Through the telecom lens: Are all training samples important?

Shruti Bothe*, **Illyyne Saffar**†, **Aurelie Boisbunon**†, **Hasan Farooq***, **Julien Forgeat***, **Md Moin Uddin Chowdhury***

Ericsson Research
Santa Clara, USA* Massy, France†

shruti.bothe, illyyne.saffar, aurelie.boisbunon, hasan.farooq, julien.forgeat, md.moin.uddin.chowdhury@ericsson.com

Abstract

The rise of AI in telecommunications, from optimizing Radio Access Networks to managing user experience, has sharply increased data volumes and training demands. Telecom data is often noisy, high-dimensional, costly to store, process, and label. Despite AI's critical role, standard workflows still assume all training samples contribute equally. On the other hand, next-generation systems require AI models that are accurate, efficient, and sustainable. This paper questions the assumption of equal importance by focusing on applying and analyzing the roles of individual samples in telecom training and assessing whether the proposed method optimizes computation and energy use. We perform sample-level gradient analysis across epochs to identify patterns of influence and redundancy in model learning. Based on this, we propose a sample importance framework that selectively prioritizes impactful data and reduces computation without compromising accuracy. Experiments on three real-world telecom datasets show that our method preserves performance while reducing data needs and computational overhead while advancing the goals of sustainable AI in telecommunications.

1 Introduction

From one generation to the next, leading up to 6G, the telecommunications industry is undergoing a major shift driven by the convergence of AI and next-generation network design, with ML models increasingly deployed for critical functions such as user traffic prediction, beamforming optimization, anomaly detection, and intelligent handover control [1]. Although these models improve automation and efficiency, their deployment brings significant challenges, including high demands for data, compute, and training time.

The telecommunications industry is increasingly focused on energy efficiency as operational costs and environmental impact become critical considerations [2]. Dense 5G and future 6G networks consume substantial energy across base stations and core processing [3], making reductions essential for lowering expenses and carbon footprint [4]. Most efforts target specific use cases (e.g., base-station optimization or transmission protocols), while far fewer address the dual challenge of optimizing both AI training and application-level energy. This paper addresses that gap with an integrated strategy that considers AI model efficiency alongside network operational energy.

Meanwhile, the AI community has also recognized the environmental cost of large-scale training. Initiatives like *Green AI* [5] call for models that are not only accurate, but also compute- and energy-efficient. In telecommunications, this challenge is compounded by the sheer scale of networks, the

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

high cost of operations, the dynamic nature of traffic, and the need to retrain frequently. Thus, a key question emerges:

"Are all training samples equally valuable in the context of telecom model training, or can we train smarter by focusing on the most impactful data?"

A notable assumption in ML pipelines is that all training samples are equally important for model convergence and generalization. This assumption originates from traditional domains such as image classification or Natural Language Processing (NLP), where data distributions are often static and well-curated [6]. In contrast, telecom data is fundamentally different: it is non-stationary, domain-specific, and frequently reflects rare but critical events. Moreover, many ML pipelines in telecom must operate under real-time or near-real-time constraints, making data curation, training efficiency, and sustainability major practical concerns.

This paper takes a gradient-centric view of sample utility in model training with a focus on telecom applications. By analyzing how the gradients of individual samples evolve over training epochs, we categorize samples into those that are critical, redundant, or even detrimental to performance. Our central insight, through experimentation is that a significant portion of the training set offers marginal returns in terms of model improvement that suggests an opportunity for selective training that is both faster and more resource-efficient. To implement this notion, we introduce a sample importance framework. This model dynamically scores and filters samples during training based on their gradient norms, which serve as a proxy for influence on the loss incurred. In standard ML pipelines, the full dataset is used indiscriminately through randomized mini-batching, shuffling without regard to temporal structure, sequentially iterating over all samples in fixed-size batches, oversampling already well-represented periods, or repeatedly training on stable low-variance intervals. This approach focuses instead on selectively training and prioritizing samples that drive meaningful parameter updates. We validate this approach on multiple telecom datasets and show that models trained using only a fraction of the data can achieve comparable performance to those trained on the full dataset while reducing training time and compute cost.

The key contributions of this paper are: (i) an analysis of per-sample gradient dynamics across training epochs using a real-world open telecom dataset, showing that depending on the dataset size, the influence of a significant portion of training data contributes minimally to loss reduction or generalization; (ii) the proposition of a lightweight, model-agnostic mechanism to estimate sample importance during training based on gradient norm behavior that dynamically filters or re-weights data based on gradient importance. Unlike influence functions or forgetting metrics, our approach does not require retraining or access to ground truth labels; and (iii) Our framework supports goals of energy-efficient AI development by reducing redundant computation and data usage during training, which is especially relevant for large-scale network operators seeking to deploy AI models across distributed infrastructure.

2 Related Works

Research in core-set selection [7] and influence functions [8] have explored ways to identify the most impactful samples, but these methods often come with high computational costs. Influence functions, for instance, measure how changing a sample's weight affects model performance without retraining, yet their applicability to deep models is limited by convexity assumptions and the expense of computing Hessian inverses [9]. [10] proposed tracking "forgetting events" to reveal persistently misclassified or unstable samples, offering insights into data difficulty or noise. However, most of this work focuses on vision and NLP benchmarks, with little attention to structured, operationally constrained telecom data.

Other approaches, like curriculum learning [11], suggest ordering training samples by difficulty to speed learning. While effective in some domains, defining "difficulty" in telecom is challenging due to noisy or missing labels. Bothe et all [12] used curriculum learning to order training samples, but their work uses the full dataset for training. Meanwhile, AI in telecommunications has expanded rapidly, powering applications from network traffic forecasting [13] to load balancing, radio access network (RAN) tuning, and fault prediction. Yet, model training is often treated as a black box, with little examination of how training data characteristics affect reliability, retraining frequency, or energy use. As networks shift toward dynamic architectures like Open-RAN and edge-cloud splits, models must adapt to continuously evolving data streams by making efficient retraining not just beneficial,

but necessary. Environmental costs are also rising. Recent studies [14, 5, 15] warn of the carbon footprint of large-scale deep learning, but domain-specific efficiency strategies for telecom remain rare. This is especially important given that telecom data is costly to collect, privacy-sensitive, and often subject to real-time constraints.

Our work addresses these challenges with a lightweight, gradient-based sample importance metric that directly ties to operational KPIs. By allowing the model's own learning dynamics to prioritize training data, we reduce computational overhead without compromising accuracy. This aligns AI efficiency with telecom's operational, financial, and sustainability goals by making retraining faster, greener, and better suited to the demands of large-scale telecom networks.

3 Problem Statement and proposed approach

Let $X \in \mathbb{R}^{t \times d}$ be a time series of length t with d features, and Y be the target to predict. In the case of forecasting, $Y \in \mathbb{R}^{t' \times d}$ represents the same time series as X with a given delay, which can have the same length t or a different one. In the case of a classification problem, $Y \in \{1, \dots, c\}$ represents the class of the time series, such as the type of activity of a signal (SMS, call, internet), where c is the number of classes. The dataset $(x_i, y_i)_{i=1}^n$ denotes the collection of observations of (X, Y) of size n.

Let f be the prediction model mapping X to Y, and $\mathcal{L} : (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$ be a loss function based on the predictions errors (e.g., the mean squared error in regression, or the cross-entropy in classification). We focus in this study on deep neural network architectures f_θ, where $\theta \in \Theta$ denotes the weights of the neural network that are optimized through $\mathcal{L}$: $f_{\theta^*} = \arg\min_{\theta \in \Theta} \mathcal{L}(X, Y)$.

3.1 Gradient Norm Computation for Importance score

To identify important samples, we track per-sample gradient norms across all epochs. At epoch e and for sample s, the gradient norm is computed as:

$$g_{e,s} = \sqrt{\sum_{j=1}^{P} \left\| \frac{\partial \mathcal{L}_{e,s}}{\partial \theta_j} \right\|_2^2}, \tag{1}$$

where θ_j denotes the jth trainable parameter in the network and P is the total number of parameters. These norms are stored in a matrix $\mathbf{G} \in \mathbb{R}^{E \times N}$, where E is the number of epochs. The *importance score* of a sample s is then defined as:

$$\mathcal{I}(s) = \frac{1}{E} \sum_{e=1}^{E} g_{e,s}. \tag{2}$$

3.2 Important Sample Selection

The goal of sample selection is to find a metric that orders the samples according to their importance, and to use that metric to select the most relevant ones for training the model. In this work, the metric we consider is the importance score based on the gradient, defined in (2).

The selection of the most $p\%$ important samples thus corresponds to the set:

$$S_p^* = \underset{S \subset D; |S| \leq k}{\arg\max} \sum_{s \in S} \mathcal{I}(s), \tag{3}$$

where $|S|$ denotes the cardinality of S and $k = \left\lceil \frac{p}{100} \times n \right\rceil$ is the number of samples corresponding to $p\%$.

4 Experiments and Results

4.1 Datasets and pre-processing

1. **Telecom Italia Big Data Challenge**: We use the publicly available *Telecom Italia* dataset [16], which provides anonymized measurements of SMS, call and Internet activity in the

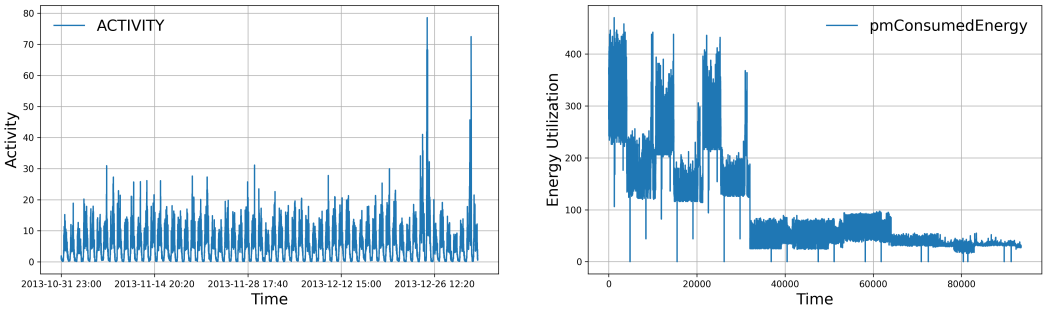
city of Milan. For this work, we select the "Internet_Activity" feature as both the input and output in a univariate forecasting task. An example of the shape of the univariate time series is shown in Figure 1a. Data is normalized in the range [0, 1] to stabilize neural network training and avoid scale-induced gradient instability. The dataset is chronologically split 80-20 into training and testing subsets. No shuffling is applied to preserve temporal order and prevent look-ahead bias. The supervised learning representation is constructed using a lag-1 sliding window:

$$X_t = [x_t], \quad y_t = x_{t+1} \tag{4}$$

where x_t denotes the Internet activity at time t. The resulting samples are reshaped to $(n_{\text{samples}}, n_{\text{timesteps}}, n_{\text{features}})$ format, where $n_{\text{timesteps}} = 1$ and $n_{\text{features}} = 1$.

2. **Proprietary data from a Telecom Vendor**: The second dataset consists of multivariate time series data collected every 15 minutes from 249 base station sites, covering both LTE and 5G New Radio (NR) technologies. It includes performance counters related to key KPIs such as physical resource block utilization, data volume, number of connected users, active sessions, throughput, and signaling overhead, with measurements from both uplink and downlink. Performance counters at the cell level are aggregated to match energy consumption data measured at the physical hardware level, requiring data pre-processing and mapping between cells and radio units. When multiple cells share a radio unit, their performance counters are aggregated accordingly and combined with corresponding energy consumption values, as shown in Figure 1b. The dataset focuses on a subset of sites selected from a European city center and 50 neighboring sites within roughly five kilometers.

3. **5G Beam Selection data**: The 5G Beam Selection dataset [17] is a synthetic 5G mmWave MIMO dataset generated by combining Simulation of Urban Mobility (SUMO) traffic simulator [18] with the Wireless InSite® ray-tracing tool to capture realistic vehicular mobility and wireless propagation in an urban canyon in Rosslyn, Virginia, where a roadside unit transmits to 10 moving vehicles, producing 116 episodes with 50 time-sampled scenes each. For every scene, detailed ray-level channel information is recorded, including complex gains, power, angles of arrival/departure, time of arrival, and interaction types (line-of-sight, reflections, scattering), along with a 2D occupancy grid that encodes vehicle and receiver positions.



(a) Telecom Italia

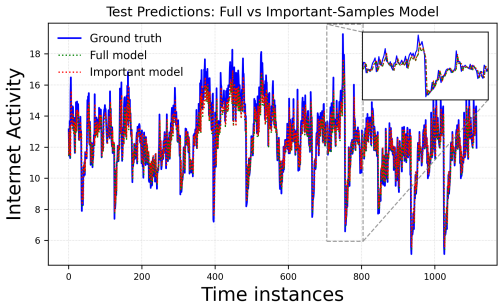
(b) Vendor Data

Figure 1: Snapshot of Internet Activity and Energy Consumption of a Base-station

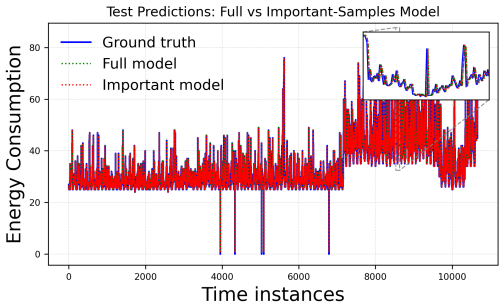
4.2 Results

The objective of this study is to investigate whether ML models can maintain comparable predictive performance when trained on a reduced set of *important samples*, selected using a gradient-based influence metric, compared to training on the full dataset. We have chosen a long-short term memory (LSTM) neural network as a benchmark for the first two datasets to compare with prior art implementation [19]. For the third dataset we use the implementation and baseline from [17]. We sort training samples by their importance scores in descending order and select the top $p\%$ for retraining, where $p \in \{10, 20, \ldots, 90\}$. For each selected subset, the LSTM is reinitialized and trained solely on the selected samples and the MAE and training time are recorded. Results are compared to the full model baseline, which is compared to prior published art.

Figure 2 presents the test prediction plots, comparing the baseline with proposed model predictions for the first two datasets. In each case, the ground truth (blue), best important samples model (red), and full model (green) are plotted over time, showing how closely the proposed model tracks real-world internet activity. Insets in the plots provide zoomed views of regions with high activity variance, illustrating fine-grained prediction. It turns out that the full model and the one based on the best important samples are almost super-imposed. This framework also captures extreme drift in the Vendor data.



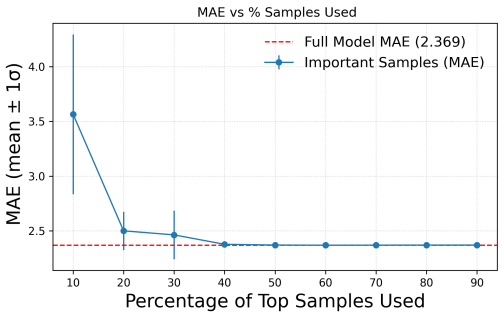
(a) Telecom Italia



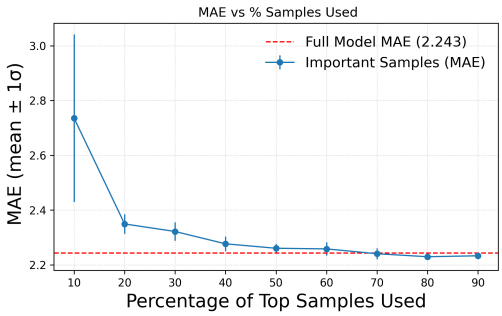
(b) Vendor Data

Figure 2: Test predictions comparing the baseline model and the sample important model

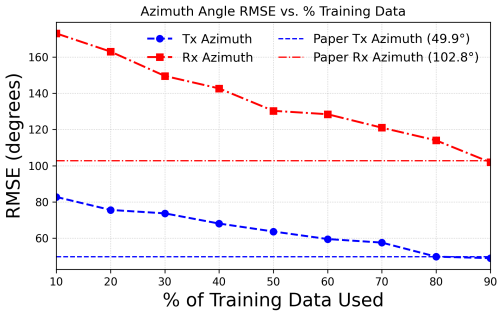
Figure 3a and Figure 3b show the Mean Absolute Error (MAE) across five runs as a function of the percentage of top-ranked training samples used. Figure 3c and Figure 3d show the Root Mean Squared Error (RMSE) difference in azimuth and elevation angles in degrees. The curves demonstrate that the proposed sample selection method achieves accuracy comparable to the full model once a sufficient proportion of informative samples is included, while requiring fewer samples. This leads to notable computational and energy efficiency gains without significant accuracy loss.



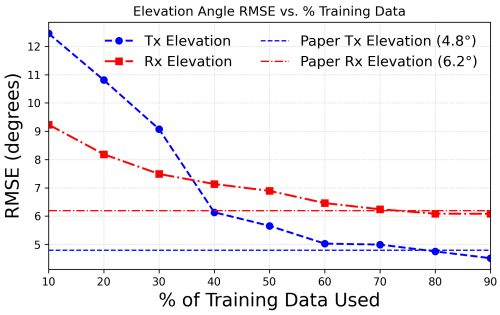
(a) Telecom Italia



(b) Vendor data



(c) 5G Beam Selection(Azimuth Angle)



(d) 5G Beam Selection (Elevation Angle)

Figure 3: Model performance of sample importance framework as compared to baseline models.

Table 1: Model Performance and Training Time Improvements for Different Telecom Italia Dataset Sizes

Dataset Size (Samples)	Percentage of Samples Used (%)	Model Performance Improvement (MAE)	Training Time Improvement (seconds)
5K	80	2.7130	3.0697
50K	70	21.0284	15.5542
300K	65	28.3479	21.0284

For the Telecom Italia dataset, comparable performance is reached with 68% of important samples, reducing the training data by 28%. For the Vendor dataset, 74% suffices to match the baseline, giving a 23% reduction. In the 5G beam selection dataset, the method outperforms the baseline on elevation angles using 75% of training samples, and matches the performance on azimuth angles with 90% data. Dataset reduction is defined as the gap between the total training data and the subset of important samples selected via gradient-based filtering. Thus, using 78% of important samples does not imply a 78% reduction, but rather reflects the proportion of the selected subset used. For statistical significance, results are averaged over $N_{\text{RUNS}} = 5$ independent runs, with error bars denoting mean ± 1σ. A bootstrap 95% confidence interval (CI) is reported for the improvement in MAE between the important-sample method and the full model, ensuring robustness without assuming strict normality.

Finally, Table 1 summarizes the improvement in model performance, the percentage of time saved, and the proportion of data required to achieve these results with varying dataset sizes, the latter being selected from the best improvements from Figure 3. It turns out that this optimal proportion of data decreases with the model size, and thus so do its improvements in time and accuracy, showing that larger datasets tend to exhibit redundancy.

We monitored the carbon emissions of our experiments using CodeCarbon [20] across the Telecom Italia, Vendor Data, and 5G Beam Selection datasets to assess whether strategically reducing training samples influences energy usage and contributes to broader sustainability goals. Emissions were measured and averaged across multiple runs for each dataset. The important-samples model consistently lowers emissions versus the full model. The Telecom Italia dataset drops from 2.1066×10^{-6} kg to 1.3031×10^{-6} kg ($-38.14\,\%$), Vendor Data from 2.0734×10^{-6} kg to 1.2666×10^{-6} kg ($-38.91\,\%$), and 5G Beam Selection from 1.9918×10^{-6} kg to 1.6927×10^{-6} kg ($-15.02\,\%$). The findings indicate an average reduction of approximately 30.69%, demonstrating that sample importance not only enhances computational efficiency but also delivers measurable sustainability benefits in large-scale data analysis workflows.

5 Conclusion and future works

To the best of our knowledge, this is the first study to apply an important-sample–based training strategy in the telecommunication domain. Unlike typical machine learning datasets, telecom data is tabular, time-ordered, and shaped by both seasonal patterns and sudden anomalies. These behaviors appear at fine temporal resolutions, yet operational constraints make it impractical to maintain separate models for different times of day. Hence, adaptive methods that generalize across traffic conditions without full retraining are needed. Our approach addresses this by selectively training on the most impactful data points. This model-agnostic strategy provides a resource-efficient and sustainable way to process large-scale telecom datasets. Evaluations on LSTM-based models show that while small datasets yield limited gains, larger datasets benefit substantially in training efficiency, reduced computation time, and sustained or even improved accuracy. The results confirm that training on carefully chosen samples is more effective than using all data indiscriminately and offer guidance for balancing accuracy against efficiency when slight performance trade-offs are acceptable.

This work represents a conceptual rather than algorithmic contribution, focusing on empirical insight and practical utility over theoretical development. Future research could extend this direction by linking gradient-norm statistics to generalization behavior, incorporating dynamic or curriculum-style selection mechanisms, and benchmarking against more advanced core-set or influence-based approaches under comparable computational budgets. Further validation across diverse model architectures and telecom tasks would also help assess generality. Finally, translating the observed

trade-offs into operational guidelines—such as how to select the optimal data fraction under given latency or cost constraints—remains an open area for exploration.

References

[1] Pradnyawant M Gote, Praveen Kumar, Prateek Verma, Prajyot Yesankar, Adesh Pawar, and Saniya Saratkar. From 5G to 6G: The role of AI, machine learning, and deep learning in wireless systems. In *Conference on Sentiment Analysis and Deep Learning (ICSADL)*, 2025.

[2] Jaewon Lee, Hyunseok Ryu, Younsun Kim, Youngbum Kim, and Junyung Yi. TREES: Toward a Real Energy Efficient System for 6G Communications. *IEEE Wireless Communications*, 32(1), 2025.

[3] GSMA. Energy efficiency: An overview, 2019. Wednesday May 8, 2019.

[4] Lopamudra Kundu, Xingqin Lin, and Rajesh Gadiyar. Toward energy efficient ran: From industry standards to trending practice. *IEEE Wireless Communications*, 32(1):36–43, 2025.

[5] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

[6] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS 2007)*. 2008.

[7] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

[8] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning (ICML)*, 2017.

[9] Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

[10] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[11] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International conference on machine learning (ICML)*, 2009.

[12] Shruti Bothe, Hasan Farooq, Julien Forgeat, and Kristijonas Cyras. Time-series prediction using nature-inspired small models and curriculum learning. In *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–6, 2023.

[13] Ai Chen, John Law, and Mikhail Aibin. A survey on traffic prediction techniques using artificial intelligence for communication networks. *Telecom*, 2(4):467–498, 2021.

[14] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *AAAI conference on artificial intelligence (AAAI)*, 2020.

[15] Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, page 85–99, 2024.

[16] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data*, 2(1):1–15, 2015.

[17] Aldebaro Klautau, Pedro Batista, Nuria Gonzalez-Prelcic, Yuyang Wang, and Robert W. Heath Jr. 5G MIMO data for machine learning: Application to beam-selection using deep learning. In *2018 Information Theory and Applications Workshop, San Diego*, pages 1–1, 2018.

[18] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018.

[19] Shan Jaffry. Network traffic prediction project. `https://github.com/ShanJaffry/Network_traffic_Prediction`, 2020. Accessed: 2025-08-28.

[20] Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stęchły, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper addresses the shortcomining of training an ML with the entire dataset without knowing the importance or contribution of the traninig samples, this is important for efficient model training and is accurately reflected in the abstract, intro and paper's contributions

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper addresses the limitations in the ways: by having a future work section, that will determine the postential expansion of this work, that is not discussed in this paper

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper has enough details to peproduce the architecture described

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This work includes data generated by propeitary simulators and will not be made public. This architecture is also tested on an open source data than can be reproduced from the information in the Experimental setup section of the manuscript

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the relevant information is presented in the experimental setup section of the manuscript

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: There is sufficient information provided for reproducibility

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper addresses societial impact in the form for resource utilization in terms of energy-efficiency to run large scale AI models

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: [NA] .

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Open source dataset has been cited

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM has not been used in any capacity

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.