Master Thesis Project Proposal

Tokenization as a Neural Compression Strategy in Automotive Embedded Systems

Tim Boleslawsky, gusbolesti@student.gu.se

Emrik Dunvald, gusdunvem@student.gu.se

September 2025

Suggested Supervisor at CSE: Yinan Yu

Suggested Supervisor at Company: Dhasarathy Parthasarathy

Relevant completed courses student 1:

- …

Relevant completed courses student 2:

- …

1 Introduction

- Modern vehicles generate vast amounts of data from multi-modal sensors such as cameras, radar, LiDAR, and in-vehicle networks (IVNs) like CAN and LIN networks.

- Legacy IVNs such as Classical CAN (1 Mbit/s) and LIN (20 Kbit/s) were never designed for continuous high-bandwidth streams.

- To avoid overload, event-triggered or selective logging schemes are used.

- These reduce bandwidth but limit observability and introduce sampling bias, degrading downstream machine learning (ML) performance.

- The proliferation of ADAS and intelligent systems further multiplies data quantity and complexity.

- Hence, there is a pressing need for adaptive and ML-aware logging frameworks that preserve informational value while respecting resource constraints.

2 Context

The In-Vehicle Embedded System: An in-vehicle embedded system is a specialized computer system integrated within a vehicle to perform dedicated functions, often in real time, and is essential for controlling, monitoring, and enhancing various automotive operations. These systems typically consist of both hardware and software components, such as electronic control units (ECUs), sensors, actuators, and communication interfaces, which are responsible for tasks like engine management, safety features, infotainment, and advanced driver assistance systems [Navet and Simonot-Lion, 2017, Fairley, 2019].

In-Vehicle Networks and Event-Triggered Logging: Modern vehicles may contain dozens or even hundreds of these embedded systems, interconnected through in-vehicle networks (e.g., CAN, LIN, FlexRay, Ethernet), enabling efficient communication, and coordination among different vehicle subsystems [Bello et al., 2019, Navet and Simonot-Lion, 2017, Fairley, 2019]. Event-triggered logging and diagnostic frameworks, which record data only when anomalies or threshold crossings occur, are often adopted to reduce data transmission and avoid bus saturation in complex systems, such as the in-vehicles embedded system. However, this selective approach can reduce holistic visibility of system health, as it may miss subtle degradation patterns or early warning signs that do not cross predefined thresholds. This complicates the detection of emerging faults and comprehensive condition monitoring [Nunes et al., 2023, Jiménez et al., 2020, Azar et al., 2022]. Additionally, the need to carefully tune event thresholds and diagnostic criteria introduces maintenance challenges, as improper settings can lead to missed events or excessive false positives, further complicating system upkeep and reliability [Nunes et al., 2023, Azar et al., 2022].

Downstream Machine Learning Tasks and Data Quantity: Two developments in recent years further underline the shortcomings of event-triggered logging in automotive systems: the massive increase in signal-based data in the in-vehicle network and the growing relevance of downstream ML tasks.

Recent industry and research reports indicate that the data quantity generated by ADAS (Advanced Driver Assistance Systems) sensors in vehicles is growing at an extremely rapid pace. According to a 2023 technical paper referencing McKinsey's 2021 automotive electronics report, by 2030, about 95% of new vehicles will be connected, up from around 50% today, and a single car can generate up to 1 terabyte (TB) of data per hour from its sensors [Bertoncello et al., 2021, Samantaray, 2023]. This explosive growth is driven by the increasing number and sophistication of sensors—such as cameras, radars, and lidars—required for advanced safety and autonomous driving features, with the complexity and volume of data presenting significant challenges for storage, processing, and transmission within embedded automotive systems [Samantaray, 2023].

Modern vehicles increasingly rely on data-driven intelligence to enhance safety, reliability and efficiency. Beyond perception and control, downstream ML tasks — those leveraging collected vehicle and sensor data for offline analysis, optimization and predictive functions — have become central to automotive-system design. These tasks include predictive maintenance [Theissler et al., 2021], anomaly and intrusion detection

[Övgü Özdemir et al., 2024], and fleet-level analytics like fuel consumption or maintenance scheduling [Chen et al., 2025].

Recent reviews highlight that while event-triggered and anomaly-based data collection can optimize resource use, they often result in fragmented or incomplete datasets, making it harder to implement robust predictive maintenance strategies and limiting the effectiveness of ML models that rely on continuous, high-resolution data streams [Nunes et al., 2023, Jiménez et al., 2020]. Multi-model and hybrid approaches are being explored to address these limitations, but the trade-off between data reduction and diagnostic completeness remains a significant challenge in both industrial and automotive contexts [Jiménez et al., 2020, Azar et al., 2022]. How modern research tries to approach this trade-off,is discussed in more detail in Section 3.

3   Background

In essence, event-triggered logging is a form of compression that reduces data volume by selectively recording only significant events. Given the limitations of this approach in automotive systems and the need for ML-ready data, one might look to traditional compression methods for an alternative.

Traditional Compression: Compression, as originated in information theory by Shannon [1948], is the process of encoding information using fewer bits than the original representation. Compression techniques can be broadly categorized into lossless and lossy methods. Lossless compression is based on two principles: distribution modelling, sometimes called entropy modeling, and entropy coding. Entropy modeling involves creating a probabilistic representation of the data, while entropy coding assigns shorter codes to more frequent symbols based on their probabilities, thereby minimizing the average code length. Lossy compression allows for some loss of information in exchange for higher compression ratios. This is typically achieved through techniques such as transform coding and quantization [Sayood, 2018]. For the purpose of this project, the focus will be on lossy compression as we focus on downstream ML tasks where some loss of fidelity is acceptable as long as the relevant information for the task is preserved.

Traditional compression methods, based on these information theory principles, often fall short in automotive applications, especially as a precursor for downstream ML tasks. For video/image compression traditional methods like JPEG or MP3 are optimized for human perception (e.g., visual quality) rather than ML tasks or efficient downstream data use [Ma et al., 2019]. For time series data, algorithmic approaches like CHIMP or Gorilla depend on manually chosen parameters like window size and are sensitive to data characteristics such as entropy and signal variability. This limits their effectiveness in capturing the nuances required for accurate ML model performance in automotive contexts [Johnsson, 2025]. These algorithmic approaches were investigated by Johnsson [2025] in a previous master thesis project. This work builds upon this thesis by exploring an alternative approach to compressing vehicle telemetry data.

Rate-Utility Trade-off and Related Research: As introduced in Section 2, constructing downstream ML models for automotive systems, or in fact Internet-of-Things (IoT) systems in general, is a constant trade-off between handling large quantities of data and maximizing model performance. Traditional compression techniques can reduce data volume, but often at the cost of losing critical information necessary for accurate ML tasks such as predictive maintenance, anomaly detection, and fleet analytics. The impact of this trade-off is well-documented in the literature. Muniz-Cuza et al. [2024] for example study the impact of lossy compression techniques on time series forecasting tasks and observe a constant trade-off between compression ratio and forecasting accuracy.

Existing research approaches these challenges from three different angles: utility-aware adaptive telemetry, neural compression, and task-aware compression.

- First, utility-aware adaptive telemetry methods aim to employ policy learning methods to dynamically adjust telemetry parameters to reduce maintenance costs while preserving data utility for downstream tasks. Although this approach is still emerging, recent research has demonstrated promising results [Zhang et al., 2023].

- Second, neural compression techniques learn data representations optimized for both compression efficiency and ML task performance. This research is heavily inspired by deep generative models like

GANs, VAEs, and autoregressive models, but focuses on compressing the data, instead of generating realistic data samples [Yang et al., 2022]. Neural compression techniques extend the introduced lossy compression principles in two key ways. First, they offer an alternative to traditional distribution modelling by leveraging deep neural networks to learn complex data distributions directly from the data, capturing intricate patterns and dependencies that traditional statistical models may miss. Second, they substitute traditional approaches to transform coding and quantization with learned representations [Yang et al., 2022]. Studies as early as 2019 have shown that neural compression methods can outperform traditional compression techniques for image and video data, especially at low bitrates [Löhdefink et al., 2019]. The same has been shown for time series data [Zheng and Zhang, 2023, Liu et al., 2024].

- Lastly, task-aware compression techniques focus on optimizing compression algorithms to retain information that is most relevant for specific tasks [Yang et al., 2022]. This idea has shown promise in handling time-series data more efficiently in IoT systems. Azar et al. [2020] and Sun et al. [2025] for example explore task-aware compression algorithms that adaptively prioritize data features based on their relevance to downstream tasks, demonstrating improved performance in resource-constrained environments.

When combining task-aware methods and neural compression methods, task-aware neural compression models have shown promise in reducing the rate-utility trade-off. These models are specifically designed to retain essential features for ML tasks while achieving high compression ratios [Yang et al., 2022]. Studies that empirically evaluate the performance of task-aware neural compression models are somewhat limited, but they do exist. In one study for example, Kawawa-Beaudan et al. [2022] use a hierarchical autoencoder-based compression network together with a recognition model and implement two hyperparameters to trade off between distortion, bitrate, and recognition performance.

An important note that needs to be made on neural compression methods is that they often rely on computationally heavy architectures like RNNs and transformers [Zheng and Zhang, 2023, Löhdefink et al., 2019, Kawawa-Beaudan et al., 2022, Liu et al., 2024], to solve the task of entropy modeling within the compression pipeline. The prevalence of these complex architectures is largely due to the complexity of the continuous latent space. Traditionally compression pipelines utilize continuous latent representations because these are much easier to optimize end-to-end with gradient descent. However, accurately modeling the entropy of continuous latents requires sophisticated context models capable of capturing long-range dependencies. Ballé et al. [2018] for example introduce a sophisticated but complex hyperprior to better capture these long-range dependencies and discuss the challenges.

4 Problems & Research Gaps

Analyzing relevant industry practices and literature on compression techniques reveals several significant unsolved challenges and research gaps, which will be addressed with this project:

- From the industry perspective, automotive systems need high-utility ML-ready data under severe bandwidth and computational limits. Existing event-triggered logging schemes introduce sampling bias and maintenance overhead.

- While there exists some exploration of task-aware neural compression techniques for image and video data, there is a notable lack of research focusing on time series data, which is the predominant data type in automotive and IoT applications. This gap is supported by a 2022 survey done on the topic of neural compression [Yang et al., 2022].

- Most of the discussed papers fail to address the computational constraints of in-vehicle embedded or IoT systems. The mentioned papers, if they use neural compression, primarily focus on achieving high compression rates while maintaining model performance. Because of this, computationally heavy neural network architectures like recurrent neural networks (RNNs) or transformers were chosen [Zheng and Zhang, 2023, Löhdefink et al., 2019, Kawawa-Beaudan et al., 2022, Liu et al., 2024].

So while task-aware approaches to modern compression techniques like neural compression have shown promising advancements in balancing the rate-utility trade-off, there remains a significant gap in analyzing their effects on time series data, specifically in vehicular contexts, where computational resources and bandwidth are often constrained. A lightweight, task-aware compression method for automotive time-series data is therefore needed.

5 Motivation

The basic motivation behind using tokenization as a neural compression strategy for automotive time series data is to produce discrete latent representations that simplify the entropy modeling task within the compression pipeline. By constraining the data to a finite set of tokens, the complexity of modeling the underlying data distribution is reduced, enabling the use of lightweight entropy models that are computationally efficient.

The idea to use discrete latent representations within the transform step of the compression pipeline to simplify the entropy modeling task and therefore enable more efficient compression has already been explored in recent years. It is the main inspiration behind the Vector Quantized-Variational AutoEncoder (VQ-VAE) architecture of van den Oord et al. [2018], one of the most prevalent methods to produce discrete latent representations. van den Oord et al. [2018] propose the use of vector quantization as a way to learn discrete latent spaces. The VQ-VAE architecture and its successors have been successfully applied to image and audio data, but their main focus remains reconstruction [van den Oord et al., 2018, Razavi et al., 2019]. This makes them suboptimal for task-aware compression tasks. Tokenization emerges as an alternative approach to produce discrete latent representations in audio and speech processing research [Schmidt et al., 2024].

Tokenization is traditionally understood as the mapping of high-dimensional, continuous inputs into a sequence of discrete symbols drawn from a finite vocabulary [Grefenstette, 1999]. Tokenization therefore can act as a form of transformation and quantization: it reduces dimensionality, decorrelates, and constrains representations to a compact code space. Additionally, tokenization can be made task-aware so that the retained tokens are maximally useful for prediction or classification. One example of this is the WavTokenizer by Ji et al. [2025], which efficiently tokenizes acoustic data for audio language modeling. We propose, that this idea can be translated to time series data to enable lightweight entropy modeling architectures. This would allow more computationally efficient compression pipelines, which, as shown, is especially relevant for in-vehicle embedded systems with limited computational resources.

6 Approach

Detailed Approach: The proposed approach is to develop a task-aware tokenization framework for automotive time series data compression that balances computational efficiency, compression rate, and ML utility.

- **Dataset:** The dataset will focus on available automotive sensor and telemetry test-fleet data supporting tasks such as predictive maintenance and anomaly detection. Alternatively, publicly available datasets such as the SCANIA Component X Dataset can be used [Kharazian et al., 2025].

- **Task 1:** Train downstream ML models on uncompressed data to quantify loss in predictive utility.

- **Task 2:** Implement baseline.

- **Task 3:** Develop a learnable tokenization module that discretizes data into semantically meaningful units optimized for downstream tasks.

- **Task 4:** Develop lightweight entropy modeling and coding schemes tailored to the tokenized representations.

- **Task 5:** Evaluate and compare the methods.

- Measure rate-utility curves across the methods.

- Evaluate trade-offs between computational efficiency.

- **Expected Outcome:** Demonstrate that task-aware tokenization achieves comparable rate-utility trade-off to established neural compression approaches, while increasing computational efficiency.

7 Goals and Challenges

To achieve this, we define the following goals and challenges:

- **Main goal:** Develop and evaluate a *task-aware tokenization framework* for automotive data that balances computational efficiency, compression rate, and ML utility.

- **Sub-goals:**

  - Quantify the loss in predictive utility when training ML models on uncompressed, tokenized and compressed data.

  - ...

- **Challenges:**

  - Agree on a downstream ML task or task type (e.g., predictive maintenance, anomaly detection).

  - Define how computational efficiency will be measured (e.g., inference time, model size).

  - Agree on a subset of the available automotive data.

References

Joseph Azar, Abdallah Makhoul, Raphaël Couturier, and Jacques Demerjian. Robust iot time series classification with data compression and deep learning. *Neurocomputing*, 398, 02 2020. doi: 10.1016/j.neucom.2020.02.097.

Kamyar Azar, Zohreh Hajiakhondi-Meybodi, and Farnoosh Naderkhani. Semi-supervised clustering-based method for fault diagnosis and prognosis: A case study. *Reliability Engineering & System Safety*, 222: 108405, 2022. doi: 10.1016/j.ress.2022.108405.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior, 2018. URL https://arxiv.org/abs/1802.01436.

L. L. Bello, R. Mariani, S. Mubeen, and S. Saponara. Recent advances and trends in on-board embedded and networked automotive systems. *IEEE Transactions on Industrial Informatics*, 15:1038–1051, 2019. doi: 10.1109/tii.2018.2879544.

Michele Bertoncello, Christopher Martens, Timo Möller, and Tobias Schneiderbauer. Unlocking the full life-cycle value from connected-car data. Technical report, McKinsey & Company, McKinsey Center for Future Mobility, Feb 2021. URL https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/unlocking-the-full-life-cycle-value-from-connected-car-data. White paper.

Fanghua Chen, Hong Jia, and Wei Zhou. Vehicle maintenance demand prediction: A survey. *Applied Sciences*, 15(20), 2025. ISSN 2076-3417. doi: 10.3390/app152011095. URL https://www.mdpi.com/2076-3417/15/20/11095.

Richard E. Fairley. *Automobile Embedded Real-Time Systems*, pages 377–389. Wiley-IEEE Press, 2019. doi: 10.1002/9781119535041.app2.

Gregory Grefenstette. *Tokenization*, pages 117–133. Springer Netherlands, Dordrecht, 1999. ISBN 978-94-015-9273-4. doi: 10.1007/978-94-015-9273-4_9. URL https://doi.org/10.1007/978-94-015-9273-4_9.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling, 2025. URL https://arxiv.org/abs/2408.16532.

Juan José Montero Jiménez, Sébastien Schwartz, R. Vingerhoeds, B. Grabot, and M. Salaün. Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics. *Journal of Manufacturing Systems*, 2020. doi: 10.1016/j.jmsy.2020.07.008.

Simon Johnsson. Large scale efficient data readout for vehicle fleets. Master's thesis, Chalmers University of Technology, 2025.

Maxime Kawawa-Beaudan, Ryan Roggenkemper, and Avideh Zakhor. Recognition-aware learned image compression. *Electronic Imaging*, 34(14):220–1–220–5, January 2022. ISSN 2470-1173. doi: 10.2352/ei.2022.34.14.coimg-220. URL http://dx.doi.org/10.2352/EI.2022.34.14.COIMG-220.

Zahra Kharazian, Tony Lindgren, Sindri Magnússon, Olof Steinert, and Oskar Andersson Reyna. Scania component x dataset: A real-world multivariate time series dataset for predictive maintenance, 2025. URL https://arxiv.org/abs/2401.15199.

Jinxin Liu, Petar Djukic, Michel Kulhandjian, and Burak Kantarci. Deep dict: Deep learning-based lossy time series compressor for iot data, 2024. URL https://arxiv.org/abs/2401.10396.

Jonas Löhdefink, Andreas Bär, Nico M. Schmidt, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. Gan-vs. jpeg2000 image compression for distributed automotive perception: Higher peak snr does not mean better semantic segmentation. *arXiv preprint arXiv:1902.04311*, 2019. doi: arXiv:1902.04311v1.

Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:1683–1698, 2019. doi: 10.1109/tcsvt.2019.2910119.

{Carlos Enrique} Muniz-Cuza, {Søren Kejser} Jensen, Jonas Brusokas, Nguyen Ho, and {Torben Bach} Pedersen. Evaluating the impact of error-bounded lossy compression on time series forecasting. In *Advances in Database Technology - EDBT*, number 3 in Advances in Database Technology - EDBT, pages 650–663. OpenProceedings, March 2024. doi: 10.48786/edbt.2024.56.

N. Navet and F. Simonot-Lion. *Automotive Embedded Systems Handbook*. CRC Press, 2017. doi: 10.1201/9780849380273.

P. Nunes, J. Santos, and E. Rocha. Challenges in predictive maintenance – a review. *CIRP Journal of Manufacturing Science and Technology*, 2023. doi: 10.1016/j.cirpj.2022.11.004.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019. URL https://arxiv.org/abs/1906.00446.

Rojalin Samantaray. Adas sensor data handling in the world of autonomous mobility. *SAE Technical Paper Series*, 2023. doi: 10.4271/2023-01-0993.

Khalid Sayood. *Introduction to Data Compression, Fifth Edition*. Morgan Kaufmann Publishers Inc., 5th edition, 2018. ISBN 978-0-12-809474-7.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. Tokenization is more than compression. *arXiv preprint*, pages 678–702, 2024. doi: 10.48550/arxiv.2402.18376.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Guoyou Sun, Panagiotis Karras, and Qi Zhang. Highly efficient direct analytics on semantic-aware time series data compression, 2025. URL https://arxiv.org/abs/2503.13246.

Andreas Theissler, Judith Pérez-Velázquez, Marcel Kettelgerdes, and Gordon Elger. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability Engineering & System Safety*, 215, 2021. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2021.107864. URL https://www.sciencedirect.com/science/article/pii/S0951832021003835.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Yibo Yang, S. Mandt, and Lucas Theis. An introduction to neural data compression. *Found. Trends Comput. Graph. Vis.*, 15:113–200, 2022. doi: 10.1561/0600000107.

Penghui Zhang, Hua Zhang, Yibo Pi, Zijian Cao, Jingyu Wang, and Jianxin Liao. Adapint: A flexible and adaptive in-band network telemetry system based on deep reinforcement learning. *IEEE Transactions on Network and Service Management*, 21:5505–5520, 2023. doi: 10.1109/tnsm.2024.3427403.

Zhong Zheng and Zijun Zhang. A temporal convolutional recurrent autoencoder based framework for compressing time series data. *Applied Soft Computing*, 147:110797, 2023. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2023.110797. URL https://www.sciencedirect.com/science/article/pii/S1568494623008153.

Övgü Özdemir, M. Tuğberk İşyapar, Pınar Karagöz, Klaus Werner Schmidt, Demet Demir, and N. Alpay Karagöz. A survey of anomaly detection in in-vehicle networks, 2024. URL https://arxiv.org/abs/2409.07505.