3 Synthetic Dataset

In order to test and measure the different components of the in Section 1.2 introduced framework a synthetic dataset was generated. The methods used to generate the synthetic data as well as the generated dataset will be explained in the following.

3.1 Methods

To achieve images that are as close to reality as possible, the recorded data at INI was analyzed and modelled. Furthermore, different approaches for the generation of synthetic datasets were investigated. In general, the synthetic data is composed of firstly, background fluctuation and secondly, neuronal data.

3.1.1 Background

The background of the data is initialized with a mean pixel value which can be altered. In the pre-processing of the PCA-ICA pipeline the data is filtered by a low and high pass. The frequencies below and above the cut-off frequencies which are usually removed from the data were extracted and added to the background. The filtered data is shown in Figure 3.1. In addition, Gaussian noise is added.

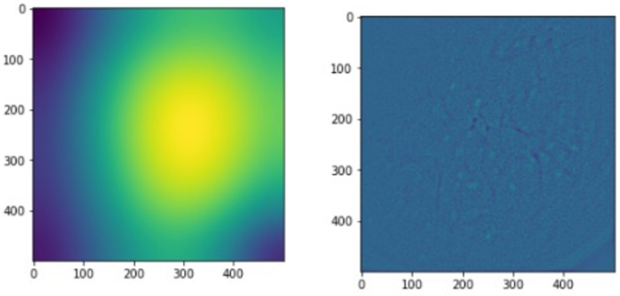

Figure 3.1: Filtered Background extracted from a real session (low pass left and high pass right).

3.1.2 Neuronal Signals

To generate the neuron ROI's and calcium traces centroids were created randomly in a first step. For the concrete shape of each neuron a 360 degrees radius with 100 values is Poisson distributed initialized and afterwards Gaussian smoothed. The resulting functions of r can be seen in Figure 3.2. Some example neuron shapes are shown in Figure 3.3.

As a subsequent step, for each neuron spiking events are randomly generated. The intensity of each activation is drawn from a skew normal distribution which

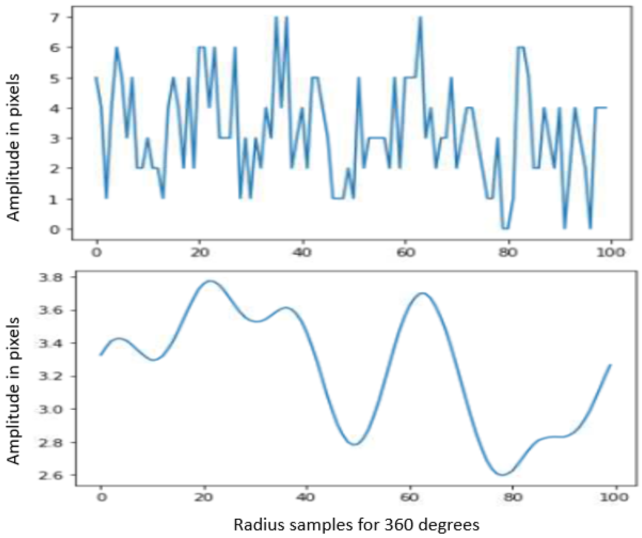

Figure 3.2: Top: Example function of neuron shape radius after being randomly initialized following a Poisson distribution. Bottom: Example function of neuron shape radius after applied Gaussian smoothing. X axis shows radius samples (100 samples for 360 degrees), Y axis shows the amplitude in pixels.

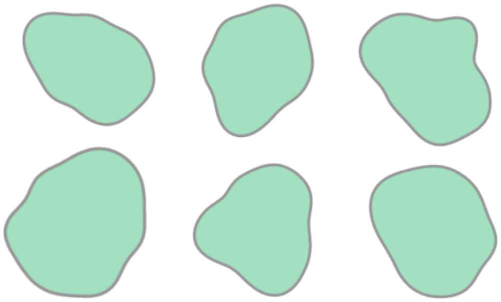

Figure 3.3: Examples of generated neuron shapes.

is modeled after the histogram of activation intensities detected by the PCA-ICA method when applied to a real session. Rise and decay of each activation vary between values that model the extracted traces from the CNMF-E and PCA-ICA method on real data. After generating the trace, noise is added and the trace is assigned to the region of the corresponding neuron shape. An example trace is shown in Figure 3.4

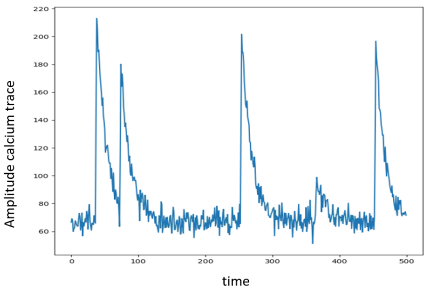

Figure 3.4: Example of a synthetic calcium trace.

3.1.3 Merging Signals and Background

When adding both background and neuronal signals a signal to noise ratio is used. Moreover, a Gaussian sliding filter is applied such that the shining neuron shapes blend n the background. The added filtered noise from a real session is randomly rotated.

3.1.4 Randomizing Parameters

Many parameters are set as ranges and then randomly selected from uniform or other distributions. Firstly, the number of datasets to be generated, sequence length and image size can be determined. For the background signal to noise ratio and mean background value can be set. With respect to the neuron data, the number of neurons, their average size, the variation of the neuron shape, the rise and decay length of their activations and the maximal reachable intensity can be chosen. Lastly, a parameter that determines how active the neurons are on average, i.e. in how many percentage of frames of the sequence the neurons are active , can be set.

3.2 Comparison of Synthetic to Real Data

In the following the generated synthetic data will be compared to either extracted data obtained by an unsupervised method or to real data directly.

3.2.1 Neuron's Region of Interest

Figure 3.5 shows 3 different coordinate maps: firstly, the generated synthetic ground truth of the neuron's ROI's, secondly, the extracted cells of the PCA-ICA method for a real session and thirdly, the extracted cells of the CNMF-E method for the same session. The main differences between the maps are that neurons are elongated on the outer circle of the microscope's field of view (FOV) in the PCA-ICA extraction map. This was not considered for the synthetic data. Note, that the CNMF-E method handles this microscope effect differently as it either fails to detect these neurons (as they are also harder to detect since the closer to the center of FOV, the brighter the ROI is) or predicts an adjusted shape. Furthermore, the neuron ROI's of the synthetic data overlap more often and additionally have a greater intersection area. This is due to the fact that the centroids are randomly sampled and since only 600 neurons are generated the data is not as evenly distributed as in the real data. Note, that the CNMF-E ROI map also shows many overlapping neurons but this is because of the many unmerged predictions for one single neuron.

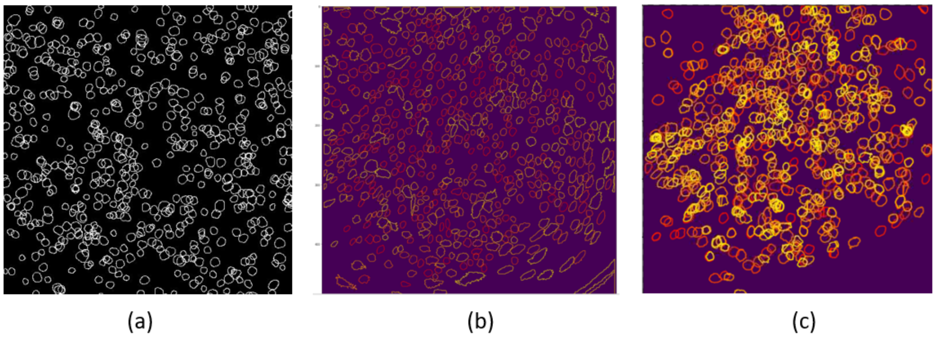

Figure 3.5: Comparison of neuron's ROI's maps. (a) ROI ground truth of synthetic data, (b) extracted ROI map of a real session by the PCA-ICA method, (c) extracted ROI map of a real session by the CNMF-E method.

3.2.2 Calcium Trace

As can be seen in Figure 3.6, the extracted calcium traces from the PCA-ICA (b) and CNMF-E (c) method differ vastly. While the PCA-ICA method averages cell activity of a neuron's ROI over time and includes basic noise removal, the CNMF-E method models the cells activity and obtains a noise free signal trace. Noise can still occur in the form of wrongly detected spiking events. Therefore, the CNMF-E output was used for rise and decay modelling and noise was modeled after the PCA-ICA trace. This type of modelling is also evident in the synthetic trace (a), as we can clearly see the decay similar to the CNMF-E method under the noise. Note, that the traces are arbitrarily picked, for all extraction methods and synthetic data there exist more or less active traces with higher and lower intensities.

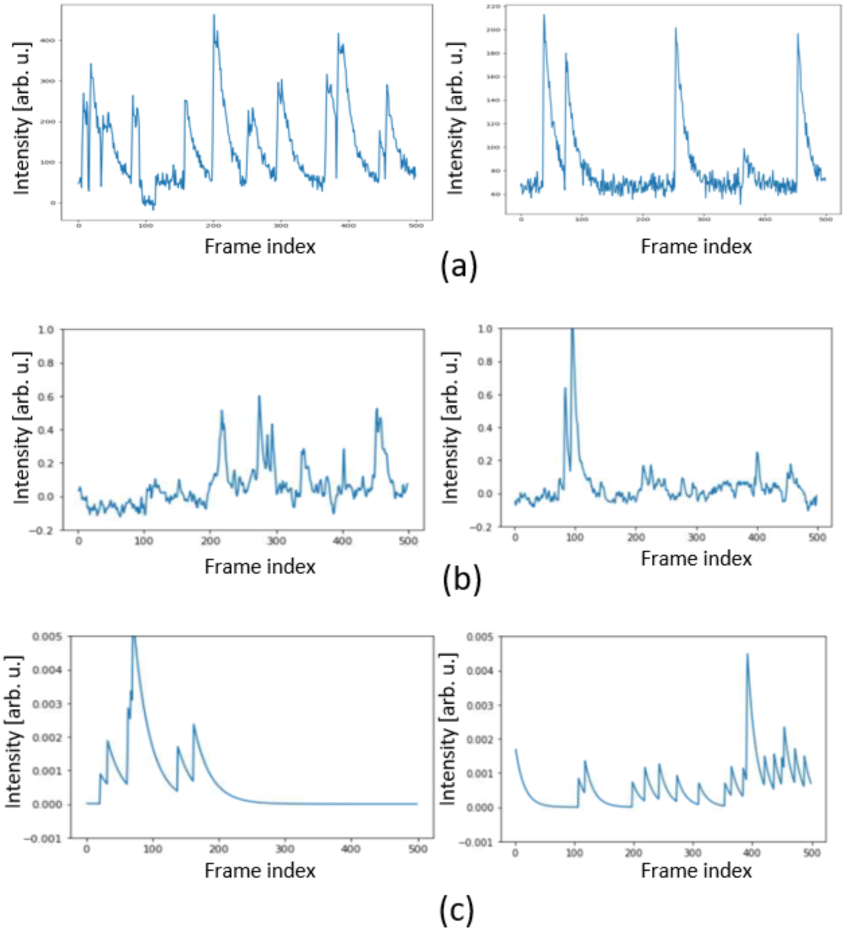

Figure 3.6: Comparison of 6 example calcium traces. (a) example calcium traces of synthetic data, (b) example calcium traces extracted of a real session by the PCA-ICA method, (c) example calcium traces extracted of a real session by the CNMF-E method.

3.2.3 Movie Frames

In Figure 3.7 three example frames are shown each for synthetic and real pre-processed data. The frames look qualitatively similar with respect to the mean average background value, number of neurons and their spiking events. To create more realistic frames one would further have to tackle the dissimilarities between synthetic and non-synthetic frames such as adding more background fluctuation and creating an even brighter spiking event. As the pre-processed movie underwent a $\frac{\Delta f}{f}$ transformation the neurons turn dark if they are below their average intensity. This was modeled for the synthetic data as well as can be seen in Figure 3.6 but not as heavily. Due to the missing and hard to reproduce background fluctuations dark spots otherwise would not have been partially covered by the background noise and would have been to prominent in the single frames. Another difference is the lack of radial blur which is complicated to implement as one would not only have to apply it to the single frames but also change the ground truth coordinates of the outer neurons. Due to time constraints this effect was not implemented. For the interpretation of the extraction methods explained in Chapter **??** and **??** the simplicity of the synthetic data is considered.

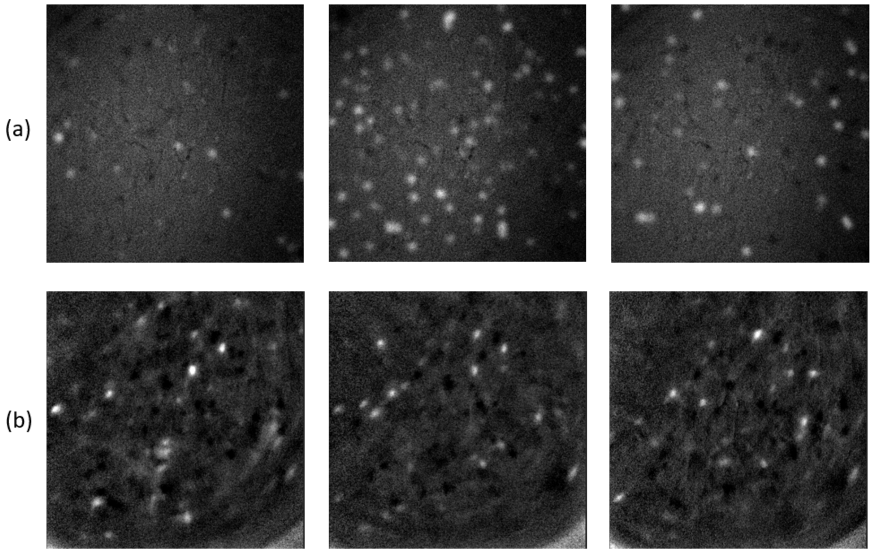

Figure 3.7: Comparison of 6 example calcium imaging frames. (a) example calcium imaging frames of synthetic data, (b) example calcium imaging frames of a real session (pre-processed movie)

3.3 Final Dataset Parameters

For benchmarking purposes two small datasets of each 800 frames sequence length were created since a much larger datasets would have slowed down the pipeline of extracting signals with two unsupervised methods and the evaluation process. If needed, i.e. for pretraining a CNN, of course a much larger dataset can be created.

The concrete parameters set for generating the data are given in Table 1. The datasets vary in the mean active value of the neurons which leads to one movie with slightly more activity than the other. This was done to evaluate the effect in performance of the different methods on the easier to detect dataset with more activity or the harder to detect dataset with less activity. As most of these parameters are ranges for random choices of distributions other properties of the datasets vary as well. Note, that a seed can be set to recreate the same dataset.

Table 1: Overview of Parameters Chosen for Synthetic Datasets

Parameters	**Dataset 01**	**Dataset 02**
Sequence Length	800	800
Image Size	(500,500)	(500,500)
Background SNR	[0.3, 0.25]	[0.3, 0.25]
Mean Background Value	50	50
Number of Neurons	600	600
Neuron Size	[1.5, 2.5]	[1.5, 2.5]
Smoothing Factor	[16, 18, 20]	[16, 18, 20]
Rise Length	[2, 4]	[2, 4]
Decay Parameter	[0.05, 0.7]	[0.05, 0.7]
Mean Intensity	50	50
Active Value	0.05	0.2