algorithms

MDPI

Article

Impact of Sliding Window Variation and Neuronal Time Constants on Acoustic Anomaly Detection Using Recurrent Spiking Neural Networks in Automotive Environment †

Shreya Kshirasagar [1,*], Andre Guntoro [1] and Christian Mayr [2]

[1] Robert Bosch GmbH (Corporate Research), 71272 Renningen, Germany; andre.guntoro@de.bosch.com
[2] Highly-Parallel VLSI Systems and Neuro-Microelectronics, Technische Universität Dresden, 01062 Dresden, Germany; christian.mayr@tu-dresden.de
* Correspondence: shreya.kshirasagar@de.bosch.com
† This paper is an extended version of our paper published in 6th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2024), Funchal (Madeira Island), Portugal, 17–19 April 2024.

Abstract: Acoustic perception of the automotive environment has the potential to advance driving potentials with enhanced safety. The challenge arises when these acoustic perception systems need to perform under resource and power constraints on edge devices. Neuromorphic computing has introduced spiking neural networks in the context of ultra-low power sensory edge devices. Spiking architectures leverage biological plausibility to achieve computational capabilities, accurate performance, and great compatibility with neuromorphic hardware. In this work, we explore the depths of spiking neurons and feature components with the acoustic scene analysis task for siren sounds. This research work aims to address the qualitative analysis of sliding windows' variation on the feature extraction front of the preprocessing pipeline. Optimization of the parameters to exploit the feature extraction stage facilitates the advancement of the performance of the acoustics anomaly detection task. We exploit the parameters for mel spectrogram features and FFT calculations, prone to be suitable for computations in hardware. We conduct experiments with different window sizes and the overlapping ratio within the windows. We present our results for performance measures like accuracy and onset latency to provide an insight on the choice of optimal window. The non-trivial motivation of this research is to understand the effect of encoding behavior of spiking neurons with different windows. We further investigate the heterogeneous nature of membrane and synaptic time constants and their impact on the accuracy of anomaly detection. On a large scale audio dataset comprising of siren sounds and road traffic noises, we obtain accurate predictions of siren sounds using a recurrent spiking neural network. The baseline dataset comprising siren and noise sequences is enriched with a bird dataset to evaluate the model with unseen samples.

Keywords: sliding window; window sizes; spiking leakages; neuronal time scales; spiking neural networks; acoustic perception; anomaly detection; siren sounds; neuromorphic computing; time-series prediction

Citation: Kshirasagar, S.; Guntoro, A.; Mayr, C. Impact of Sliding Window Variation and Neuronal Time Constants on Acoustic Anomaly Detection Using Recurrent Spiking Neural Networks in Automotive Environment. *Algorithms* **2024**, *17*, 440. https://doi.org/10.3390/a17100440

Academic Editor: Sergey Y. Yurish

Received: 15 August 2024
Revised: 20 September 2024
Accepted: 20 September 2024
Published: 1 October 2024

1. Introduction

AI applications on edge devices need to perform efficiently with resource and power constraints. If applications need to be deployed at the sensory edge, these constrains could be hard to satisfy. Therefore, a trade-off between performance and resource demands needs to be made. Neuromorphic computing is a fairly new research field whose focus is to bring ultra low-power solutions without any compromise in efficiency to the future resource constrained edge devices. This field introduces biologically inspired neural networks, which have found recent applications in robotics [1], gesture recognition [2], constraint satisfaction problems [3], image classification through temporal coding [4], predictive medical systems [5], keyword spotting [6] and audio applications like scene

classification [7], speech recognition [8] etc. End-to-end neuromorphic keyword spotting has been demonstrated on neuromorphic hardware processors like Loihi [9] and large scale systems like SpiNNaker [10], showing benefits over conventional hardware accelerators. Spiking neural networks closely mimic sparse and asynchronous biological information processing. The models in SNNs are naturally operated in terms of time and are based on the principles of brain-inspired computing. Temporal perception of complex auditory scenes like speech signals is processed within the range of tens of hundreds of milliseconds (ms) [11], contributing to our investigation on temporal processing of siren audio sequences of approaching emergency vehicles with naturally adept, spiking neural networks.

Well-known approaches look at the problem of siren sound detection using deep learning methods [12–16]. The authors of [14], extensively explore 2D CNNs to detect siren signals based on the spectrum that is generated by combining multiple windowed FFTs to generate an image used as an input to the network, this subsequently leads to a higher computational effort. SNNs, on the other hand, exhibit temporal processing directly in their neurons, this motivates us to exploit time series tasks efficiently. Furthermore, with the advent of neuromorphic hardware [9,17], these small-scale networks with sparsity introduced through the optimal choice of time constants could potentially save orders of magnitude of energy as shown in [9]. This motivated us to take a closer look at the pre-processing stage and the way of optimizing anomaly detection task even further. The right level of biological abstraction of neuronal time constants could complement the performance of the audio scene classification tasks and give directions to build suitable neuromorphic hardware with informed parameterization.

This research work attempts to understand the intricacies of variation in sliding windows on the performance of the temporal detection of siren sounds in order to trade off the accuracy against the onset latency of the prediction. We model spiking neurons for real-time applications by inferring the impact of windowing in terms of encoding information. Through this empirical study, we aim to investigate the optimal window size with and without overlapping windows and encapsulate its impact on the task performance. A lack of understanding of the relation between neuronal decays and sliding windows in the literature for spiking architectures for the emergency vehicle detection task motivates us to answer a relevant research question that could help shape the hardware aspects of the processing pipeline for better performance. In an effort to validate the robustness of the proposed spiking architecture to solve this task, we augment a large scale audio dataset [18] with environmental actors that cover various noises, road backgrounds, human speech, bird songs, insect calls, rain, wind etc., from the bird [19] dataset. This helps us understand how the proposed model generalizes to new and unseen samples.

The remainder of this paper is organized as follows: Section 2 presents literature review on temporal detection of siren sounds and sliding windows employed in different tasks in detail. We present the method employed using sliding windows and SNN training approach in Section 3. We detail the experimental setup for an empirical study and subsequently present the results in Section 4. Finally, in Section 5 we present an outlook of the work, discussion and conclusion.

2. Related Work

Deep learning models namely—DNN (deep neural networks), CNN (convolutional neural networks), LSTM (long short-term memory) and hybrid CNN-LSTM are employed to solve human activity recognition (HAR) in [20]. The authors study the effect of sliding windows for preprocessing time-series data using four models and show improvement in accuracy, latency, and processing costs. Furthermore, the authors in [21] provide an extensive characterization of windowing technique. They show the impact of diverse window sizes for HAR task. Other interesting techniques like adaptive sliding windows are studied in [22] for assisted living application. The authors of [23], explore pose pattern recognition for sensors and extend the study to evaluate the impact on sliding windows. Their study is in alignment with the prior research that shows the introduction of overlapping windows increases the accuracy of pattern recognition.

Overall, through a literature study, we garner results on the significance of the choice of the optimal window size and sliding windows. To the best of authors' knowledge, there is scarce research on sliding window variation for anomaly detection task using SNNs. Through this empirical study, we attempt to understand the impact of sliding windows on the performance for anomaly detection task incorporated into bio-inspired networks like SNNs. We evaluate the relation between spiking neurons and sliding windows in terms of accuracy and onset latency. We further expand and elaborate the results on mel channel variation to investigate the intricate parameterization on the feature extraction front that impacts the performance of the siren sound detection task.

In parallel with preprocessing the data, SNNs inherit other properties which are engraved in neuroscientific studies. Prior research works have shown that leak channels exist in various synaptic transmissions in the visual cortex [24] and in sodium ion leak channels [25]. On neuromorphic datasets NMNIST and SHD, for different spiking neurons, leakages are studied for spatio-temporal pattern recognition in [26]. Authors in [26] explore the impact of synaptic and membrane time constants for three different spiking neuron models on pattern recognition and conclude the significance of neuronal leakage for both temporal features and the explicit presence of recurrent connections. Authors in [27] showed the importance of leak for LIF neurons in terms of robustness to noise by acting as high frequency filter. In parallel, authors also comment on the statistical relationship of sparsity introduced through leaky models and hardware efficiency through synaptic operations. In this paper, our aim is to understand if there is a relation between sliding windows and neuronal decays. This work attempts to partially answer this question by conducting an empirical evaluation of neuronal time constants in recurrent SNNs for acoustic anomaly detection.

3. Materials and Methods

We introduce the concept of overlaps in sliding windows for the acoustic anomaly detection task. The parameterization and tuning of the preprocessing stage leads to an impact on the performance of the acoustic event detection system. The goal of this research work was to improve the network parameters to further expand on the relation between the tweaking of the specific neuronal parameters and that of sliding windows.

We used artificial siren sequences generated from the publicly available siren dataset [18] to train our models. The artificial audio sequences were sampled at a sampling frequency of 48 kHz. We employed a small FFT window to minimize the hardware effort. Taking into account that our signal of interest, i.e., siren sounds, had a fundamental frequency between 400 Hz to 600 Hz, we started with a window size of 4096 which corresponds to 85.33 ms and we reduced the window size further further to 2048, . . ., 512 for this empirical evaluation. Windowing was applied on the audio sequences of 30s before FFT calculation as shown in Figure 1. Feature extraction was carried out using a mel spectrogram. The input to the hidden layer of the SNN was varied as 32, 64 and 128 mel channels. The SNN had a topology of 100 neurons with recurrent connections in the hidden layer. We kept constant parameters (structural/topological) throughout the experiments for homogeneity.

3.1. Dataset

thw dataset in [18] is comprised of siren sounds and road noises. The dataset consists of different types of siren sounds, namely, wail, yelp, hi-lo. We modified the publicly available dataset to perform temporal predictions using artificially generated audio sequences. More specifically, we utilized single channel siren and traffic noise recordings from the dataset presented in [18] and split the samples of each class (siren and noise) with a 80/20 ratio into train and test samples. All samples were resampled to a shared sample rate of 48 kHz. Based on the noise samples, a continuous sequence was generated. To each of these sequences of 30s duration, a random single siren sound of random length was added at a random time. To ensure accurate measurement of onset latency and network stability we

enforced the first 1s of the artificial sequences to exclude siren signals. This formed our baseline dataset with siren and noise sounds in the audio clips.

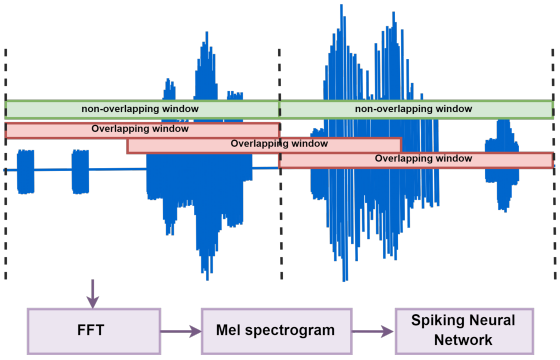

Figure 1. Sliding window method in the processing pipeline. Block diagram highlights the sliding windows on acoustic anomaly sequences used as input features in the spiking architecture. Different window sizes (with and without overlaps) are provided in the form of time slices as an input to FFT; mel features are extracted to inject as current input to the recurrent SNN.

Augmentation Dataset

Bird dataset [19]—We further included bird audio in combination with the baseline siren data [18]. The Bird dataset is considered as evaluation dataset that consists of three different sub-datasets. The Chernobyl dataset has 6620 audio clips collected from unattended remote monitoring equipment in the Chernobyl Exclusion Zone (CEZ). Poland NFC has 4000 recordings with different weather conditions and background noise comprising of wind, rain, sea noise, insect calls, human voice and deer calls. The Crowdsourced dataset is a held-out of 2000 recordings from the Warblr bird recognition app. The analysis of the proposed method on evaluation dataset aided in the further understanding of the robustness of the spiking architecture. Since the commonly known types of siren sounds were already included in the baseline dataset, our idea was to enrich the noise samples with other environmental actors to provide a qualitative outlook on the robustness of SNNs towards various actors.

3.2. Feature Extraction

We used windowing technique to deconstruct temporal features into spatial features to analyze different frequencies. A Hann window was used for smoothening of edges in FFT calculations. In this work, since our focus was on understanding the performance of windowing for the anomaly detection task, we used sliding windows with and without overlap. We used log-scaled mel spectrograms as input features to our SNN model (referred to as SpikeSireNet henceforth in this article). The window length and hop length were varied in the order of 2^x to obtain optimal design choices for better performance. For the mel transformation, we imposed a lower frequency limit of 50 Hz to cover the noise signals so that the network could easily differentiate between noise characteristics. The upper limit was set according to respective window size. For example, when w = 512 signals were extracted, the frequency was only within 0–513 Hz. We considered a total of 64 mel channels to constrain the feature range in most of our experiments, and we also varied the mel channels in a range between 32 to 128 channels. Furthermore, the features were converted to dB scale and a min-max normalization was applied to each time slice.

3.3. Network Architecture

We adopted the model architecture from our previous work in [7,28] to achieve predictions using a recurrent spiking neural network, as shown in Figure 2. Recurrent networks were chosen as they exhibit excellent ability to handle time-series data much better than feed-forward networks. Recurrences aid to excelling in tasks that need to uncover the realms of contextual understanding such as audio processing. The SpikeSireNet model

comprised a hidden layer with 100 spiking leaky-integrate-and-fire (LIF) neurons with recurrent connections. For our experiments with sliding-window variation and time constants, we adopted a homogeneous network structure with parameters specifically from our research work on designing a model to detect siren sounds in [7]. With the aim to understand the impact of dataset augmentation, we further slightly modified the design of the SNNs in terms of input features and neuronal time constants of the hidden layer. The network in Figure 2 comprises a hidden spiking layer and a single readout for predicting siren or not. The information processing in the hidden layer of SpikeSireNet is in terms of spikes. LIF neurons are analogous to biological neuronal processing. When the input stimulus crosses the threshold voltage, neuronal firing occurs. We aimed to understand the effect of sliding windows and neuronal processing on network predictions with a great focus on the spectral feature components. Therefore, we conducted experiments with various membrane (τ_{mem}) and synaptic (τ_{syn}) time constants of the neurons to evaluate the impact of different windows on the task performance.

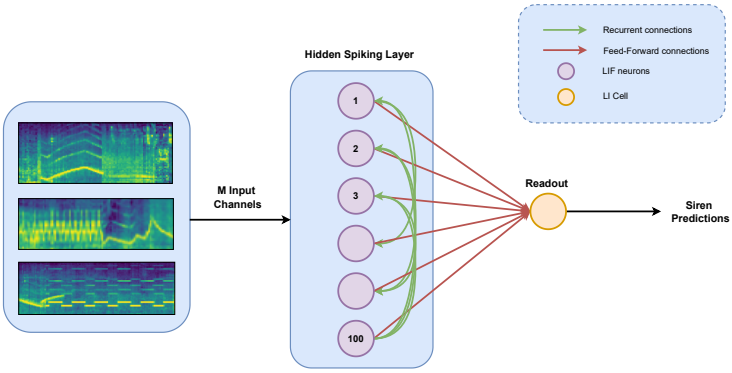

Figure 2. SpikeSireNet: Overview of the recurrent SNN used in this work is highlighted. The features extracted are given as M input channels (where M = number of mel bins) to the SNN. The spiking architecture comprises 100 hidden neurons with recurrent connections and a single readout neuron for siren predictions.

The design parameters and the hyper parameters set for the evaluation of the anomaly detection task are described in Table 1. A surrogate gradient based method was used to approximate the derivative of the LIF recurrent cell [29]. We employed a Leaky-integrator (LI) cell as a readout neuron, which had a continuous-valued output. The SNNs used in this work were built using the Norse [30] framework, an extension of PyTorch [31]. The differential equations and dynamics of the current-based LIF neuron are extensively discussed and presented in Equations (1)–(3) in [26]. Neurons have a membrane potential that decays with a membrane time constant τ_{mem}. Synaptic currents follow specific temporal dynamics. The exponentially decaying current triggered by the pre-synaptic input leads to the second dynamics of LIF neurons. This exponential decay of synapses is termed as synaptic time constant τ_{syn}. The specific dynamics of CUrrent BAsed (CUBA)- LIF neurons for the exponential decay of synaptic currents and membrane potential are presented in [26] in Equations (3) and (4).

Table 1. Design parameters of SNN architecture and hyperparameters.

Design Parameters	Values	Hyper Parameters	Values
Network structure	M [1]-100-1	Learning rate	1×10^{-3}
Threshold voltage	1 V	Batch size	16
Reset potential	0 mV	Optimizer	Adamax
Membrane time constant	2 s		
Synaptic time constant	2 s		

[1] M indicates mel channels.

4. Results

4.1. Experimental Setup

We employed the surrogate gradient method [29] to train our recurrent SNN model. In this work, we trained the models for 100 epochs, with a batch size of 16 on Nvidia V100. We performed the experiments on the modified audio dataset in [18]. Henceforth, we refer to the modified audio dataset as the Siren dataset. We obtained the results in terms of accuracy and onset latency. The LI cell outputted predictions based on the threshold value. Accuracy was based on total correct predictions vs total predictions at every time step. Onset latency was calculated as the time to obtain the detection of the siren onset event if a siren prediction was expected.

First experiment setup: In order to evaluate the sliding windows with different sample sizes, we set constant parameters for LIF neurons in the hidden layer of the SNN. The threshold voltage of neuron was set to 1 V, both the time constants (membrane and synaptic) were set to 2 ms. We designed experiments with variation in window length (w = 2^x, i.e., 4096, 2048, . . ., 512) and hop lengths (h). We chose three setups for our evaluation, h = w (no overlap), h = 0.5 w (50% overlap) and h = 0.25 w (75% overlap) with three different mel channels.

4.2. Feature Resolution for Sliding Windows

We demonstrated feature resolution using different sliding windows on the log-scaled mel spectrogram as depicted in Figure 3. From the experimental setup detailed in Section 4, the window and hop sizes were varied to understand the behavior of the network and obtain first impressions on spike activity in the hidden layer. LIF parameters and time constants were set according to Table 1.

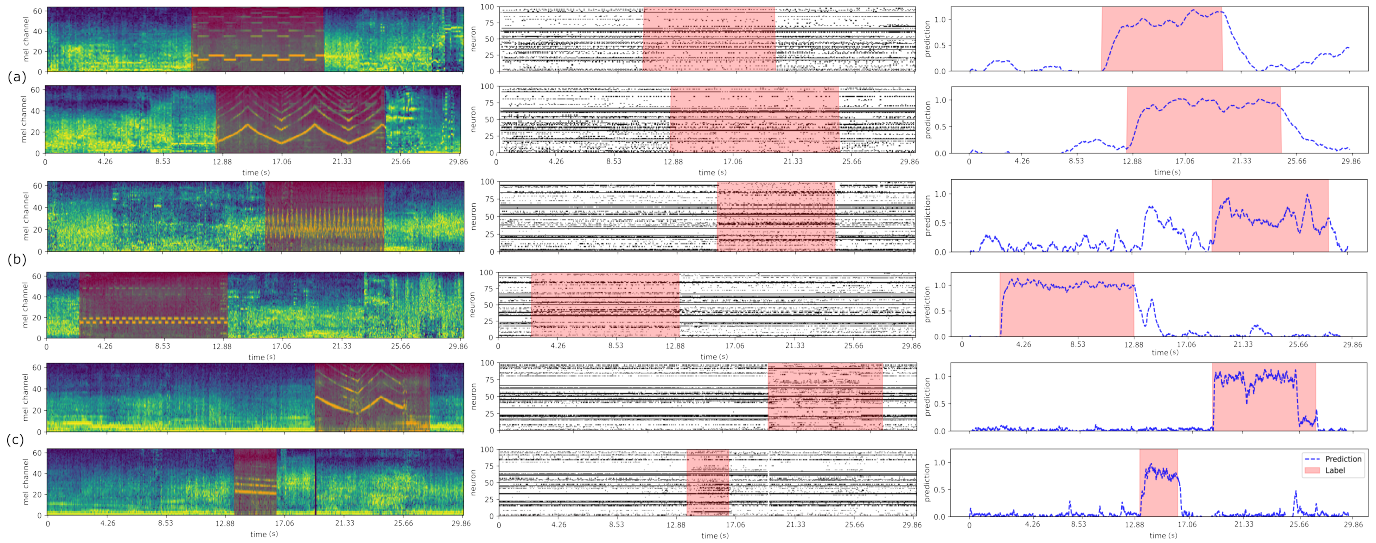

Figure 3. Feature resolution and their effect on performance. Demonstration of feature resolution on mel channels for sliding window variation on the anomaly detection task (presence of siren in each panel is highlighted in red). (**a**) First two rows: w = 4096 (h = w); (**b**) Next two rows: w = 2048 (h = 0.5 w); (**c**) Last two rows: w = 1024 (h = 0.25 w). The panels in each row showcase stages of information processing in the network pipeline of the proposed method. SpikeSireNet was trained on [18] for this set of experiments.

With w = 4096 and h = w, in the first row, the features of interest were in a lower frequency range due to higher energy concentration. With smaller window sizes and smaller overlaps (w = 2048, h = 0.5 w), the prediction strength became stronger due to the finer resolution in data points. It is interesting to note the increased spike activity in the hidden layer of SNN. For w = 1024 (h = 0.25 w), we saw a noticeable difference in feature resolution due to the increased granularity with distinguishable noise and siren sounds.

Overlap ratios lead to finer resolution and this was reflected in terms of spike activity which was observed to slightly increase. However, the focus of this work was not to examine the spike activity with different feature resolutions. Deduction of the spike activity for sliding windows in terms of spike sparsity could be part of the future work. This sort of analysis will be beneficial in terms of further compression of the network, without loss in performance quantifiers.

4.3. Effect of Overlapping Windows

The focus of this work was to analyze and demonstrate the effect of sliding windows with and without overlaps on accuracy and onset latency for siren predictions. We further elaborated the results in terms of mel channel variations as depicted in different sub-figures in Figure 4.

We expected that having more samples would increase the frequency density of the processed signal which allow us to have higher accuracy. Likewise, the introduction of overlaps should allow us to reduce the onset latency as the network would see new data more often. In alignment with our expectations and as evident from Figure 4, incorporating overlapping windows for the same window size helped to improve the training accuracy and our results showcased an increase in performance for overlaps within sliding windows. A slight drop in accuracy with smaller windows, and for their respective overlapping hop lengths was observed. This effect was observed in the window with sample size of 512 for hop length h = 0.5 w, which corresponded to a frequency range of 255 Hz. Siren sounds have a typical characteristic frequency of 400–600 Hz. The covered frequencies were below the signal of interest for window sizes below 256, thus explaining the slight drop. However, we need to keep in mind that we need to feed the input more frequently to the network, e.g., for $h = 0.5w$, the network needs to process the input twice as fast.

This paper strengthens the idea of diversifying input features to obtain a range of performance variations for an acoustic scene analysis. The feature component was restricted to different bin sizes i.e., 32, 64 and 128 mel channels. The subplots in Figure 4 have mel channel variation indicated as the title of the plot. Smaller bin sizes indicate higher accuracy values albeit with a slight increase in onset latency to detect the siren sounds within the sound sample. This is intuitive and due to the fact that the feature set provided to the input of the SpikeSireNet has improved feature granularity, thus making details of spectral components prominent. Maintaining overlaps further emphasizes that there is little compromise oin feature integrity; therefore, this explains the higher accuracy values for h = 0.5 w and h = 0.25 w than with no overlaps (h = w).

We investigated the time to detect siren sounds using the onset of events by predicting the neuronal state change. This gave us an intuition of how the fine temporal resolution of the spiking neurons influenced latency. Hence, we designed experiments to vary the sliding window with overlap and fixed windows to measure the time to first event, given an audio sequence being processed and knowing the ground truth (label). We observed the latency values of validation samples in the last epoch and averaged them over the batch size.

As observed in Figure 4b,d,e, the introduction of overlapping windows had a modest influence on latency within each window. It is our understanding that the hop length introduces faster processing, through a reduction in latency. Based on processing time alone, we expected 2× reduction in latency h = 0.5 w, 4× for h = 0.25 w. A latency improvement of 5× was achieved for window size of 4096, this is explainable from the neuronal sensitivity to detect siren sound events in windows with increased information granularity.

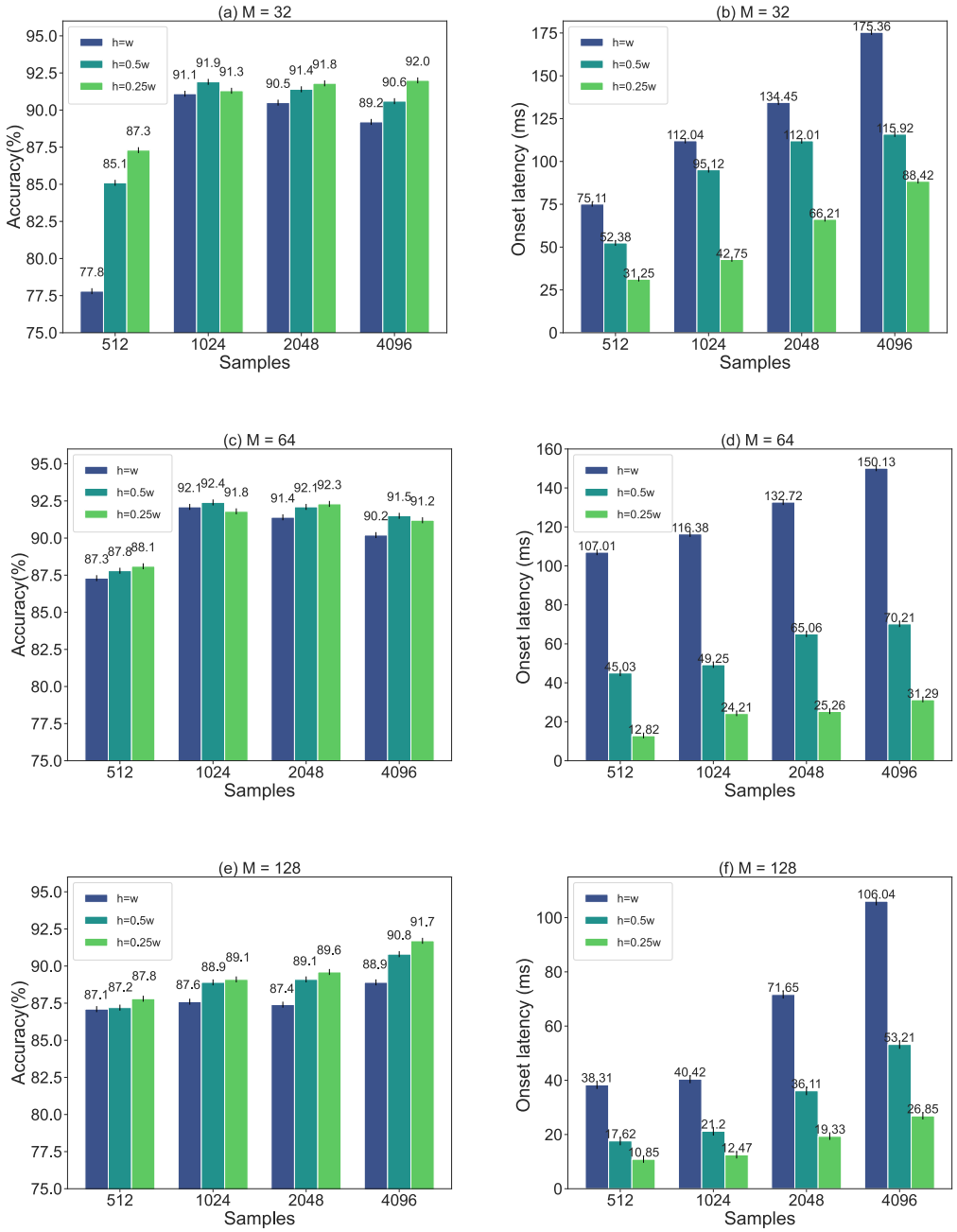

Figure 4. Sliding windows with different sample size impacts performance. Evaluation of SpikeSireNet for sliding window with mel channel variation on the anomaly detection task with membrane and synaptic time constants set to 2 s. (**a**) Introduction of hop length improves the prediction accuracy (**b**) Onset latency reduces for overlapping windows. From (**a**–**f**), the mel channel variation is seen from M = 32, 64 and 128 respectively. SpikeSireNet is trained on [18] for this set of experiments.

4.4. Relation between Sliding Windows and Time Constants

Second experiment setup: We performed experiments to understand the correlation between sliding windows and neuronal processing speed. The membrane time constant τ_{mem} and synaptic time constant τ_{syn} were varied with different window sizes and hop sizes to obtain results for the accuracy and onset latency for the siren prediction task. We further investigated the impact of windowing without any overlap for different time constants to obtain an optimal window size and to understand the impact of individual neuronal leakages.

We explored the relation between neuronal time constants of charge-based LIF and their impact on the performance with sliding windows. To garner results in this direction, we performed experiments with a variation in the overlap ratios for a window size of 4096 with different τ_{mem} and τ_{syn} ranging from 1 s to 100 s. From Figure 5, a clear trend indicates the accuracy was best for neuronal time constants ranging between 2–4 s. With larger time constants, the neurons decayed at a much slower rate, and this effect led to a slight degradation in accuracy for overlapping windows. An accuracy drop of nearly 5% occurred for higher time constants because fast responses or the high sensitivity of neurons with a slower membrane decay led to missing crucial information within overlapping windows.

However, for smaller time constants, this was reflected as a benefit in terms of modest deviation in accuracy. The results from the onset latency plots in Figure 5 suggest that there is a linear trend with time constants. Smaller time constants led to higher accuracy and a nearly 4× processing speed with the addition of overlapping windows for a fixed window size of 4096. In order to understand the impact of neuronal time constants on the spiking activity of the neurons and extrapolate these results for acoustic siren sounds, we performed a grid search method on Siren dataset [18] on SpikeSireNet for tuning of different time constant values and quantify in terms of performance metrics like accuracy. We expanded our experiments to accommodate individual variables for each set of membrane and synaptic time constants as shown in Table 2.

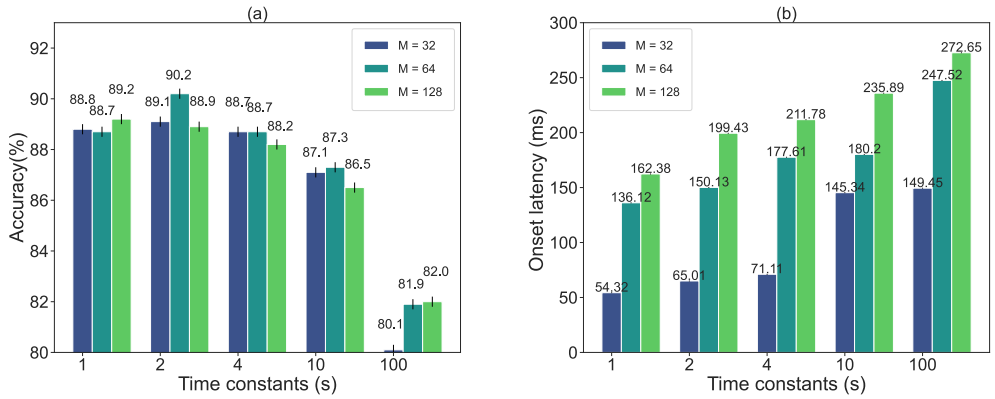

Figure 5. Smaller time constants result in higher performance. Neuronal time constants τ_{mem} and τ_{syn} are varied to evaluate anomaly detection task with w = 4096 and no overlaps (h = w). From (**a**), best accuracy values are obtained for time constants 2–4 s. From (**b**), $\tau = \tau_{mem} = \tau_{syn}$. Onset latency reduces for smaller time constants. Best performing mel channel variants are in the order as follows: 64, 32 and 128 for τ = 2 s, whereas mel channels reorder as 32, 64 and 128 for τ = 4 s. SpikeSireNet is trained on [18] for this set of experiments.

We varied time constants with a variable window size of 4096, . . ., 512 and without overlap to analyze the correlation between neuronal decays for windowing. The case of no overlaps was considered to constrain the experiment space and direct the focus towards understanding the statistical relation between the time constants for this particular task. The accuracy was higher for window sizes ranging between 85–50 ms with time constants in the range of 2–4 s. Smaller windows meant coarse features, and with no overlap, there was a chance of spectral leakage. In the window size range of 5–10 ms, having a time membrane and synaptic time constants in the range of 10–100 s, meant the decay was faster for both the time constants. This implied a higher firing rate and missing of information, leading to a slight degradation in performance. With both time constants in the higher bracket, the spikes retained were constrained to the specific window, which may have been the reason for the slight performance reduction. From Table 2, it is evident that the best performance was achieved for smaller neuronal time constants with an optimal window between 85.33–42.66 ms.

For higher time constants, the performance was slightly poor, emphasizing the importance of right parameterization of the spiking CuBA-LIF neurons in SpikeSireNet to achieve performance quantifiers. This provides further insights about the right biological modelling required to create efficient models on neuromorphic hardware. The neuronal leakages could impact data with a rich temporal structure; therefore, the results might completely vary with another dataset for SpikeSireNet. The best performing accuracy was obtained for τ_{mem} = 4 s and τ_{syn} = 2 s, i.e., 91.2% for a window size of 1024.

Table 2. Accuracy values for time constant heterogeneity. τ_{syn} is varied for different $\tau = \tau_{mem}$ values across variation in window sizes 4096, 2048, . . ., 512 with no overlap and with $n_{mel} = 64$. Smaller time constants and larger windows tend to achieve accurate siren predictions. All experiments performed in this table are trained and validated on Siren dataset [18].

τ_{syn}	h = w	τ = 1 s	τ = 2 s	τ = 4 s	τ = 10 s	τ = 100 s
1 s	4096	87.1%	87.8%	87.4%	85.2%	77.2%
	2048	89.2%	89.6%	89.3%	86.6%	77.5%
	1024	90.0%	**90.5%**	85.1%	87.6%	77.9%
	512	88.2%	85.4%	85.2%	82.9%	77.5%
2 s	4096	89.6%	**90.9%**	87.8%	87.0%	77.7%
	2048	89.6%	88.9%	90.3%	88.7%	79.5%
	1024	**90.6%**	88.7%	**91.2%**	90.0%	81.1%
	512	84.3%	87.3%	87.0%	85.2%	77.1%
4 s	4096	86.7%	88.0%	86.4%	87.5%	80.0%
	2048	83.2%	89.5%	88.1%	89.4%	82.0%
	1024	82.6%	**90.4%**	89.4%	**90.9%**	83.0%
	512	82.4%	85.8%	86.1%	86.3%	81.3%
10 s	4096	85.1%	85.8%	85.9%	85.6%	80.2%
	2048	87.5%	88.1%	88.2%	87.4%	83.3%
	1024	89.4%	89.6%	89.7%	86.0%	85.6%
	512	84.0%	82.9%	83.1%	86.0%	83.6%
100 s	4096	78.3%	77.7%	78.2%	78.8%	80.8%
	2048	80.3%	82.4%	82.5%	82.5%	83.2%
	1024	84.3%	84.5%	85.3%	86.3%	81.7%
	512	77.8%	77.3%	80.9%	80.3%	82.6%

Best accuracy is highlighted in bold.

4.5. Data Enrichment

In this set of experiments, we focued on understanding the robustness of SpikeSireNet and evaluating on cross-data. Data augmentation techniques were proven to be beneficial in terms of providing robustness and generalization to new data and larger data samples. We evaluated the performance of the SpikeSireNet architecture in a cross-augmented dataset setting. The baseline model was trained on Siren dataset as in [18]. We explored further data enhancement by introducing noise signals with other environmental actors. Furthermore, SpikeSireNet was pre-trained on Siren dataset and validated on the Bird dataset [19] to understand how the proposed model generalizes to new data. In this case the bird songs were quite similar to the siren sounds but had a low frequency range. We further evaluated this sort of data augmentation on SpikeSireNet using three settings by varying the mel channels of the input features; however, to understand the effect of the spectral components, we used constant network hyperparameters. The dataset attributes are mentioned in Table 3.

Figure 6 explains the accuracy obtained for each of the settings. Enriching data with various data augmentation techniques represents a true real-world scenario. This helps in real-time acoustic emergency detection task, and contributes towards model robustness to unknown environmental actors. The baseline dataset with siren + noise samples was trained, and the test accuracy was marked as purple circles (see Figure 6). The pre-trained

model was presented with the evaluation dataset consisting of samples from the Bird dataset introduced as a noise class (see Figure 6 marked by yellow triangles) alongside the siren dataset. With increasing mel bins, the performance improved. As the augmented dataset for evaluation was challenging because of the various noises and sounds like bird chirps, insect calls, rain, wind, etc., the performance increased only slightly. The distribution of noise and siren samples was uneven, and this might have impacted the performance. Given the challenges in the evaluation dataset and the fact that the model was not trained on the bird data, SpikeSireNet performed well towards the new unseen samples, therefore explaining the better accuracy values than for the bird augmented dataset. From the small variation between accuracy values, it was inferred that the model stabilized to siren and non-siren signals well and cross-data generalization was achieved.

Table 3. Dataset Attributes.

	Siren Dataset [18]	**Bird Dataset [19]**
Class	Siren sounds (wail, yelp, hi-lo), traffic noise	Bird audio, noise, insect calls, white noise, wind, human speech, rain
Average Clip Duration	≈3 to 15 s	3 to 30 s
Total Duration	≈7 h	50 h
Sampling Rate	48 kHz	44.1 kHz

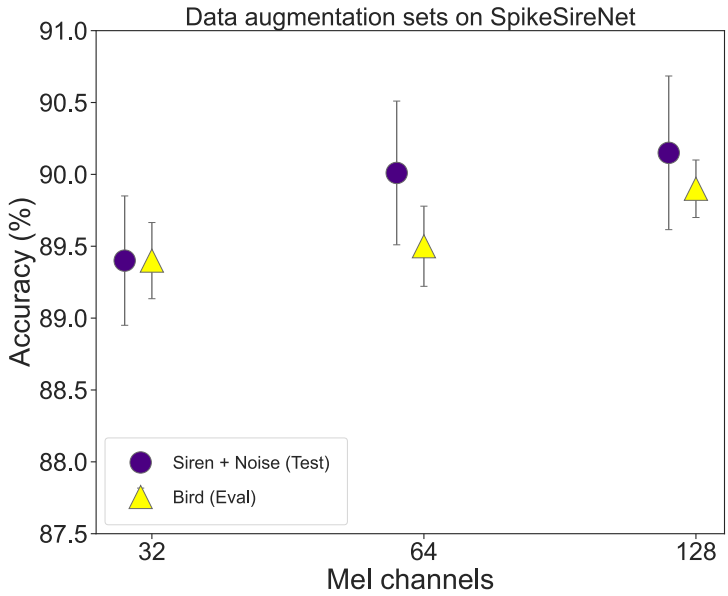

Figure 6. Validation on data induced with various environmental actors. Purple circles indicates accuracy obtained for testing on Siren dataset [18] and yellow triangles indicate evaluation accuracy on augmented siren data with bird dataset [19]. SpikeSireNet shows a similar performance with data augmentation and new unseen samples in the audio sequences. Mean accuracy and standard deviation over 10 random seeds is reported.

4.6. Discussion

We trained a recurrent SNN in different experimental setups to detect siren sounds on a modified public dataset from [18] to firstly understand the impact of sliding windows on the task performance and secondly, to provide a basis for the correlation between sliding windows and the leaky behavior of spiking neurons employed in SpikeSireNet. Our results indicated a performance boost in terms of accuracy using sliding windows with overlaps.

This was intuitive as with the enlarged features (high data points) that bacame available with the overlaps, the features were also retained, hence the accuracy was higher. With overlaps, the feature retention was increased and translated to less probability of missing information. Interestingly, a higher overlap ratio for hopping windows in smaller window regimes showed the best performance with $92.4 \pm 0.8\%$ accuracy for M = 32 and 64, closely matching the performance for larger window sizes. We particularly noted slight drop in accuracy with smaller windows below our signal of interest (siren sounds). We conducted experiments with different mel channels keeping the network parameters like hidden size and readout constant. The overall performance was better for 128 mel channels in terms of accuracy, also reaching the best onset latency. From our experimental results, a trade-off was achieved with 64 mel channels sufficing to predict accurate siren sounds for the proposed SpikeSireNet. Therefore, we conclude that sliding windows with an overlapping windows translate into improved task performance. Collectively, our results appear consistent with the body of literature that show the impact of sliding windows for various detection tasks like HAR [20], and pose pattern recognition task [23]. We further believe this is one of the first works to demonstrate a sliding window variation for an acoustic anomaly task, specifically using a spiking architecture.

Some authors have shown the significance of time constants of spiking neuron models to help improvise the performance of spatio-temporal pattern recognition task [26]. Neurons tend to show the behavior which confirms that leakages exist in neuronal models as underpinned by biology and by authors in [24,25]. The relation between the leaky behavior of spiking neurons in terms of exponential decay as time constants and that of the sliding windows was understood step-by-step through experiments in this work. At a higher level of abstraction we understood the variation of sliding windows and the impact on onset latency and accuracy for various time constants (in this case, keeping $\tau_{mem} = \tau_{syn}$). With smaller neuronal time constants, the response time of neurons was faster. The obtained prediction accuracy was highest for overlapping windows and thus led to faster processing speed. It is noteworthy, as the time constants fell in the range of audio sequences, this adversely impacted the accuracy by nearly 10%. We also noted that for smaller time constants and smaller windows, accuracy and onset latency were 5.5% higher and 2.4$\times$ lower, respectively, in contrast to higher time constants.

The evaluation dataset included the siren data [18] augmented with Bird dataset [19]. This cross-data validation gave us insights into the variation of performance; however, it is important to note the size of the data, split size and randomness can all contribute to masking the results. From Figure 6, we can confirm the variation within the performance values across datasets was heavily constrained in 2% range. This showcases the ability of SpikeSireNet to generalize to new unseen samples of environmental noise like the ones in Bird data and predict siren signals accurately.

5. Conclusions

In this work, we trained recurrent SNNs for an auditory anomaly detection task in an automotive environment. We conducted an empirical study on the impact of sliding windows and neuronal time constants on accuracy and onset latency. Intuitive selection of window length could lead to biased results. Through this study we showcased that right selection of sliding windows, and the expansion of window lengths significantly improved the accuracy and performance for acoustic anomaly detection task. Our study substantiated that the detection of acoustic cues, such as siren sounds, is influenced by window sizes. By fine-tuning the knob of overlap length, we achieved performance gains. We investigated sliding windows on different mel channels to provide an informative parameterization for designing architectures. This could lead to computational efficiency for pre-processing pipelines in the AI workflows. SpikeSireNet uses leaky neuronal models represented by membrane and synaptic time constants. We explored the landscape of neuronal time constant parameterization to reflect on the relation between spiking neurons and sliding windows. Our study showed that neuronal time constants had a great impact

on the acoustic anomaly detection task. For 4× processing speed and with smaller time constants, we observed gains in accuracy. The best performing accuracies were in the range of 1 s – 4 s for membrane and synaptic time constants. Prior studies have shown the effect of leakages on hardware compatible with the neuronal time constants. Our comprehensive results on neuronal leakages for an application like anomaly detection task could facilitate compute-efficient bio-inspired hardware with minimal effort and faster design time. It would be interesting to further evaluate our findings to understand the effect of time constants on spike sparsity to potentially build resource-constrained edge applications. We carried out a cross-data validation on the proposed SpikeSireNet for robustness against different real world environmental actors like human speech, bird chirp, rain, insect calls etc., by enriching the baseline data with different audio clips from the Bird dataset. The results obtained in this paper confirm that SpikeSireNet can generalize to new, unseen data samples, thus being robust to environmental noise and accurately predicting sirens.

Author Contributions: Conceptualization, S.K. and A.G.; methodology, S.K. and A.G.; software, S.K.; validation, S.K.; formal analysis, S.K.; investigation, S.K.; resources, A.G.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, A.G. and C.M.; visualization, S.K.; supervision, A.G. and C.M.; project administration, A.G. and C.M.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work received funding from European Union's Horizon Europe research and innovation programme under the grant agreement No. 101070374.

Data Availability Statement: Research conducted in this study is entirely based on publicly available datasets. The data can be found here: Siren dataset: https://figshare.com/articles/media/Large-Scale_Audio_Dataset_for_Emergency_Vehicle_Sirens_and_Road_Noises/19291472 (accessed on 10 April 2023). Bird dataset: https://zenodo.org/records/1298604 (accessed on 20 September 2024).

Conflicts of Interest: Authors Shreya Kshirasagar and Andre Guntoro were employed by the company Robert Bosch GmbH (Corporate Research), Renningen 71272, Germany. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Polykretis, I.; Tang, G.; Michmizos, K.P. An Astrocyte-Modulated Neuromorphic Central Pattern Generator for Hexapod Robot Locomotion on Intel's Loihi. In Proceedings of the International Conference on Neuromorphic Systems 2020 (ICONS 2020), Oak Ridge, TN, USA, 28–30 July 2020; Association for Computing Machinery: New York, NY, USA, 2020; Volume 23, pp. 1–9. [CrossRef]
2. Amir, A.; Taba, B.; Berg, D.; Melano, T.; McKinstry, J.; Di Nolfo, C.; Nayak, T.; Andreopoulos, A.; Garreau, G.; Mendoza, M.; et al. A low power, fully event-based gesture recognition system. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; Volume 1, pp. 7388–7397.
3. Mostafa, H.; Müller, L.K.; Indiveri, G. An event-based architecture for solving constraint satisfaction problems. *Nat. Commun.* **2015**, *6*, 8941. [CrossRef] [PubMed]
4. Stöckl, C.; Maass, W. Optimized spiking neurons can classify images with high accuracy through temporal coding with two spikes. *Nat. Mach. Intell.* **2021**, *3*, 230–238. [CrossRef]
5. Custode, L.L.; Mo, H.; Ferigo, A.; Iacca, G. Evolutionary Optimization of Spiking Neural P Systems for Remaining Useful Life Prediction. *Algorithms* **2022**, *15*, 98. [CrossRef]
6. Blouw, P.; Choo, X.; Hunsberger. E.; Eliasmith, C. Benchmarking keyword spotting efficiency on neuromorphic hardware. In Proceedings of the NICE '19: Proceedings 7th Annual Neuro-Inspired Computational Elements Workshop, New York, NY, USA, 26–28 March 2019; ACM: New York, NY, USA, 2019; pp. 1–8.
7. Kshirasagar, S.; Cramer, B.; Guntoro, A.; Mayr, C. Auditory Anomaly Detection using Recurrent Spiking Neural Networks. In Proceedings of the 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS), Abu Dhabi, United Arab Emirates, 22–25 April 2024; IEEE: New York, NY, USA, 2024; pp. 278–281.
8. Dominguez-Morales, J.P.; Liu, Q.; James, R.; Gutierrez-Galan, D.; Jimenez-Fernandez, A.; Davidson, S.; Furber, S. Deep spiking neural network model for time-variant signals classification: A real-time speech recognition approach. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE: New York, NY, USA, 2018; pp. 1–8.

9. Davies, M.; Srinivasa, N.; Lin, T.H.; Chinya, G.; Cao, Y.; Choday, S.H.; Dimou, G.; Joshi, P.; Imam, N.; Jain, S.; et al. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro* **2018**, *38*, 82–89. [CrossRef]
10. Furber, S.B.; Galluppi, F.; Temple, S.; Plana, L.A. The SpiNNaker project. *Proc. IEEE* **2014**, *102*, 652–665. [CrossRef]
11. Mauk, M.D.; Buonomano, D.V. The Neural Basis of Temporal Processing. *Annu. Rev. Neurosci.* **2004**, *27*, 307–340. [CrossRef]
12. Fazenda, B.; Atmoko, H.; Gu, F.; Guan, L.; Ball, A. Acoustic based safety emergency vehicle detection for intelligent transport systems. In Proceedings of the ICCAS-SICE 2009: ICROS-SICE International Joint Conference, Fukuoka, Japan, 18–21 August 2009; pp. 4250–4255.
13. Carmel, D.; Yeshurun, A.; Moshe, Y. Detection of alarm sounds in noisy environments. In Proceedings of the 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017; pp. 1839–1843.
14. Tran, V.-T.; Tsai, W.-H. Acoustic-Based Emergency Vehicle Detection Using Convolutional Neural Networks. *IEEE Access*. **2020**, *8*, 75702–75713. [CrossRef]
15. Cantarini, M.; Brocanelli, A.; Gabrielli, L.; Squartini, S. Acoustic Features for Deep Learning-Based Models for Emergency Siren Detection: An Evaluation Study. In Proceedings of the 12th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, Croatia, 13–15 September 2021; pp. 47–53. [CrossRef]
16. Marchegiani, L.; Newman, P. Listening for Sirens: Locating and Classifying Acoustic Alarms in City Scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 17087–17096. [CrossRef]
17. Merolla, P.A.; Arthur, J.V.; Alvarez-Icaza, R.; Cassidy, A.S.; Sawada, J.; Akopyan, F.; Jackson, B.L.; Imam, N.; Guo, C.; Nakamura, Y.; et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **2014**, *345*, 668–673. [CrossRef]
18. Asif, M.; Usaid, M.; Rashid, M.; Rajab, T.; Hussain, S.; Wasi, S. Large-scale audio dataset for emergency vehicle sirens and road noises. *Sci. Data* **2022**, *9*, 599. [CrossRef]
19. Stowell, D.; Pamuła, H.; Wood, M. Evaluation datasets for DCASE 2018 Bird Audio Detection (1.0) [Data set]. *Zenodo* **2018**. [CrossRef]
20. Jaén-Vargas, M.; Reyes Leiva, K.M.; Fernandes, F.; Barroso Gonçalves, S.; Tavares Silva, M.; Lopes, D.S.; Serrano Olmedo, J.J. Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models. *PeerJ Comput. Sci.* **2022**, *8*, 1052. [CrossRef] [PubMed]
21. Banos, O.; Galvez, J.M.; Damas, M.; Pomares, H.; Rojas, I. Window size impact in human activity recognition. *Sensors* **2014**, *14*, 6474–6499. [CrossRef] [PubMed]
22. Ma, C.; Li, W.; Cao, J.; Du, J.; Li, Q.; Gravina, R. Adaptive sliding window based activity recognition for assisted livings. *Inf. Fusion* **2020**, *53*, 55–65. [CrossRef]
23. Wang, G.; Li, Q.; Wang, L.; Wang, W.; Wu, M.; Liu, T. Impact of sliding window length in indoor human motion modes and pose pattern recognition based on smartphone sensors. *Sensors* **2018**, *18*, 1965. [CrossRef] [PubMed]
24. Snutch, T,P.; Monteil, A. The sodium "leak" has finally been plugged. *Neuron* **2007**, *54*, 505–507. [CrossRef]
25. Ren, D. Sodium leak channels in neuronal excitability and rhythmic behaviors. *Neuron* **2011**, *72*, 899–911. [CrossRef]
26. Bouanane, M.S.; Cherifi, D.; Chicca, E.; Khacef, L. Impact of spiking neurons leakages and network recurrences on event-based spatio-temporal pattern recognition. *Front. Neurosci.* **2023**, *17*, 1244675. [CrossRef]
27. Chowdhury, S.S.; Lee, C.; Roy, K. Towards understanding the effect of leak in spiking neural networks. *Neurocomputing* **2021**, *464*, 83–94. [CrossRef]
28. Kshirasagar, S.; Guntoro, A.; Mayr, C. An Empirical Evaluation of Sliding Windows on Siren Detection Task Using Spiking Neural Networks. In Proceedings of the 6th International Conference on Advances in Signal Processing and Artificial Intelligence, Madeira Island, Portugal, 17–19 April 2024; pp. 112–118.
29. Neftci, E.O.; Mostafa, H.; Zenke, F. Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks. *IEEE Signal Process. Mag.* **2019**, *36*, 51–63. [CrossRef]
30. Pehle, C.; Pedersen, J.E. Norse: A library to do deep learning with spiking neural networks. *GitHub* **2019**. [CrossRef]
31. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *GitHub* **2019**. [CrossRef]