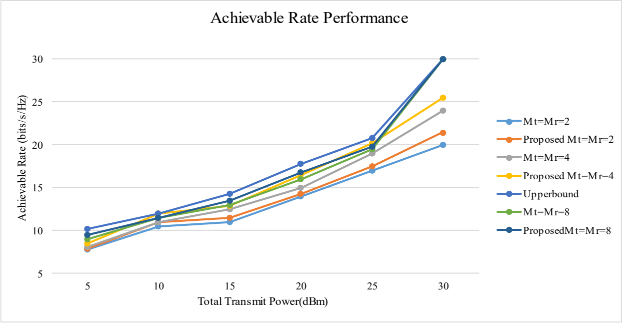

Fig. 7. Achievable rate performance.

Further, we measure the performance of proposed approach in terms of spectral efficiency for varied SNR levels. Below given Fig. 8 depicts the obtained performance and comparative representation of the spectral efficiency. The upper bound is obtained with the help of fully digital precoding by applying ZF with perfect CSI. The increased SNR leads to increase in spectral efficiency.

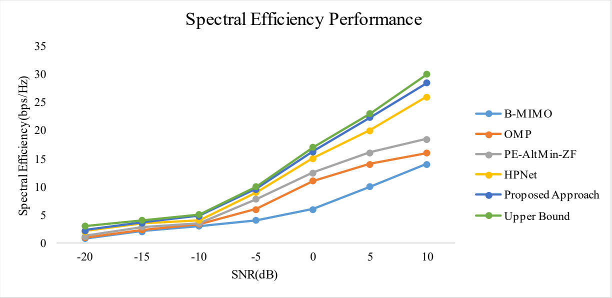

Fig. 8. Spectral efficiency.

The proposed approach achieves the near optimal solution because its performance is not degraded due to stochastic noise increase. The average spectral efficiency is obtained as 5.68, 7.643, 8.914, 11.357, and 12.457 (all values in bits/s/Hz) by using Beamspace MIMO (B-MIMO), Orthogonal Matching Pursuit (OMP), Phase Extraction Alternate Minimization Zero Forcing (PE-AltMin-ZF), Hybrid Precoding Network (HPNet), and Proposed Approach, respectively as mentioned in [50]. The comparative analysis is obtained from Table II.

TABLE II. SPECTRAL EFFICIENCY PERFORMANCE

| SNR in dB | B-MIMO | OMP | PE-AltMin-ZF | HPNet | Proposed Approach | Upper Bound |
|---|---|---|---|---|---|---|
| −20 | 0.8 | 0.9 | 1.2 | 2 | 2.2 | 3 |
| −15 | 2 | 2.3 | 2.8 | 3.5 | 3.6 | 4 |
| −10 | 3 | 3.3 | 3.5 | 4 | 4.8 | 5 |
| −5 | 4 | 6 | 7.8 | 9 | 9.6 | 10 |
| 0 | 6 | 11 | 12.5 | 15 | 16.2 | 17 |
| 5 | 10 | 14 | 16.1 | 20 | 22.3 | 23 |
| 10 | 14 | 16 | 18.5 | 26 | 28.5 | 30 |

Finally, we compare the sum rate performance for varied number of users. Below given Fig. 9 the comparative analysis of sum-rate performance. In this experiment, the average sum rate performance is obtained as 6.042 bits/Hz, 6.4, 7.28, 7.12, 7.68, and 8.02 by using SOMP [43], Two-Stage Hybrid Beamforming (TS-HB) [43], Multi-layer Perceptron (MLP) [43], Low-Resolution Hybrid Beamforming (LRHB) [43], Convolution Neural Network MIMO (CNN-MIMO) [43], and Proposed Approach, respectively. This experiment shows that the increasing number of users degrade the sum rate performance however, proposed limited feedback mechanism helps to maintain the sum rate by reducing the congestion. This comparative analysis is presented in Table III.

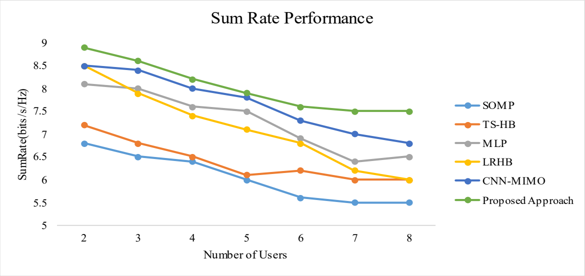

Fig. 9. Sum rate performance.

TABLE III. SUM RATE PERFORMANCE

| No. of Users | SOMP | TS-HB | MLP | LRHB | CNN-MIMO | Proposed Approach |
|---|---|---|---|---|---|---|
| 2 | 6.8 | 7.2 | 8.1 | 8.5 | 8.5 | 8.9 |
| 3 | 6.5 | 6.8 | 8 | 7.89 | 8.4 | 8.6 |
| 4 | 6.4 | 6.5 | 7.6 | 7.4 | 8 | 8.2 |
| 5 | 6 | 6.1 | 7.5 | 7.1 | 7.8 | 7.9 |
| 6 | 5.6 | 6.2 | 6.9 | 6.8 | 7.3 | 7.6 |
| 7 | 5.5 | 6 | 6.4 | 6.2 | 7 | 7.5 |
| 8 | 5.5 | 6 | 6.5 | 6 | 6.8 | 7.5 |

## V. Conclusion

In this work, we have concentrated on massive MIMO communication systems and identified the various issues faced by these systems. The existing studies have reported that the performance of mMIMO systems can be upgraded by applying precoding scheme. Currently, the hybrid precoding schemes have gained attention in this domain. Moreover, deep learning based schemes are also widely adopted to improve the communication performance. Therefore, we present a deep learning based hybrid precoding scheme for mMIMO system. The proposed DNN architecture accepts the channel matrix as input and generates the outputs with the help of analog precoder and combiner. Moreover, we have articulated that the excessive feedback increases the congestion therefore, we introduce encoder and decoder based limited feedback model to improve the overall performance of the system. Similarly, the compression and reconstruction also affect the communication performance at BS. The outcome of proposed approach is compared with traditional schemes in terms of achievable rate for varied transmit power, spectral efficiency, and Sum rate. The comparative investigation illustrates that the proposed DL based methodology realizes better performance when compared with the existing systems.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

Dr. K Ramesha and Shruthi N together initially worked towards identifying the new approach. Shruthi N worked on the DNN framework and understanding the datasets as well as analysis and simulation. Dr. K Ramesha guided Shruthi N on Literature survey and did overall review of the research work. Shruthi N worked on the drafting and formatting of the journal article. All authors had approved the final version.

## References

[1] M. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: A survey," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3109–3132, 2019.

[2] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3027–3032, 2019.

[3] Peter Jonsson *et al.* (November 2023). Mobile data traffic forecast–mobility report. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-traffic-forecast

[4] S. Bhairanatti and S. M. Kumar, "Evolution of 6G era: A brief survey of massive MIMO, mm Wave, NOMA-based 5G and 6G communication protocols, role of deep learning and inherent challenges," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 1, pp. 24–40, 2023.

[5] M. A. Albreem, A. H. A. Habbash, A. M. Abu-Hudrouss, and S. S. Ikki, "Overview of precoding techniques for massive MIMO," *IEEE Access*, vol. 9, pp. 60764–60801, 2021.

[6] R. Chataut and R. Akl, "Massive MIMO systems for 5G and beyond networks—overview, recent trends, challenges, and future research direction," *Sensors* (Switzerland), vol. 20, no. 10, 2020.

[7] S. Ehsanfar, M. Chafii, and G. P. Fettweis, "On UW-based transmission for MIMO multi-carriers with spatial multiplexing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 9, pp. 5875–5890, 2020.

[8] W. Jin, J. Zhang, C. Wen, and S. Jin, "Model-driven deep learning for hybrid precoding in millimeter wave MU-MIMO system," *IEEE Transactions on Communications*, vol. 71, no. 10, pp. 5862–5876, 2023.

[9] E. Bjornson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality–what is next? Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, no. 3–20, 2019.

[10] M. Cho, H. Lee, K. Oh, and J. Kim, "Low complexity hybrid precoding using beam steering for mmWave MIMO systems," in *Proc. IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–5.

[11] R. Ilyas, A. Malik, A. A. Alammari, and M. Sharique, "5G and mmWave MIMO channel models: simulations and analysis," in *Proc. International Conference on Wireless Communications, Signal Processing and Networking*, 2021, pp. 435–440.

[12] T. Q. Duong, X. Chu, and H. A. Suraweera, "Ultra-dense networks for 5G and beyond: Modelling, analysis and applications," *John Wiley & Sons*, 2019.

[13] Z. Zheng and H. Gharavi, "Spectral and energy efficiencies of millimeter wave MIMO with configurable hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5732–5746, 2019.

[14] L. Yan, C. Han, and J. Yuan, "Joint two-level spatial multiplexing and beamforming in terahertz ultra-massive MIMO systems," in *Proc. IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 873–878.

[15] F. Alraddady, I. Ahmed, and F. Habtemicail, "Robust hybrid beam-forming for non-orthogonal multiple access in massive MIMO downlink," *Electronics* (Switzerland), vol. 11, no. 1, 2022.

[16] M. Li, Z. Wang, H. Li, Q. Liu, and L. Zhou, "A hardware-efficient hybrid beamforming solution for mmWave MIMO systems," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 137–143, 2019.

[17] M. J. Shundi, G. G. Mihuba, W. G. Zakayo, and T. Adkhamjon, "Improvements in spectrum sharing towards 5G heterogeneous networks," *International Journal of Electronics and Communication Engineering*, vol. 6, no. 3, pp. 1–9, 2019.

[18] P. Sameera, "Design and analysis of MIMO radar GLRT detector," *SSRG International Journal of Electronics and Communication Engineering (SSRG-IJECE)*, vol. 2, no. 10, pp. 34–38, 2015.

[19] T. K. Tandra, M. S. Anzum, F. Tajrian, and A. Bin Shams, "Downlink performance of beamforming for high mobility users in 5G cellular network," *Journal of Computer and Communications*, vol. 10, no. 8, pp. 104–116, 2022.

[20] Y. Zhao, "Exploitation of robust AoA estimation and low overhead beamforming in mmWave MIMO system," M. E. thesis, Department of Electrical and Computer Engineering, The University of Western Ontario, Ontario, Canada, 2019.

[21] E. Zeydan, O. Dedeoglu, and Y. Turk, "Experimental evaluations of TDD-Based massive MIMO deployment for mobile network operators," *IEEE Access*, vol. 8, pp. 33202–33214, 2020.

[22] M. Boloursaz Mashhadi, and D. Gunduz, "Deep learning for massive MIMO channel state acquisition and feedback," *Journal of the Indian Institute of Science*, vol. 100, no. 2, pp. 369–382, 2020.

[23] T. Kebede, Y. Wondie, and J. Steinbrunn, "Massive MIMO linear precoding techniques performance assessment," in *Proc. 2021 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–8, 2021.

[24] T. Kebede, Y. Wondie, J. Steinbrunn, H. B. Kassa, and K. T. Kornegay, "Precoding and beamforming techniques in mmWave-Massive MIMO: Performance assessment", *IEEE Access*, vol. 10, pp. 16365–16387, 2022.

[25] C. Han, L. Yan, and J. Yuan, "Hybrid beamforming for terahertz wireless communications: challenges, architectures, and open problems," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 198–204, 2021.

[26] M. Vu and A. Paulraj, "MIMO wireless linear precoding," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 86–105, 2007.

[27] X. Wei, L. Xiang, L. Cottatellucci, T. Jiang, and R. Schober, "Cache-Aided massive MIMO: Linear precoding design and

performance analysis," in *Proc. 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[28] S. R. Aghdam, S. Jacobsson, U. Gustavsson, G. Durisi, C. Studer, and T. Eriksson, "Distortion-Aware linear precoding for massive MIMO downlink systems with nonlinear power amplifiers," arXiv preprint, arXiv:2012.13337, 2020.

[29] Y. Liu, J. Liu, Q. Wu, Y. Zhang, and M. Jin, "A near-optimal iterative linear precoding with low complexity for massive MIMO systems," *IEEE Communications Letters*, vol. 23, no. 6, pp. 1105–1108, 2019.

[30] Y. Zhang, P. Mitran, and C. Rosenberg, "Joint resource allocation for linear precoding in downlink massive MIMO systems," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3039–3053, 2021.

[31] V. N. Ha, Z. Abdullah, and G. Eappen *et al.*, "Joint linear precoding and DFT beamforming design for massive MIMO satellite communication," *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1121–1126.

[32] Z. Wang, R. M. Gower, C. Zhang, S. Lyu, Y. Xia, and Y. Huang, "A statistical linear precoding scheme based on random iterative method for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10115–10129, 2022.

[33] A. Krishnamoorthy and R. Schober, "Downlink massive MU-MIMO with successively-regularized zero forcing precoding," *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 114–118, 2022.

[34] W. Jiang and H. D. Schotten, "Deep learning-aided delay-tolerant zero-forcing precoding in cell-free massive MIMO," in *Proc. 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, 2022, pp. 1–5.

[35] L. Kibona, J. Liu, and Y. Liu, "BER analysis using MRT linear precoding technique for massive MIMO under imperfect channel state information," *Photonics & Electromagnetics Research Symposium-Fall (PIERS - Fall)*, pp. 500–506, 2019.

[36] A. Benzin, G. Caire, Y. Shadmi, and A. M. Tulino, "Low-complexity truncated polynomial expansion DL precoders and UL receivers for massive MIMO in correlated channels," *IEEE Trans Wirel Commun*, vol. 18, no. 2, pp. 1069–1084, 2019.

[37] J. Jee, G. Kwon, and H. Park, "Regularized zero-forcing precoder for massive MIMO system with transceiver I/Q imbalances," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1028–1031, 2019.

[38] Z. Zou and A. Dutta, "Capacity achieving by diagonal permutation for MU-MIMO channels," in *Proc. GLOBECOM 2023-2023 IEEE Global Communications Conference*, 2023, pp. 2536–2541.

[39] X. Zeng, S. Fang, Y. Yang, R. Huang, and H. Wang, "A novel transceiver for vector perturbation precoding based on channel correlation matrix," in *Proc. IEEE 19th International Conference on Communication Technology (ICCT)*, 2019, pp. 150–154.

[40] A. Flores, R. C. de Lamare, and B. Clerckx, "Tomlinson-harashima precoding with stream combiners for MU-MIMO with rate-splitting," arXiv preprint, arXiv:2103.08009, 2021

[41] Y. Kabalci, M. Ahmadi, and M. Ali, "Optimal hybrid precoder design for millimeter-wave massive MIMO systems," *Computers and Electrical Engineering*, vol. 99, 107746, 2022.

[42] W. Ma, C. Qi, Z. Zhang, and J. Cheng, "Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 2838–2849, 2020.

[43] A. M. Elbir and A. Papazafeiropoulos, "Hybrid precoding for multi-user millimeter wave massive MIMO systems: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 552–563, 2019.

[44] N. R. G and C. V. Ravikumar, "Developing novel channel estimation and hybrid precoding in millimeter-wave communication system using heuristic-based deep learning," *Energy*, vol. 268, 126600, 2023. doi: 10.1016/j.energy.2022.126600

[45] H. Liu and K. Long, "A deep learning channel estimator for millimeter-wave hybrid massive MIMO systems," *IEEE Wireless Communications Letters*, 2023.

[46] B. Rajarajeswarie and R. Sandanalakshmi, "Intelligent based hybrid precoder for millimetre wave massive MIMO system," *Wireless Networks*, pp. 1–8, 2023.

[47] A. H. Ismail, T. A. Soliman, M. Rihan, and M. I. Dessouky, "Deep learning-based beamforming for millimeter-wave systems using parametric ReLU activation function," *Wireless Personal Communications*, vol. 129, no. 2, pp. 825–836, 2023.

[48] Q. Sun, H. Zhao, J. Wang, and W. Chen, "Deep learning-based joint CSI feedback and hybrid precoding in FDD mmWave massive MIMO systems," *Entropy*, vol. 24, no. 4, 2022.

[49] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 800–805.

[50] M. Chai, S Tang, M Zhao, and W Zhou, "HPNet: A compressed neural network for robust hybrid precoding in multi-user massive mimo systems," in *Proc. GLOBECOM 2020-2020 IEEE Global Communications Conference*, 2020, pp. 1–7.