

AI-Driven Equalization for Compressive Optical Fronthaul in Next-generation Radio Access Network

Rui Chen, Tongyun Li, Michael Crisp and Richard V. Penty

Centre for Photonics Systems, Department of Engineering
University of Cambridge, 9 J J Thomson Ave, Cambridge, CB3 0FA, UK
Email: rc714@cam.ac.uk

Abstract—The increasing demands for data throughput and efficiency in next-generation Radio Access Networks (RAN) necessitate innovative solutions for the optical fronthaul. This work presents an AI-driven equalization framework integrated with digital automatic gain control (DAGC) for compressive optical fronthaul systems. The proposed system uniquely performs joint phase and timing recovery while mitigating channel impairments, noise, and compression-induced signal loss. Leveraging a Convolutional Neural Network (CNN) architecture optimised through iterative pruning and quantisation, the model achieves high sparsity and reduced computational complexity. Experimental results demonstrate significant enhancements in Error Vector Magnitude (EVM) over the conventional technique, enabling efficient support for high-order modulation schemes. By addressing key challenges in lossy compression and signal distortion, this approach ensures robust performance, scalability, and low latency in practical 5G and beyond fronthaul scenarios, making it particularly suited for Open RAN architectures.

Index Terms—radio access networks, optical fiber communication, digital signal processing, neural networks, data compression

I. INTRODUCTION

The rapid evolution of wireless communication systems, especially with the rollout of 5G and future networks, has placed increasing demands on Radio Access Network (RAN) architectures. In these networks, the fronthaul link, as in Fig. 1 connecting the Central Unit (CU), Distributed Unit (DU), and Radio Unit (RU) is critical for efficient operation. Traditional fronthaul links, such as those based on Common Public Radio Interface (CPRI) and its more advanced counterpart evolved CPRI (eCPRI), require significant bandwidth to transmit raw radio signals from the RU to the DU for centralized processing. This bandwidth demand grows significantly with future wireless technologies including higher carrier frequencies, wider bandwidths, and more complex modulation schemes.

To address this fronthaul bandwidth challenge, fronthaul compression has emerged as a key solution to reduce the data rate on the fronthaul while maintaining the integrity of the RF signals [1]. The disaggregation of the RAN architecture further introduces additional connections and links, potentially increasing the infrastructure burden. Compression techniques enhancing the scalability, flexibility, and cost-effectiveness of the RAN can ensure efficient utilization of network resources without compromising performance. However, to meet future wireless network needs, further investigation of efficient compression techniques will be vital to ensuring fronthaul data

demands remain feasible without compromising latency or reliability.

In traditional CPRI-based fronthaul systems, raw I/Q data is transmitted without compression, resulting in high data rates that scale with the number of antenna elements and carriers. For example, a 4×4 MIMO system with a 100 MHz bandwidth can require fronthaul data rates of nearly 25 Gbps for a single RU [2]. This is unsustainable as network operators deploy massive MIMO with many antennas and mmWave technologies with much wider bandwidths. To mitigate this, eCPRI introduces new functional splits with support for compression and more efficient data handling, significantly reducing the required bandwidth by transmitting only the essential data.

Despite the improvements brought by the new functional splits in eCPRI, they also present trade-offs between RU complexity and fronthaul capacity. To maintain an optimal balance, lower-layer splits between 7 and 8 are generally recommended for CU/DU and RU interfaces. Conventional eCPRI/O-RAN compression methods [3] include:

- block floating point (BFP) compression
- block scaling compression
- μ -law compression
- beamspace compression
- modulation compression

Techniques like BFP and block scaling suffer from precision loss due to coarse block-level quantization, making them unsuitable for signals with rapid fluctuations or highly dynamic amplitude ranges. μ -law compression, while effective for low-amplitude signals, introduces significant distortion in high-amplitude signals, leading to degraded Evm, especially in high-order modulations such as 256-QAM. Beam-space compression focuses only on beamforming patterns, which is not suitable for systems without beamforming. Modulation compression currently can only work on downlink as prior knowledge of the modulation format is required [4]. These limitations highlight the need for more dynamic and robust compression techniques capable of maintaining high signal fidelity in diverse and challenging scenarios.

In this paper we show that AI-enabled equalization can mitigate the impact of lossy compression on signal quality. By leveraging machine learning models to compensate for signal distortions introduced during compression and transmission, AI equalization ensures that even highly compressed fronthaul

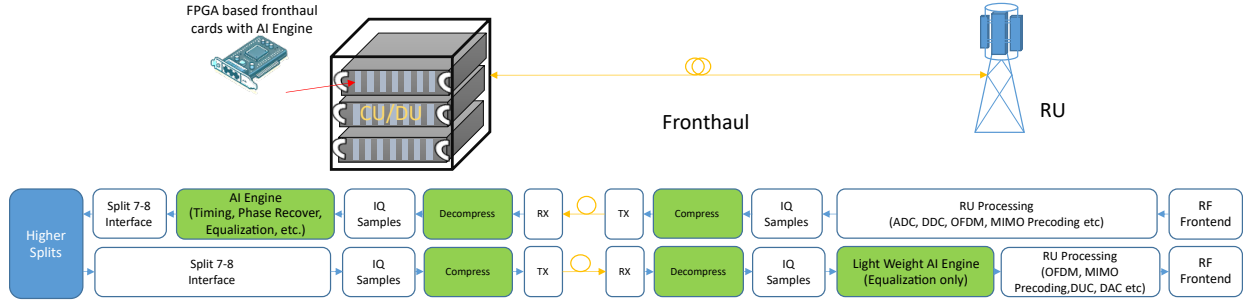


Fig. 1. System Setup of an actual fronthaul. This work focuses on compression and AI equalisation in the uplink, where the signal variation is more dynamic

signals can be accurately reconstructed at the receiver end with minimal degradation.

Our proposed system achieves high compression ratios by combining a digital automatic gain controller (DAGC) enabled data compression with an AI equaliser. One of the key novelties of our system is the development of compression and equalisation algorithms, designed specifically for next-generation fronthaul scenarios. These algorithms incorporate pruning to reduce computational overhead. By selectively pruning redundant neural network weights, the AI algorithm jointly optimizes model complexity as well as RF signal recovery. The pruning strategy ensures that the model remains lightweight and capable of operating in real-time, making it suitable for deployment in practical network environments. Experimental results demonstrate that our algorithm reduces Error Vector Magnitude (EVM) compared to a compression only approach and significantly reduces the complexity of the equalization process through pruning and quantisation. This allows a high compression ratio with minimal signal degradation, showing our AI solution addresses the critical trade-off between compression efficiency and signal fidelity.

By integrating AI-driven equalization with the compressive fronthaul, our system enhances scalability of fronthaul by allowing a greater aggregate RF to be carried on an optical link. This is essential for the future of wireless networks, where distributed and disaggregated RAN architectures will play a pivotal role in delivering ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) services to a growing number of users and devices.

II. SYSTEM SETUP

The experimental setup incorporates two main subsystems: the transmitter (RU) and receiver (CU) designed to emulate the fronthaul uplink under varying power levels. These components are connected by a 500-meter optical fiber link, as shown in Fig. 2. Key components and processes of the setup are detailed below:

A. Transmitter Side

The transmitter side involves signal generation and pre-processing on a PC, where a pseudorandom binary sequence (PRBS) of length 6000 bits is generated. The data is then modulated using a 256-QAM scheme with 8 bits per symbol in

30 M symbols/s (MSpS), achieving high spectral efficiency as defined by 5G NR standards. The uplink scenario is emulated by varying the signal power to evaluate the system's dynamic range performance. It is important to note that orthogonal frequency division multiplexing (OFDM) is not included in this testing system since compression is applied at functional split 7.2, where OFDM demodulation has already taken place. Additionally, OFDM does not impact the compression and equalization methods. In fact, The following methods can be effectively applied to any functional splits between 7 and 8. A fully integrated OFDM system will be explored in future studies.

The resulting signal undergoes filtering through a root-raised cosine (RRC) filter with a 0.1 roll-off factor to minimize inter-symbol interference (ISI). The filtered signal's in-phase and quadrature (I/Q) components are then upconverted to an intermediate frequency (IF) of 37.5 MHz. This signal is produced by a Vector Signal Generator. The IF is chosen to correspond to the center frequency of the first Nyquist zone at a 150 MHz clock rate of the ADC attached to the Tx side of Intel's Statix IV FPGA. The 30 MSpS rate is selected as it is a convenient integer divisor of the clock rate, ensuring accurate symbol detection.

Within the transmitter, the IF signal is sampled at 150MHz and quantised to 16 bit resolution before processing using a digital automatic gain control (DAGC) algorithm, as shown below, specifically designed to optimize signal scaling in a dynamic environment with fast variation of signal powers. This step ensures that the signal is rescaled to fully utilize the 16-bit fixed point representation before being passed to the digital down-conversion (DDC) stage. DDC minimizes the sampling rate to match the signal bandwidth, ensuring efficient processing.

In the lossy compression stage, the least significant bits (LSBs) of the I/Q samples, corresponding to Split 7.2 data, are truncated to reduce the quantization depth, and down sampling is applied to minimize the sampling window, thereby enhancing transmission efficiency. For a 256-QAM input signal, a 2:1 compression ratio is achieved compared to the original Split 7.2 rate. The comparison between uncompressed and compressed data rates is calculated as follows:

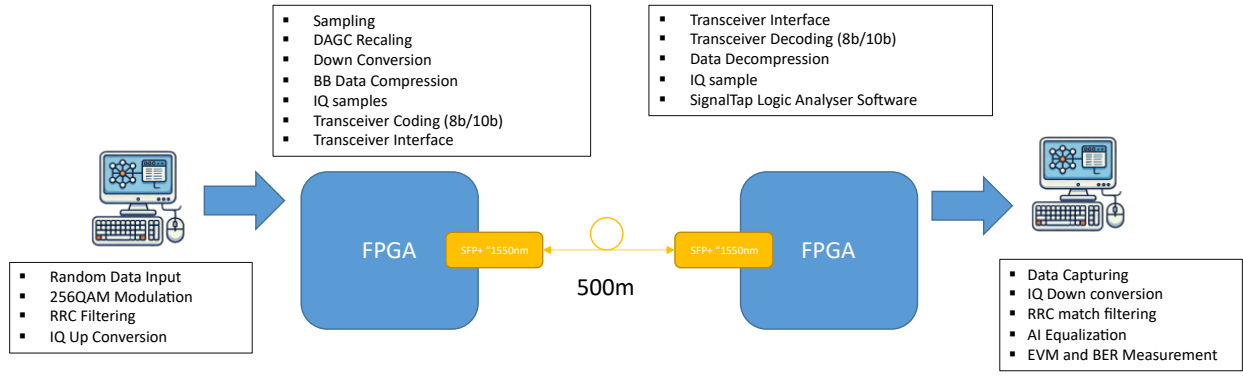


Fig. 2. Experimental setup

Algorithm 1 Dynamic Automatic Gain Control (DAGC)

```

1: function DAGC
2:   Initialise: Gain  $\leftarrow$  initial_gain
3:   Step 1: Measure Signal Power
4:   Power  $\leftarrow$  Data_Input2
5:   Power_Avg  $\leftarrow$  low_pass_filter(Power)
6:   Step 2: Decision Logic
7:   if Power_Avg < thresholds.T_low then
8:     Decision  $\leftarrow$  "Increase Gain"
9:   else if Power_Avg > thresholds.T_high then
10:    Decision  $\leftarrow$  "Decrease Gain"
11:   else
12:     Decision  $\leftarrow$  "No Change"
13:   end if
14:   Step 3: Update Gain
15:   if Decision = "Increase Gain" then
16:     Gain  $\leftarrow$  Gain + Gain_Step
17:   else if Decision = "Decrease Gain" then
18:     Gain  $\leftarrow$  Gain - Gain_Step
19:   end if
20:   Step 4: Adjust Signal Amplitude
21:   Output  $\leftarrow$  Gain  $\times$  Data_Input
22:   Return: Output, Gain
23: end function

```

- Raw data rate without compression:

$$150 \text{ MSps} \times 16 \text{ bits} = 2.4 \text{ Gbps}$$

- Split 7.2 rate with 30MHz useful bandwidth:

$$30 \text{ MSps} \times (16 \text{ bits I} + 16 \text{ bits Q}) = 960 \text{ Mbps}$$

- Compressed data rate:

$$30 \text{ MSps} \times (8 \text{ bits I} + 8 \text{ bits Q}) = 480 \text{ Mbps}$$

Following compression, 8b/10b synchronisation coding is applied to ensure word alignment and assist synchronization before the signal is transmitted over the optical link.

B. Optical Link and Receiver

A small form factor pluggable (SFP+) module at 1550 nm capable of 10 km range is used to convert the electrical signal to optical. After transmission over a 500 m fibre link, chosen for experimental convenience, the processed optical signal is received by another SFP+ module at the receiver end. While the 500 m distance facilitates a controlled experimental setup, the proposed system is designed to support significantly longer transmission distances with low error rates. The architecture ensures robust signal recovery and maintains minimal EVM, indicating that extended fibre links would not compromise system performance. Such scalability aligns with the stringent error performance requirements of next-generation RANs, ensuring feasibility for practical deployments in real-world scenarios.

The received signal at the Central Unit (CU) undergoes multiple stages of processing to ensure accurate signal recovery and analysis. The processing begins with transceiver decoding (8b/10b), which recovers the original transmitted data stream, ensuring proper word alignment and synchronization. Following this, the system decompresses the baseband signal by restoring the least significant bits (LSBs) through zero-padding and applying an RRC matched interpolation filter. This process reverses the compression applied during transmission, effectively reconstructing the I/Q components, which are critical for subsequent processing and analysis. Intel's SignalTap logic analyzer captures intermediate signal stages directly from the FPGA memory, providing valuable insights for debugging and performance evaluation to ensure the system functions as intended.

The data post-processing stage on the PC begins with loading the baseband I/Q samples and performing interpolation by a factor of 5. This is followed by matched filtering using another RRC filter to align the received signal with the input filter. Symbol level synchronisation is achieved through cross-correlation between the transmitted and received signals, accurately identifying the starting point of the signal for proper alignment. Subsequently, an AI-based equalisation scheme is applied, enabling joint phase and timing recovery while mitigating channel impairments and noise. Finally, system

performance is evaluated by calculating key metrics, including EVM and Bit Error Rate (BER), to ensure reliability and accuracy.

The proposed AI-based equalisation scheme is currently implemented offline on a PC. As demonstrated in a later section, after complexity reduction, it is feasible to be implemented on the FPGA for real-time operation. As is typical in digital fronthaul, the EVM performance is limited by ADC/DAC and compression rather than link BER, so we anticipate that with suitable optics the method can be applied over all link spans currently considered for fronthaul.

III. AI-BASED EQUALISATION SCHEME AND EXPERIMENT RESULTS

Convolutional neural networks (CNNs) have been widely used in applications where a combination of feature extraction and classification is needed [5]. Conventional CNNs take a two-dimensional (2D) form and have excelled in applications where 2D signals such as images or video frames are used [6], [7]. In one-dimensional (1D) signal processing applications, 1D CNNs are a preferred choice and have demonstrated superior performance [8], [9]. Besides not requiring a 1D to 2D signal conversion, 1D CNNs can be efficiently trained with a small dataset without overfitting [5]. Moreover, since 1D CNNs require only linear convolutions, they are suitable for real-time, low-cost hardware implementations. In the proposed neural network, the received and synchronised I/Q samples are segmented into overlapping sequences of length 25 (5 symbols, each with an SPS of 5) for input to the model. The resulting input signal takes a shape of $8 \times 25 \times 2$ where 8 is the batch size while 2 is for the I and Q components, respectively. The layer configurations are as below:

- **Three convolutional layers:**
 - 1) 64 filters, kernel size of 7, padding of 3
 - 2) 128 filters, kernel size of 7, padding of 3
 - 3) 256 filters, kernel size of 7, padding of 3
- **One adaptive average pooling layer:** Reduces the output to match the number of symbols
- **Three fully connected layers:**
 - 1) Maps pooled features into a dense representation of 100 units
 - 2) Maps features into a dense representation of 10 units
 - 3) Outputs predictions for I/Q components (size 2)
- **Activation:** the ReLU activation function is applied across all layers except the output layer, which remains linear
- **Loss function:** mean squared error (MSE) loss weighted by the error magnitude to emphasise symbols with large errors.

The entire data flow diagram is shown in Fig. 3, the output is in a shape of $8 \times 5 \times 2$, indicating 8 batches of 5 I and Q symbols. Finally, the EVM is evaluated as a cost function for training, or to analyse the system performance.

The dataset consists of 3000 received symbols and is divided into test and validation sets at an 8:2 ratio. The model is

initially trained at a power level of -55 dBm using the Adam optimiser, with a learning rate of 5×10^{-5} . The initial training takes 30 seconds and 50 epochs to reach convergence on an NVIDIA RTX 3070 laptop GPU. To adapt the model for different power levels, fine-tuning is required. This fine-tuning process is efficient, taking only 5 epochs for each new power level, ensuring the model remains optimised for varying conditions.

In this experiment, the setup in Fig. 2 is used. The transmit power of the signal is varied from -60 dBm to 10 dBm, and the EVM/BER performance of the proposed AI equaliser is compared to a baseline carrier synchroniser [10], as shown in Fig. 4 and 5. This algorithm represents a minimal processing approach commonly employed in digital fronthaul systems, where only basic carrier synchronisation (phase and frequency correction) is performed. The comparison demonstrates the significant improvements enabled by the proposed AI-driven equaliser, particularly in mitigating compression artifacts and maintaining signal fidelity under severe conditions.

IV. COMPLEXITY ANALYSIS

A. Pruning and Quantisation for Complexity Reduction

Pruning is a widely used technique in neural networks to eliminate unnecessary weights, potentially reducing the model's size by over 90% without compromising its accuracy. This approach significantly decreases storage requirements and improves the computational efficiency of neural networks [11]. Its applications range from deploying models on resource-constrained devices to optimising high-complexity networks, such as those used in optical communication and signal processing tasks.

Pruning involves identifying and removing less significant weights or connections in a neural network, resulting in a significantly smaller subnetwork. However, training these smaller subnetworks directly often results in suboptimal performance. Smaller models lack the redundancy and structural robustness of their larger counterparts, which are essential for gradient descent to effectively converge. Instead, starting with a dense, randomly initialised network and training it allows the discovery of smaller, trainable subnetworks, as articulated by the lottery ticket hypothesis. This hypothesis posits that a dense network contains "winning tickets", or subnetworks with fortuitous initialisations that make them easier to train effectively [11]. Attempting to train a small subnetwork independently, without leveraging the initial conditions provided by the dense network, often leads to poor generalisation and degraded performance.

The iterative pruning technique shown in Fig. 6 is employed to achieve high sparsity while maintaining performance. Our approach builds upon the "weight rewinding" methodology, which involves pruning weights iteratively, rewinding the surviving weights to their earlier training values, and retraining [12]. This method has shown promise in uncovering high-performing sparse architectures. Key steps in our pruning implementation include:

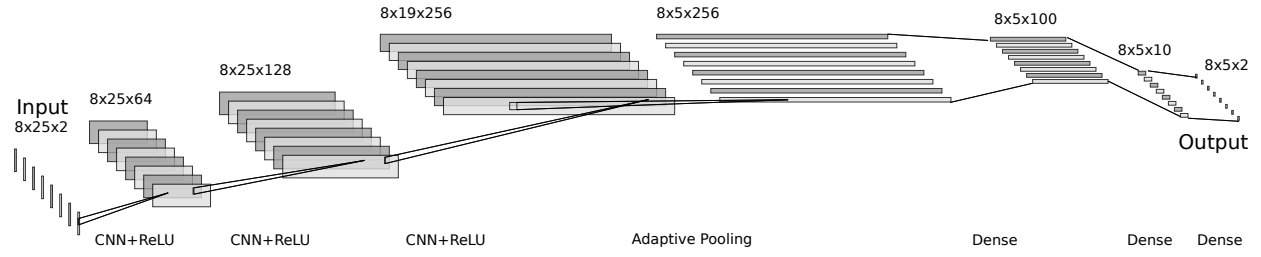


Fig. 3. Data flow diagram of the proposed neural network

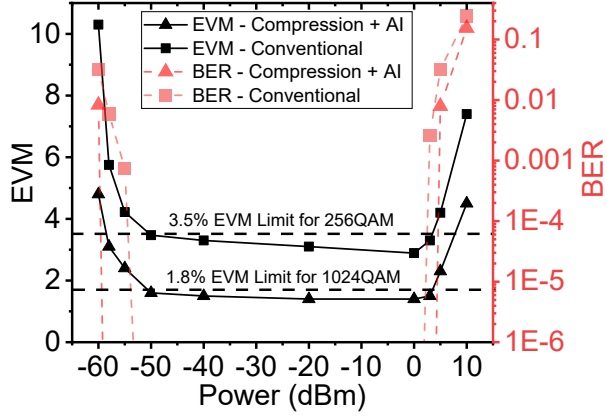


Fig. 4. EVM and BER of the proposed compression scheme with AI receiver (equalisation+timing/phase recovery) vs compression with conventional carrier synchronisation

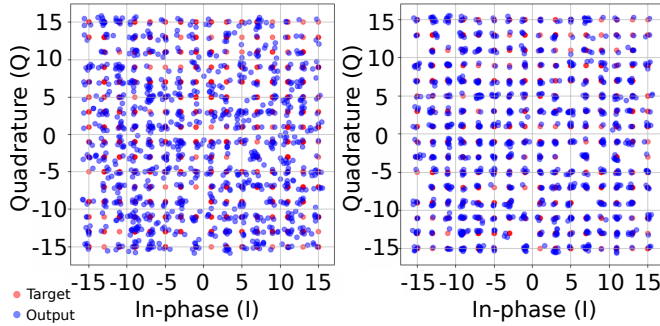


Fig. 5. Constellation diagram (at -58 dBm) before vs. after the proposed AI equaliser (right)

- **Initial Training:** The full network is trained to a certain baseline performance (EVM < 3% in this work).
- **Pruning:** 20% of weights with the smallest magnitudes are pruned in each round.
- **Weight Rewinding:** The remaining weights are reset to their values from an earlier training epoch.
- **Retraining:** The network is retrained, starting with the rewind weights, to recover performance lost in pruning.

This iterative process is repeated until the desired level of sparsity is achieved. In our implementation, we adopt unstructured pruning for both convolutional and fully connected layers

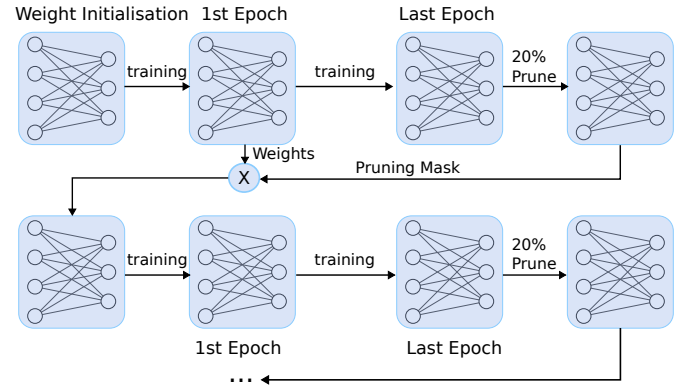


Fig. 6. Diagram of the iterative pruning method

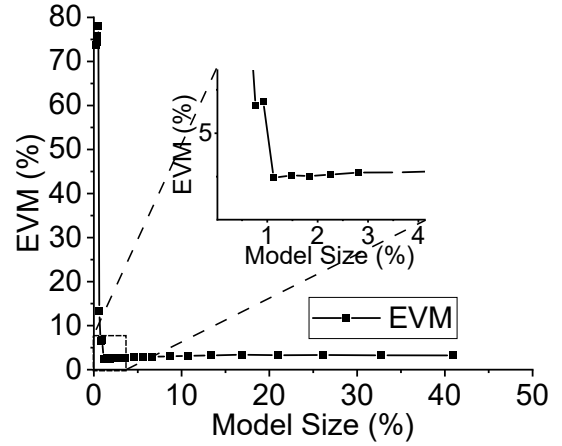


Fig. 7. EVM vs. model size (percentage) in the pruning process, at -55 dBm

[12].

The pruning result for the proposed model is shown in Fig. 7. The model achieves < 3% EVM at only 1% of its original size. After pruning, the model weights are quantised to 16 bit while the EVM number stays the same. Quantisation further improves resource utilisation, improves processing speed and enhances energy efficiency.

B. Complexity Analysis

One commonly used parameter to assess the computational complexity of equalisation models is real multiplications per recovered symbol (RMpS). The RMpS for CNN and fully connected layers can be calculated using the following equations,

considering parallel recovery of all input symbols, as in the proposed approach [12]:

$$C_{CNN} = n_i n_f n_k \times \left(\frac{1 + 2 \times p - d \times (n_k - 1) - 1}{s} + 1 \right)$$

$$C_{dense} = n_{in} \times n_{out}$$

Where n_i is the number of input features (2 in this work for the first layer, for I and Q respectively), n_f is the filter number, n_k is the kernel size, p is padding number, d is dilation (1 for all three CNN layers in this work), s is stride while n_{in} and n_{out} are the input and output feature sizes for the fully connected layer. Substituting in numbers in section 3, the RMpS value is calculated to be 290226, with pruning, the RMpS drops to 2902. At 30 MSps symbol rate, 0.087T 16-bit multiplications per second are required, while a mid-range Intel FPGA can perform 26T 16-bit multiplications per second, suggesting that the model is suitable for FPGA implementations [13].

Our results demonstrate that the iterative pruning approach, combined with weight rewinding and quantisation, can retain model performance even at high sparsity levels. This capability is essential for deploying resource-efficient neural networks in real-time signal processing scenarios.

V. CONCLUSION

This paper introduces an AI-driven equalisation framework tailored for compressive optical fronthaul systems in next-generation Radio Access Networks (RAN). By combining digital automatic gain control (DAGC) and a convolutional neural network (CNN)-based equaliser, the proposed system effectively addresses the critical challenges of bandwidth limitations, signal degradation, and compression-induced distortions.

A key contribution of this work is the robust integration of joint phase and timing recovery directly into the compression pipeline. This innovation ensures accurate signal alignment while mitigating channel impairments, noise, and compression artefacts. In addition, the framework incorporates iterative pruning and quantisation techniques that significantly reduce computational complexity, enabling real-time deployment without compromising performance. The resulting system demonstrates significant improvements in Error Vector Magnitude (EVM) compared to conventional methods, making it highly suitable for practical 5G and beyond fronthaul scenarios.

The results underscore the transformative potential of AI-driven equalisation in enhancing scalability, bandwidth efficiency, and signal fidelity within disaggregated RAN architectures. This work is pivotal for enabling ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) services, addressing the growing demand for robust and efficient connectivity across diverse and dynamic network environments.

Future directions include evaluating the system's adaptability to more diverse splits and modulation schemes including OFDM, real-time implementation of the AI scheme on an

FPGA, optimising hardware implementations for edge devices, and expanding its applicability to emerging Open RAN frameworks. These advancements will further establish AI-driven equalisation as a foundational technology for next-generation RAN.

ACKNOWLEDGEMENT

This work is supported by EPSRC Future Communications Hub in All-spectrum Connectivity (HASC EP/X040569/1).

REFERENCES

- [1] W. Li, A. Chen, T. Li, R. V. Penty, I. H. White, and X. Wang, "Novel digital radio over fiber (drof) system with data compression for neutral-host fronthaul applications," *IEEE Access*, vol. 8, pp. 40680–40691, 2020.
- [2] A. Ericsson, "Cpri specification v7. 0-common public radio interface (cpri); interface specification," 2015.
- [3] O. Alliance, "O-ran working group 4 (open fronthaul interfaces wg) control, user and synchronization plane specification," *O-RAN Alliance, Tech. Rep.*, vol. 10, no. 10, 2022.
- [4] S. Lagén, L. Giupponi, A. Hansson, and X. Gelabert, "Modulation compression in next generation ran: Air interface and fronthaul trade-offs," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 89–95, 2021.
- [5] S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci, and M. Gabbouj, "1-d convolutional neural networks for signal processing applications," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8360–8364.
- [6] S. Kiranyaz, M.-A. Waris, I. Ahmad, R. Hamila, and M. Gabbouj, "Face segmentation in thumbnail images by data-adaptive convolutional segmentation networks," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2306–2310.
- [7] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural computation*, vol. 22, no. 12, pp. 3207–3220, 2010.
- [8] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific eeg classification by 1-d convolutional neural networks," *IEEE transactions on biomedical engineering*, vol. 63, no. 3, pp. 664–675, 2015.
- [9] S. Kiranyaz, A. Gastli, L. Ben-Brahim, N. Al-Emadi, and M. Gabbouj, "Real-time fault detection and identification for mmc using 1-d convolutional neural networks," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8760–8771, 2018.
- [10] M. Rice, *Digital Communications: A Discrete-Time Approach*. Pearson/Prentice Hall, 2009.
- [11] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.
- [12] P. J. Freire, A. Napoli, B. Spinnler, M. Anderson, D. A. Ron, W. Schairer, T. Bex, N. Costa, S. K. Turitsyn, and J. E. Prilepsky, "Reducing computational complexity of neural networks in optical channel equalization: From concepts to implementation," *Journal of Lightwave Technology*, vol. 41, no. 14, pp. 4557–4581, 2023.
- [13] Altera. (2024, mar) Agilx 5 fpgas: Enhanced dsp with ai tensor block. Accessed on 15.01.2025. [Online]. Available: <https://www.intel.com/content/www/us/en/content-details/776602/agilx-5-fpgas-enhanced-dsp-with-ai-tensor-block.html>