Designing and Training Energy Efficient Machine Learning Models: Spiking Neural Networks

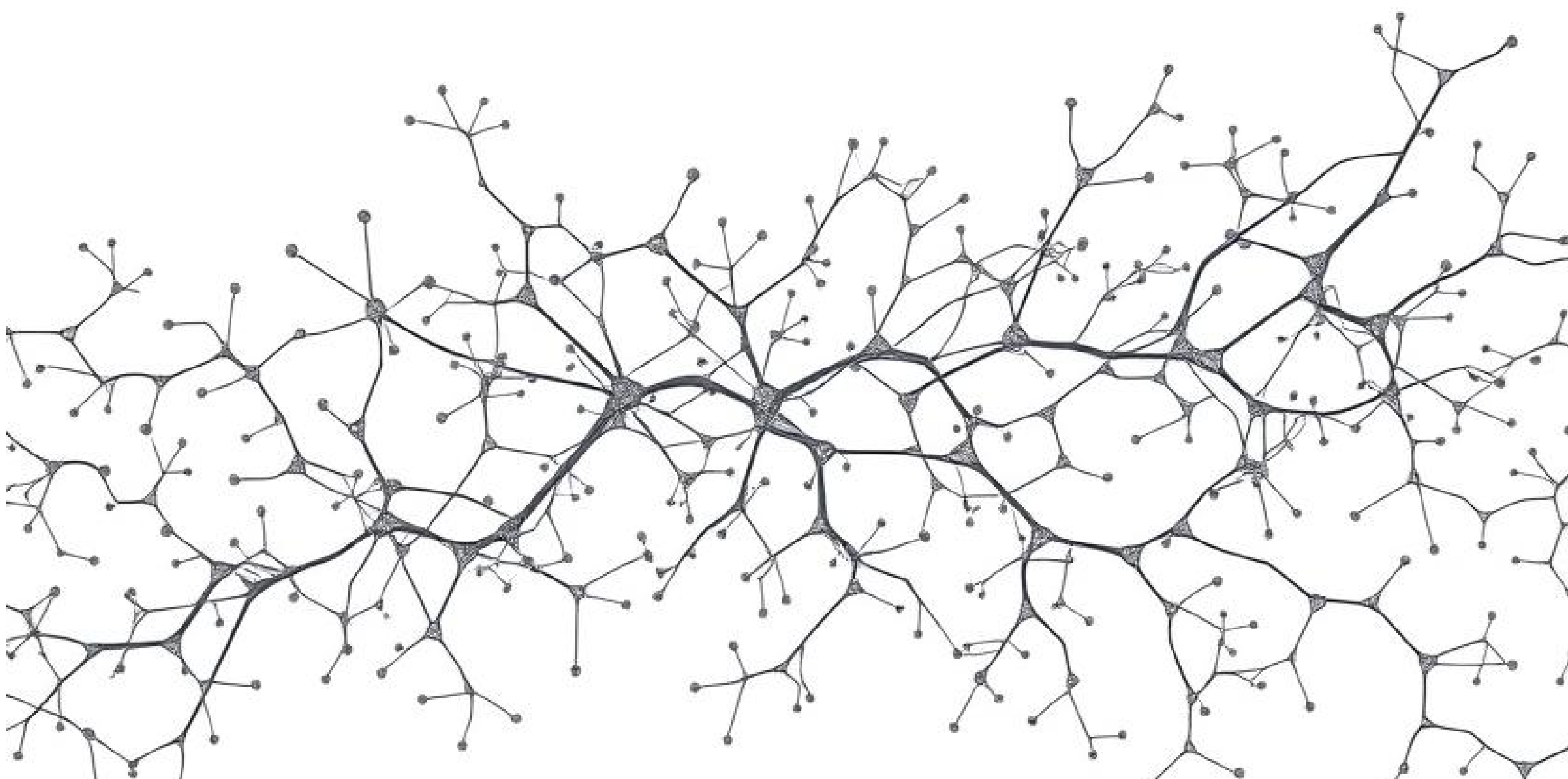
BEng Scoping & Planning Report

Rihards Klotiņš - 11083

Computer Systems Engineering

Supervisor: Dorian Florescu

2024/25 – words: 1,397



Introduction

Artificial Neural Networks are computational models designed to loosely mimic neurons in the brain. ANNs can be trained to model a wide range of phenomena, and once trained, they can perform inference — the process of predicting an output based on a given input. While ANNs have long been employed in fields such as computer vision and statistics, their adoption has surged in recent years, driven by the development of generative AI models like ChatGPT. Notably, the use of AI has increased fourfold since 2017 [1]. However, training and running these models is notoriously power intensive, their global power consumption is increasing at an exponential rate [2]. Not only does this make them expensive to operate it also raises sustainability concerns.

Artificial Neural Networks (ANNs), such as GPT-4, require substantial computational power, consuming far more energy than the human brain. Spiking Neural Networks (SNNs) offer a more biologically plausible alternative by using discrete "spikes" to transmit information only when a neuron's membrane potential reaches a threshold. This event-driven mechanism allows for sparse and asynchronous computation, significantly reducing power consumption. SNNs leverage temporal dynamics and precise spike timing to encode information, making them particularly promising for low-power applications and specialized neuromorphic hardware designed to exploit their efficiency [3].

Due to the popularity of ANNs, literature on training them is quite advanced, so a natural and popular choice for training SNNs has been by converting a trained ANN model into an SNN model. Such ways of training have had good results for some tests, however such a method incurs large computational costs during conversion and is limited by the architecture of ANNs which are less adaptable to dynamic data [4].

Thus, to fully harness the benefits of SNNs — from energy efficiency to novel architectures — effective *direct* training methods are essential. The most prominent approach is backpropagation-through-time (BPTT), often paired with the surrogate gradient (SG) method. Since spikes are inherently non-differentiable, SG provides an approximation of their gradients, enabling the use of BPTT to optimize the model's parameters. This combination has become a prominent approach for training SNNs, allowing models to learn from temporal data and allowing further exploration of novel SNN architectures. However, BPTT is not considered a biologically plausible learning method, as neurons in the brain are believed to learn using local information [5] — that is, signals received directly from neighbouring neurons. As a result, BPTT may not fully leverage the efficiencies observed in biological systems. For example, implementing local learning rules could reduce data movement across the chip, potentially enhancing energy efficiency.

My research examines advanced, biologically plausible methods for training spiking neural networks. I will analyse the mathematical and high-level strengths and weaknesses of each algorithm and present my findings in the final report. Additionally, I plan to implement these learning methods computationally—training and testing them to evaluate their performance relative to one another—and then analyse the data to draw meaningful comparisons. Focusing on auditory tasks is particularly compelling, as it leverages the inherent aptitude of SNNs for processing temporal, event-driven data while requiring considerably less computational power than applications like large language models or image pattern recognition.

Project Objectives

The objective of this project is to identify and assess the SNN training methods with the greatest long-term potential. By emphasizing biologically plausible approaches that exploit the inherent strengths of SNNs, I hypothesize that these methods will eventually prevail over conventional techniques. In this study, I will select and justify three of the most promising methods—such as those incorporating spike-timing-dependent plasticity (STDP) and liquid state machines (LSM) [6]—for detailed investigation. These models will be trained and evaluated on audio-based, temporally coded datasets, allowing for a comprehensive analysis of their accuracy and overall performance. Ultimately, the project will conclude with theoretical and evidence-based recommendations to inform researchers considering the adoption of these training methods.

Project Plan

This section presents the milestones for this research journey, starting with the literature review, moving on to the analysis of the selected methods, then the computational implementation, and finally the writing of the final report.

Literature Review

I will conduct a thorough literature review to identify the three most promising SNN training approaches, drawing from academic research and insights from current experts in the field. I will compile a list of potential methods and, with guidance from my supervisor and other specialists, narrow it down to the top three to ensure no significant techniques are overlooked. I will record my findings in an informal report, this will help me when writing the final report, and will allow my supervisor to check if I am on the right track.

Method Analysis

I will carefully study each method to understand its mathematical foundations and underlying guiding principles for why they have been developed. This knowledge will inform the computational implementation and performance analysis of these methods. As a milestone deliverable, I will produce an informal report summarizing my findings, which will also be incorporated into the final report to provide context for the computational results.

Computational Implementation

To maximize efficiency within the project's time constraints, I will leverage existing software whenever possible rather than implementing each method from scratch. Additionally, I will use established auditory datasets — such as N-TIDIGITS [7], SHD [8], and SSC [8] — specifically designed for SNNs. Care will be taken to ensure that all models are trained and tested fairly.

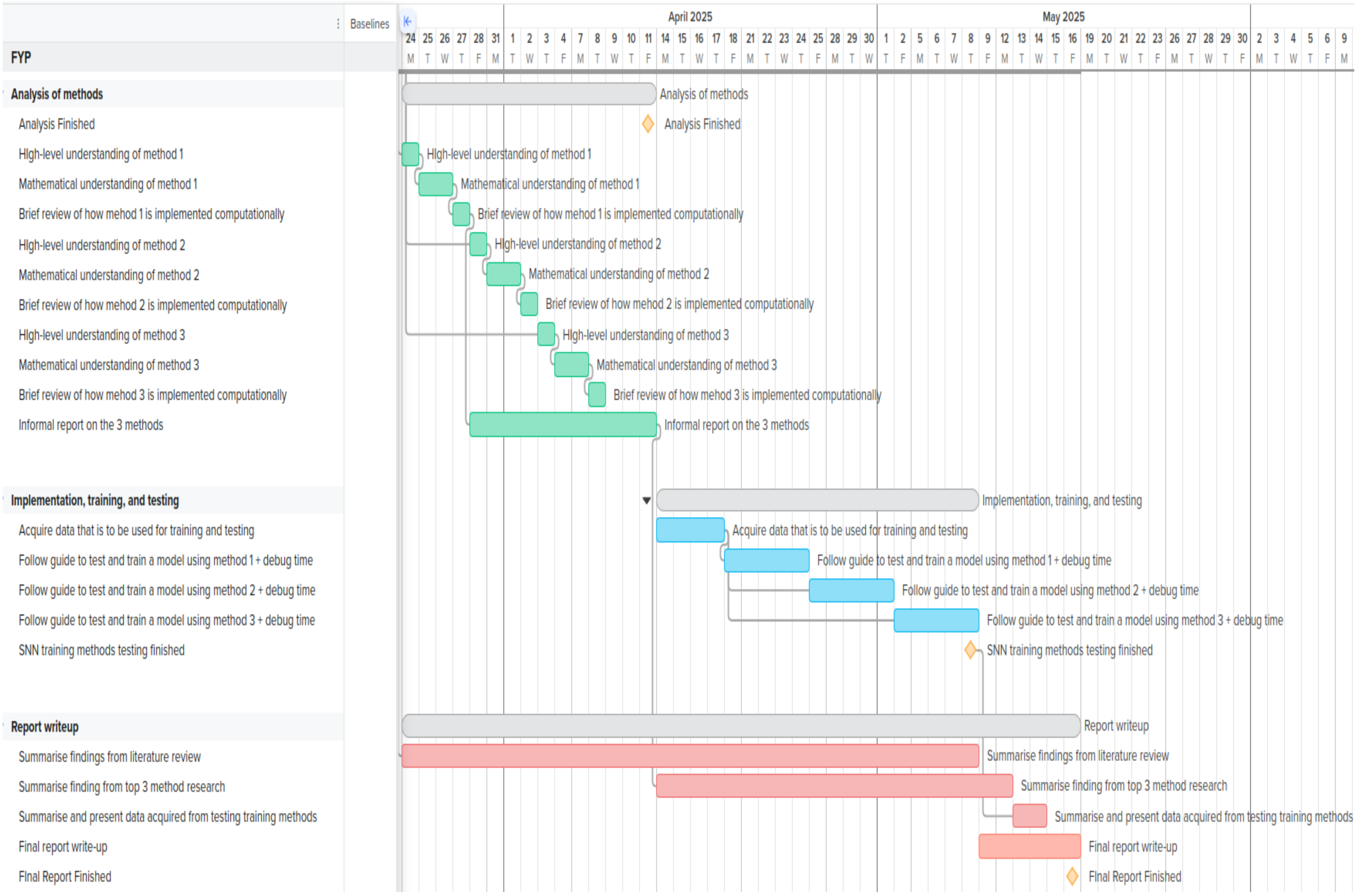
Final Report

Writing reports throughout the project will help simplify the creation of the final report. The final document will first introduce SNNs and their training methods, then clearly present the results of their training and testing.

Milestones and Deliverables

1. **Selection of Training Methods (Week 1)**
   - **Goal:** Identify the top 3 training methods of focus.
   - **Tasks:**
     - Survey various training methods.
     - Evaluate pros and cons of each approach.
     - Select the top 3 methods.
   - **Deliverable:** Short informal report summarizing the surveyed methods, their pros and cons, and the final selection of the top 3.
2. **Analysis of Selected Methods (Weeks 2–4)**
   - **Goal:** Understand the justification, inspiration, and mathematics of each selected method.
   - **Tasks:**
     - Compare and critique the methods.
   - **Deliverable:** Summarise, compare and critique the theoretical foundations of each method in an informal report.
3. **Implementation, Training, and Testing (Weeks 5–8)**
   - **Goal:** Implement the selected methods computationally and evaluate performance.
   - **Tasks:**
     - Code and implement each method.
     - Train and test using relevant datasets.
     - Collect performance data.
   - **Deliverable:** Short informal report detailing implementation, data of results from computational testing.
4. **Final Report (Week 9)**
   - **Goal:** Summarize findings and insights from the project.
   - **Tasks:**
     - Compile results from selection, analysis, and implementation.
     - Reflect on findings and provide conclusions.
   - **Deliverable:** Comprehensive Final Report.

Timeline - Gantt Chart

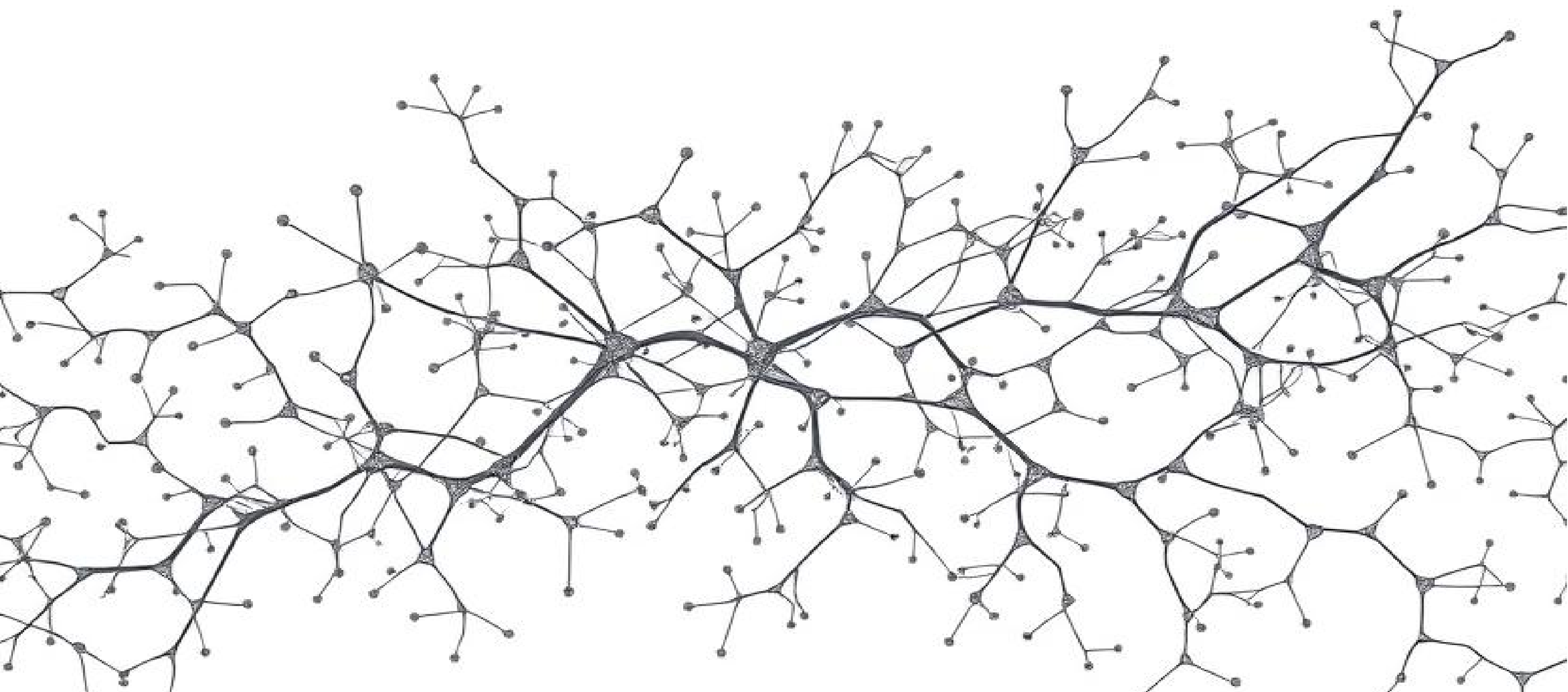


Resource planning and cost estimation

The resources required for this project include access to academic journals and online databases for method research, computational resources for implementing and testing the selected methods, and software tools such as Python and relevant machine learning libraries (e.g. PyTorch). Additionally, datasets for training and testing will be sourced from public repositories [7,8]. Costs for academic resources will be nil due to my attendance of university. Costs for computing will be negligible due to sufficient compute on my personal computer.

Risk management

Risk	Impact	Mitigation
Incomplete Literature review	May overlook promising methods or base the study on outdated approaches, reducing the project's validity.	Set clear search criteria, use multiple databases, and consult with supervisors or peers for additional insights.
Insufficient Understanding of Methods	Misinterpretation of theoretical concepts could lead to flawed analysis.	Dedicate time to discussing methods with experts and to check understanding with supervisors.
Implementation Challenges	Bugs in code taken from other people may prove difficult to debug, delaying project timelines.	Include debugging time in the planning.
Dataset Limitations	Audio dataset may be too small, or in the wrong format, affecting performance evaluation.	Set aside time to preprocess data if needed, and plan back-up sources of data.
Time Management Issues	Falling behind schedule could compromise the depth of analysis or delay the project.	Use project management tools to track progress, set interim deadlines, and conduct weekly reviews to stay on track.
Lack of Computational Resources	Training SNNs can be computationally expensive, leading to long runtimes or even failures due to insufficient resources.	Train models for tasks which don't need to be too computationally expensive and implement checkpoints to resume training after crashes.



References

[1] “The State of AI in 2025: Global survey | McKinsey.” Accessed: Mar. 12, 2025. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[2] B. Kindig, “AI Power Consumption: Rapidly Becoming Mission-Critical,” Forbes. Accessed: Mar. 12, 2025. [Online]. Available: https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

[3] Y. Cao, Y. Chen, and D. Khosla, “Spiking Deep Convolutional Neural Networks for Energy-Efficient Object Recognition,” *Int J Comput Vis*, vol. 113, no. 1, pp. 54–66, May 2015, doi: 10.1007/s11263-014-0788-3.

[4] G. Bellec, F. Scherr, E. Hajek, D. Salaj, R. Legenstein, and W. Maass, “Biologically inspired alternatives to backpropagation through time for learning in recurrent neural nets,” Feb. 21, 2019, *arXiv*: arXiv:1901.09049. doi: 10.48550/arXiv.1901.09049.

[5] C. Zhou *et al.*, “Direct training high-performance deep spiking neural networks: a review of theories and methods,” *Front. Neurosci.*, vol. 18, Jul. 2024, doi: 10.3389/fnins.2024.1383844

[6] D. Florescu and D. Coca, “Learning with Precise Spike Times: A New Decoding Algorithm for Liquid State Machines,” *Neural Computation*, vol. 31, no. 9, pp. 1825–1852, Sep. 2019, doi: 10.1162/neco_a_01218.

[7] J. Anumula, D. Neil, T. Delbruck, and S.-C. Liu, “Feature Representations for Neuromorphic Audio Spike Streams,” *Front. Neurosci.*, vol. 12, Feb. 2018, doi: 10.3389/fnins.2018.00023.

[8] B. Cramer, Y. Stradmann, J. Schemmel, and F. Zenke, “The Heidelberg spiking datasets for the systematic evaluation of spiking neural networks,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 33, no. 7, pp. 2744–2757, Jul. 2022, doi: 10.1109/TNNLS.2020.3044364.