WISE: Wireless Analog Computing at Radio Frequency for Disaggregated Deep Learning Inference

Zhihui Gao
Duke University
zhihui.gao@duke.edu

Sri Krishna Vadlamani
MIT
srikv@mit.edu

Kfir Sulimany
MIT
kfir@mit.edu

Dirk Englund
MIT
englund@mit.edu

Tingjun Chen
Duke University
tingjun.chen@duke.edu

Abstract

The emerging deep learning enables various applications on today's edge devices, such as drones, smart wearables, and autonomous vehicles, while their energy- and memory-constrained nature demands efficient computing architectures to support real-time inference. Hereby, we propose WISE, an analog computing architecture at radio frequency that performs deep learning inference between the remote model weights on the central radio, and inputs on the edge. Specifically, WISE is featured by two facts: (*i*) over-the-air model broadcasting enabling simultaneous inference across multiple edge devices, and (*ii*) analog computation of flexible and massive computing scales driven by a single frequency mixer. Extensive experiments on the software-defined testbed demonstrate that the deep learning based on WISE achieves 97.1% classification accuracy on the MNIST dataset with an energy consumption of 31.01 fJ/MAC.

1 Introduction

Recent years have witnessed a substantial rise in the adoption of deep learning (DL) techniques on edge devices to enable diverse intelligent applications, such as Internet-of-Things (IoT), computer vision, and large language models (LLMs) [10, 8, 23, 18, 3]. Many of these edge devices also rely on wireless connectivity (e.g., cellular or wireless local area networks) for control signaling, data transfer, and Internet access. Advanced DL models typically involve a large number of matrix-vector multiplications (MVMs) of the form $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, where edge devices perform inference tasks using pre-stored DL models ($\mathbf{W}$) and locally generated input data ($\mathbf{x}$). However, frequent access to locally stored model weights/inputs of digital MVM computations can be energy-consuming [16], posing significant challenges for energy-, memory-, and compute-constrained edge devices. Moreover, offloading DL inference tasks to or fetching model weights on demand from the cloud requires significant wireless bandwidth and introduces potential privacy concerns [19].

To address these challenges, we present WISE (WIreless Smart Edge networks), the first edge computing architecture designed for energy-efficient complex-valued MVMs via analog computing directly at the radio frequency (RF) [7, 21]. In WISE, a central radio broadcasts RF signals that encode model weights ($\mathbf{W}$) and leverages the shared wireless channel for simultaneous, disaggregated model access across multiple edge clients. Model weights are frequency-encoded to an RF carrier. Each edge client performs inference on local data ($\mathbf{x}$) upon receiving the broadcasted RF signals.

Compared to existing analog computing architectures [2, 12, 1, 17], WISE offers three unique advantages. First, WISE realizes analog computing using passive frequency mixers, namely, *computing*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

mixers, which are ubiquitous RF electronics employed in standard RF front ends of edge devices used for wireless connectivity. Second, we propose a frequency encoding algorithm that allows a single computing mixer to efficiently handle MVM computations with flexible and large scales, up to those required by state-of-the-art large models [20]. Third, we demonstrate that WISE enables truly disaggregated deployment via wireless broadcast of model weights from the central radio, allowing multiple edge devices to perform local DL inference without requiring local model storage.

We evaluate WISE with general-purpose MVM computation and DL model inference on a software-defined testbed. Experimental results demonstrate that WISE achieves an energy consumption of 31.01 fJ/MAC per multiply-and-accumulate (MAC) operation, and 97.1% classification accuracy on the MNIST dataset [11]. This infers to approximately two orders of magnitude improvement compared to state-of-the-art ASICs operating at 1 pJ/MAC [9].

2 System Design

2.1 RF Computing Algorithm

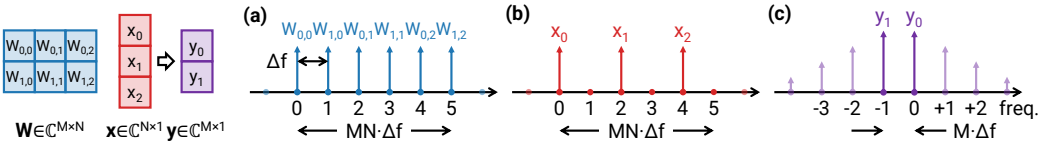

Figure 1: The frequency encoding algorithm for the complex-valued MVM, $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, including the tone mapping for (a) the weight matrix $\mathbf{W}$, (b) the input vector $\mathbf{x}$, and (c) the output vector $\mathbf{y}$.

Without loss of generality, we consider a complex-valued matrix-vector multiplication (MVM) as $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, where $\mathbf{W} \in \mathbb{C}^{N \times M}$ is the weight matrix, $\mathbf{x} \in \mathbb{C}^{N}$ is the input vector, and $\mathbf{y} \in \mathbb{C}^{M}$ is the output vector. To leverage this MVM, the central radio encodes the weight matrix $\mathbf{W}$ into time-domain waveforms $w(t)$, and broadcasts over-the-air. On the edge client, the input vector $\mathbf{x}$ is encoded into a waveform $x(t)$, and is streamed to the computing mixer. The computing mixer wirelessly receives the broadcast $w(t)$, and mixes it with $x(t)$. The output waveform of the computing mixer, $y(t)$, goes through a filter, and is decoded into the output vector $\mathbf{y}$.

Weight waveform encoding. We adopt a frequency encoding scheme to encode the weight matrix $\mathbf{W}$ into time-domain waveforms $w(t)$. In the frequency encoding, the frequency domain signal can be formulated by a series of equally spaced tones. Hereby, we assume the tone spacing of Δf, which requires the shortest time duration of these waveforms as $T = 1/\Delta f$ to distinguish the nearest two tones [5]. As shown in Fig. 1(a), the element on the m-th row and n-th column of $\mathbf{W}$, denoted as $W_{m,n}$, is mapped on a frequency at the $(nM + m)$-th tone at the frequency of $(nM + m) \cdot \Delta f$. Putting all the NM elements in $\mathbf{W}$ together, we have a total of NM tones in the frequency domain, which occupy a bandwidth of $B_w = NM \cdot \Delta f$. The time domain waveform $w(t)$ is expressed by

$$w(t) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} W_{m,n} \cdot e^{j2\pi \cdot (nM+m) \cdot \Delta f t}, \quad \text{where } t = [0, T). \tag{1}$$

Generally, this waveform can be obtained by an NM-point inverse discrete Fourier transform (IDFT). Given a fixed weight matrix $\mathbf{W}$, this IDFT for generating $w(t)$ can be pre-computed and reused. This waveform $w(t)$ is wirelessly broadcast by the central radio to any edge client within the coverage.

Input waveform encoding. The frequency encoding for the input vector is shown in Fig. 1(b), where the n-th element of $\mathbf{x}$, denoted as x_n, is encoded on the nM-th tone at the bandwidth of $B_x = nM \cdot \Delta f$. Besides, other unused tones are set to zeros in the frequency domain. Similarly, the bandwidth of the time domain waveform $x(t)$ is also $NM \cdot \Delta f$. which is obtained by

$$x(t) = \sum_{n=0}^{N-1} x_n \cdot e^{j2\pi \cdot nM \cdot \Delta f t}, \quad \text{where } t = [0, T). \tag{2}$$

It can be proven that this input waveform $x(t)$ is periodic with a period of T/M. Therefore, we can conduct a N-point IDFT for one period, and repeat it for M times to generate the whole waveform $x(t)$. On the edge client, the waveform $x(t)$ is sent to the computing mixer via a wired channel.

Waveform mixing at the computing mixer. The computing mixer essentially performs a time-domain multiplication between the waveforms $w(t)$ and $x(t)$. According to the convolution theorem [15, 5, 7], the time-domain multiplication is equivalent to the frequency-domain convolution. Therefore, the middle M tones of the mixed waveform are proportional to the output vector $\mathbf{y}$, as shown in Fig. 1(c). The output waveform $y(t)$ can be written as

$$y(t) \propto \overline{w(t)} \cdot x(t) = \sum_{m=0}^{M-1} y_m \cdot e^{-j2\pi \cdot m \cdot \Delta f t} + y_{\text{other}}(t), \quad \text{where } t = [0, T), \tag{3}$$

where $y_{\text{other}}(t)$ only contains the tones other than the middle M tones of interest.

Output waveform decoding. On the RX of the edge client, we first remove the unwanted tones $y_{\text{other}}(t)$ by applying a bandpass filter with a passing bandwidth of $B_y = M \cdot \Delta f$. Given the bandwidth B_y after the bandpass filter, the residual waveform can be received by an RX operating at a corresponding low sampling rate of B_y without frequency aliasing. Finally, the output vector $\mathbf{y}$ can be decoded from the output of the RX by an M-point DFT to the frequency domain tones.

2.2 Energy Consumption Analysis

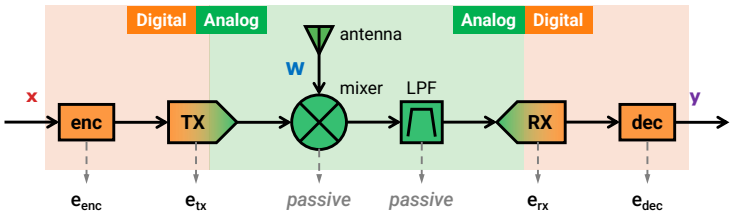

Figure 2: The energy consumption of WISE can be broken down into four terms: e_{enc} for the encoding module, e_{tx} for the TX, e_{rx} for the RX, and e_{dec} for the decoding module, while others are all passive.

The overall diagram of the RF computing algorithm w.r.t. energy consumption, labeled for each module, is shown in Fig. 2. On the edge client, the energy consumption per MVM, E, can be decoupled into four components: (*i*) the input encoding module of the N-point IDFT in digital computing, denoted as E_{enc}; (*ii*) the TX module to transmit the input waveform $x(t)$, denoted as E_{tx}; (*iii*) the RX module to receive the output waveform $y(t)$, denoted as E_{rx}; and (*iv*) the output decoding module of the M-point DFT in digital computing, denoted as E_{dec}. Besides those, other components (the computing mixer, the bandpass filter, etc.) are passive at no energy consumption.

Energy consumption of the encoding and decoding modules. Both the encoding and decoding modules are performed in digital computing. Specifically, an N-point IDFT requires $2N \log_2 N$ real-valued multiply-accumulates (MACs), and symmetrically, an M-point DFT requires $2M \log_2 M$ MACs. Denoting the digital computing energy consumption per MAC as ϵ_{dig}, we have $E_{\text{enc}} = 2N \log_2 N \cdot \epsilon_{\text{dig}}$ and $E_{\text{dec}} = 2M \log_2 M \cdot \epsilon_{\text{dig}}$.

Energy consumption of the TX module. Denoting the transmitting power of the input waveform as P, its total energy consumption over the duration T is given by $P \cdot T$. After passing through the computing mixer and the bandpass filter, the power of the M tones of interest is reduced to $\eta \cdot \frac{1}{N} \cdot P$, where the factor η comes from the hardware efficiency (e.g., the insertion loss of the computing mixer), and $\frac{1}{N}$ is derived from the power ratio of the M tones out of all the tones on the $y(t)$. The thermal noise power within the bandwidth $B_y = M\Delta f$ is given by $M\Delta f \cdot k_b T_0$, where k_b is the Boltzmann constant and T_0 is the room temperature in Kelvin. Therefore, the signal-to-noise ratio (SNR), denoted as γ, of the output waveform $y(t)$ is given by

$$\gamma = \frac{\eta \cdot \frac{1}{N} \cdot P}{M\Delta f \cdot k_b T_0} = \frac{\eta \cdot P}{NM \cdot \Delta f \cdot k_b T_0}. \tag{4}$$

Plugging it into the energy consumption of the TX module, we can rewrite E_{tx} as a function of γ as

$$E_{\text{tx}} = P \cdot T = \frac{NM \cdot \Delta f \cdot k_b T_0 \cdot \gamma}{\eta} \cdot T = NM \cdot \gamma \cdot (\eta)^{-1} k_b T_0. \tag{5}$$

Energy consumption of the RX module. The energy consumption of the RX module is determined by the sampling process of the ADCs. Given the waveform duration of $T = 1/\Delta f$, and the sampling rate of $B_y = M \cdot \Delta f$, the total number of complex-valued samples is $T \cdot B_y = M$. Denoting the energy consumption per real-valued sample as ϵ_{adc}, we have $E_{\text{rx}} = 2M \cdot \epsilon_{\text{adc}}$.

Total energy consumption. Putting them together, we have the energy consumption per MVM as

$$\begin{aligned} E &= E_{\text{enc}} + E_{\text{tx}} + E_{\text{rx}} + E_{\text{dec}} \\ &= \underbrace{2N \log_2 N \cdot \epsilon_{\text{dig}}}_{E_{\text{enc}}} + \underbrace{NM \cdot \gamma \cdot (\eta)^{-1} k_b T_0}_{E_{\text{tx}}} + \underbrace{2M \cdot \epsilon_{\text{adc}}}_{E_{\text{rx}}} + \underbrace{2M \log_2 M \cdot \epsilon_{\text{dig}}}_{E_{\text{dec}}}. \end{aligned} \tag{6}$$

Note that each complex-valued MVM involves NM complex-valued MACs, i.e., $4NM$ real-valued MACs. Therefore, we can derive the energy consumption per real-valued MAC as

$$\begin{aligned} e &= \frac{E}{4NM} = e_{\text{enc}} + e_{\text{tx}} + e_{\text{rx}} + e_{\text{dec}} \\ &= \underbrace{\frac{1}{2M} \log_2 N \cdot \epsilon_{\text{dig}}}_{e_{\text{enc}}} + \underbrace{\frac{1}{4} \cdot \gamma \cdot (\eta)^{-1} k_b T_0}_{e_{\text{tx}}} + \underbrace{\frac{1}{2N} \cdot \epsilon_{\text{adc}}}_{e_{\text{rx}}} + \underbrace{\frac{1}{2N} \log_2 M \cdot \epsilon_{\text{dig}}}_{e_{\text{dec}}}. \end{aligned} \tag{7}$$

To conclude, the energy consumption is impacted by three categories of parameters: (*i*) the MVM scale (N and M), (*ii*) the hardware (ϵ_{dig}, ϵ_{adc} and η), and (*iii*) the SNR for computing accuracy (γ).

3 Evaluation

3.1 Experimental Setup

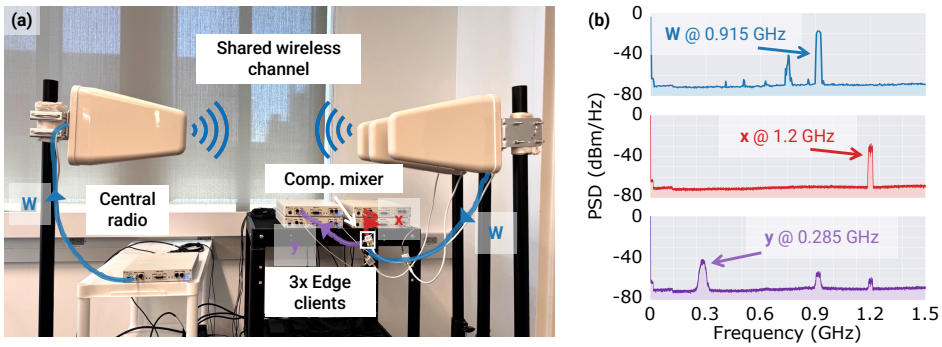

Figure 3: (a) The experimental setup of WISE with a single central radio broadcasting to three edge clients. (b) The spectrum of the input, weight, and output waveforms.

The experimental setup for evaluating WISE is shown in Fig. 3(a), where a central radio is wirelessly supporting the simultaneous RF computing on three identical edge clients.

Central radio setup. The central radio consists of a USRP X310 [6] as the transmitting radio, whose generated weight waveform $w(t)$ is connected to a Tupavco TP514 Yagi directional antenna. The waveform is wirelessly broadcast through the unlicensed industrial, scientific, and medical (ISM) band centered at 0.915 GHz with a bandwidth of 25 MHz, as shown in Fig. 3(b).

Edge client setup. The edge client employs USRP X310 as the radio tranceiver, and ZEM-4300+ [13] as the computing mixer; the encoding and decoding processes are performed on a CPU server. As shown in Fig. 3(b), the input waveform $x(t)$ is transmitted at a center frequency of 1.2 GHz, which is streamed to the computing mixer through a cable, and the weight waveform $w(t)$ is received

by the Yagi antenna; the output waveform $y(t)$ is then received at 0.285 GHz by the USRP X310, whose embedded anti-aliasing filter serves as the bandpass filter. In our evaluation, we average the performance over the three edge clients.

Hardware-related parameter setup. The energy efficiency of WISE is impacted by three hardware parameters: (*i*) the hardware efficiency, which is the combination of the transmitter, the insertion loss of the computing mixer, and the noise figure of the USRP X310 receiver, which is measured as $\eta = 1.48 \times 10^{-4}$ [7], (*ii*) the energy efficiency of the ADC assumed to be $\epsilon_{\text{adc}} = 1$ pJ/sample [14], and (*iii*) the energy efficiency of the digital computing assumed as $\epsilon_{\text{dig}} = 1$ pJ/MAC [9].

3.2 General-purpose MVM computation

We first benchmark WISE's computing accuracy by general-purpose complex-valued MVMs, $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, with randomly generated $\mathbf{x}$ and $\mathbf{W}$. Specifically, we consider $N = M$, and each element x_n in $\mathbf{x}$ and $W_{m,n}$ in $\mathbf{W}$ are independently generated with uniformly distributed amplitudes within $[0, 1]$, and uniformly distributed phases within $[0, 2\pi)$. We evaluate the computing accuracy of WISE by comparing the output vector $\mathbf{y}$ obtained by digital computing with $\widehat{\mathbf{y}}$ by WISE's RF computing. We first define the root mean square error (RMSE) between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ normalized by $\frac{1}{\sqrt{N}}$, which ensures the real and imaginary components of the output elements y_m are mainly distributed within $[-1, +1]$, regardless of the MVM scale N. We further define the resolution bit as $-\log_2(\mathsf{RMSE}/2)$.

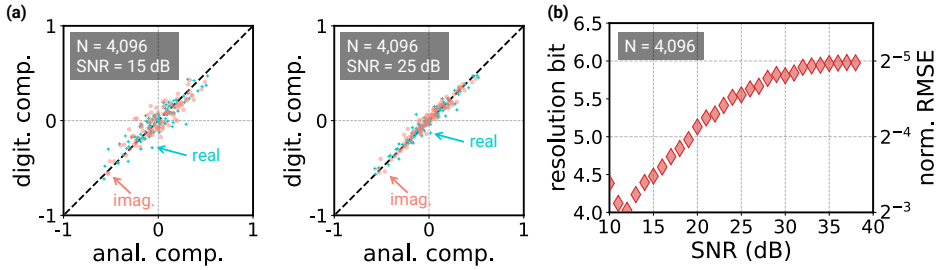

Figure 4: (a) The real and imaginary components between y_m by digital computing and $\widehat{y}_m$ by WISE's analog computing at 15/25 dB SNRs. (b) The RMSE over different SNRs with $N = 4,096$.

Computing accuracy over SNR. The SNR γ on the received $y(t)$ is the key factor affecting the computing accuracy of WISE: a higher γ leads to a higher energy consumption e, while contributing to a better computing accuracy. Fig. 4(a) shows the comparison of the real and imaginary components of y_m by digital computing and $\widehat{y}_m$ by WISE's analog computing, where $N = 4,096$, and $\gamma = 15$ dB and $\gamma = 25$ dB, respectively. Under these two SNRs, the RMSEs are 0.091 and 0.047, corresponding to the resolution bits of 4.46 and 5.40, respectively. Throughout SNRs ranging from 10–40 dB, we show the experimental computing accuracy of WISE with $N = 4,096$ in Fig. 4(b). When the SNR is low (e.g., $\gamma < 25$ dB), the RMSE is approximately reduced by half, i.e., one more resolution bit, when the SNR is increased by 6 dB. When increasing SNR beyond 25 dB, the computing accuracy is gradually saturated with an RMSE of 0.031, i.e., approximately 6-bit resolution, as the hardware imperfection (e.g., the non-linearity of the computing mixer) becomes the bottleneck. Actually, this 6-bit resolution is usually good enough for DL inference [4].

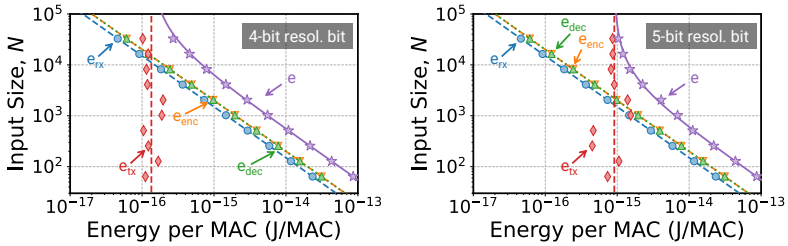

Figure 5: The minimum energy consumption and the breakdown to achieve 4/5-bit resolution over different MVM scales N (or M).

MVM computation scalability. We also examine the scalability of WISE in terms of the MVM scale N (or M). Fig. 5 shows the minimum energy consumption per MAC, e, to achieve the 4-bit or

5-bit resolution over different N, with its breakdown into four terms as $e = e_{\text{enc}} + e_{\text{tx}} + e_{\text{rx}} + e_{\text{dec}}$. Specifically to achieve 4-bit resolution, the total e is reduced from 1.46 fJ/MAC to 0.27 fJ/MAC when increasing N from 4,096 to 32,768, including the reduction of the encoding/decoding energy from $e_{\text{enc}} = e_{\text{dec}} = 0.49$ fJ/MAC to $e_{\text{enc}} = e_{\text{dec}} = 0.06$ fJ/MAC, and the receiving energy e_{rx} from 0.37 fJ/MAC to 0.05 fJ/MAC; on the other hand, the transmitting energy e_{tx} remains almost unchanged around 0.10–0.20 fJ/MAC throughout different N, whose occupancy in the total e is gradually increased from 8.08% to 37.78% with $N = 4,096$ to $N = 32,768$. As for the 5-bit resolution case, the energy consumption per MAC e is 2.26 fJ/MAC and 1.03 fJ/MAC, respectively for $N = 4,096$ and $N = 32,768$, where only the term e_{tx} is increased to around 0.82–1.54 fJ/MAC due to the higher SNR requirement.

3.3 Image classification on the MNIST dataset

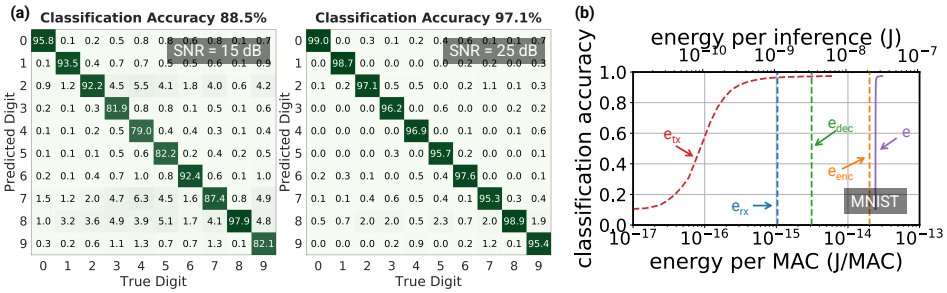

Figure 6: (a) The confusion matrices at 15/25 dB SNRs for on the MNIST dataset. (b) The energy consumption of WISE and the breakdown over classification accuracy.

We also evaluate the performance of WISE on the complex-valued DL models, LeNet-300-100 [11], for the image classification tasks based on the MNIST dataset. Specifically, the images in the MNIST dataset are first flattened to a vector of $N = 784$ as the input; the employed LeNet-300-100 model contains three fully-connected layers with the middle two hidden layers of 300 and 100 neurons. This results in a total of 266,200 complex-valued parameters and 1,064,800 MACs per inference. The model is trained in digital with full-precision weights and activations.

As shown in Fig. 6(a), the classification accuracies are 88.5% and 97.1% with SNRs of 15 dB and 25 dB, compared to that of 98.1% by digital computing. These correspond to an energy consumption of $e = 28.88$ fJ/MAC and $e = 31.01$ fJ/MAC. Fig. 6(b) shows the energy consumption of WISE over classification accuracy. To achieve a classification accuracy of 95%, WISE consumes an energy consumption of $e = 24.79$ fJ/MAC, which is dominated by encoding module with $e_{\text{enc}} = 20.04$ fJ/MAC. This further indicates a total energy consumption of 23.38 nJ per inference.

4 Conclusion

In this paper, we present an analog computing architecture for over-the-air DL inference on edge clients, leveraged by ubiquitous frequency mixers. The proposed analog computing architecture allows the edge devices to access the DL models wirelessly and conduct the computation with no memory and extremely low energy costs. In the future, WISE can be extended to more complicated DL applications, such as convolutional neural networks [11, 10] and transformers [22, 20].

Acknowledgment

Z.G. and T.C. acknowledge partial support from the NSF Athena AI Institute for Edge Computing (CNS-2112562). S.K.V. and D.E. acknowledge support from the DARPA NaPSAC program. K.S. acknowledges the support of the Israeli Council for Higher Education and the Zuckerman STEM Leadership Program. D.E. acknowledges partial support from the NSF EAGER program (ECCS-2419204) and the DARPA QuANET program. The authors thank Marc Bacvanski for the useful discussion and for providing feedback on the manuscript.

References

[1] Stefano Ambrogio, Pritish Narayanan, Atsuya Okazaki, Andrea Fasoli, Charles Mackin, Kohji Hosokawa, Akiyo Nomura, Takeo Yasuda, An Chen, A Friz, et al. An analog-AI chip for energy-efficient speech recognition and transcription. *Nature*, 620(7975):768–775, 2023.

[2] Saumil Bandyopadhyay, Alexander Sludds, Stefan Krastanov, Ryan Hamerly, Nicholas Harris, Darius Bunandar, Matthew Streshinsky, Michael Hochberg, and Dirk Englund. Single-chip photonic deep neural network with forward-only training. *Nature Photonics*, 18(12):1335–1343, 2024.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

[5] Ronald Davis III, Zaijun Chen, Ryan Hamerly, and Dirk Englund. RF-photonic deep learning processor with shannon-limited data movement. *arXiv preprint arXiv:2207.06883v2*, 2024.

[6] Ettus Research. USRP X310 - Ettus Research Product Page, 2024. Accessed: 2025-03-23.

[7] Zhihui Gao, Sri Krishna Vadlamani, Kfir Sulimany, Dirk Englund, and Tingjun Chen. Disaggregated deep learning via in-physics computing at radio frequency. *arXiv preprint arXiv:2504.17752*, 2025.

[8] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[9] Mark Horowitz. Computing's energy problem (and what we can do about it). In *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2014.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] Shi-Yuan Ma, Tianyu Wang, Jérémie Laydevant, Logan G Wright, and Peter L McMahon. Quantum-limited stochastic optical neural networks operating at a few quanta per activation. *Nature Communications*, 16(1):359, 2025.

[13] Mini-Circuits. Coaxial frequency mixer, 300–4300 MHz. `https://www.minicircuits.com/pdfs/ZEM-4300+.pdf`.

[14] Boris Murmann. ADC performance survey (1997-2024). [Online]. Available: `https://github.com/bmurmann/ADC-survey`.

[15] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[16] Flavio Ponzina, Miguel Peon-Quiros, Andreas Burg, and David Atienza. E 2 cnns: Ensembles of convolutional neural networks to improve robustness against memory errors in edge-computing devices. *IEEE Transactions on Computers*, 70(8):1199–1212, 2021.

[17] Guillem Reus-Muns, Kubra Alemdar, Sara Garcia Sanchez, Debashri Roy, and Kaushik R Chowdhury. AirFC: Designing fully connected layers for neural networks with wireless signals. In *Proc. ACM International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, 2023.

[18] Somaiieh Rokhsaritalemi, Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi. A review on mixed reality: Current trends, challenges and prospects. *Applied Sciences*, 10(2):636, 2020.

[19] Kfir Sulimany, Sri Krishna Vadlamani, Ryan Hamerly, Prahlad Iyengar, and Dirk Englund. Quantum-secure multiparty deep learning. *arXiv preprint arXiv:2408.05629*, 2024.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] Sri Krishna Vadlamani, Kfir Sulimany, Zhihui Gao, Tingjun Chen, and Dirk Englund. Machine intelligence on wireless edge networks. *arXiv preprint arXiv:2506.12210*, 2025.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[23] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.