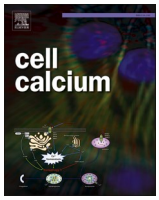
Cell Calcium 121 (2024) 102893

Contents lists available at ScienceDirect

Cell Calcium

journal homepage: www.elsevier.com/locate/ceca



A deep learning-based approach for efficient detection and classification of local Ca^{2+} release events in Full-Frame confocal imaging

Prisca Dotti [a,b], Miguel Fernandez-Tenorio [a], Radoslav Janicek [a], Pablo Márquez-Neila [b], Marcel Wullschleger [a], Raphael Sznitman [b], Marcel Egger [a,*]

[a] Department of Physiology, Universität Bern, Bern, Switzerland
[b] ARTORG Center, Universität Bern, Bern, Switzerland

ARTICLE INFO

Keywords:
Deep learning
AI
Full-Frame confocal imaging
Ca^{2+} events
Ca^{2+} sparks
Ca^{2+} puffs
Ventricular myocytes

ABSTRACT

The release of Ca^{2+} ions from intracellular stores plays a crucial role in many cellular processes, acting as a secondary messenger in various cell types, including cardiomyocytes, smooth muscle cells, hepatocytes, and many others. Detecting and classifying associated local Ca^{2+} release events is particularly important, as these events provide insight into the mechanisms, interplay, and interdependencies of local Ca^{2+}release events underlying global intracellular Ca^{2+}signaling. However, time-consuming and labor-intensive procedures often complicate analysis, especially with low signal-to-noise ratio imaging data.

Here, we present an innovative deep learning-based approach for automatically detecting and classifying local Ca^{2+} release events. This approach is exemplified with rapid full-frame confocal imaging data recorded in isolated cardiomyocytes.

To demonstrate the robustness and accuracy of our method, we first use conventional evaluation methods by comparing the intersection between manual annotations and the segmentation of Ca^{2+} release events provided by the deep learning method, as well as the annotated and recognized instances of individual events. In addition to these methods, we compare the performance of the proposed model with the annotation of six experts in the field. Our model can recognize more than 75 % of the annotated Ca^{2+} release events and correctly classify more than 75 %. A key result was that there were no significant differences between the annotations produced by human experts and the result of the proposed deep learning model.

We conclude that the proposed approach is a robust and time-saving alternative to conventional full-frame confocal imaging analysis of local intracellular Ca^{2+} events.

1. Introduction

Ca^{2+} signaling pathways are crucial in various physiological processes in almost all cell types [1]. In this context, local Ca^{2+} release events (e.g., Ca^{2+} sparks, Ca^{2+} puffs, and Ca^{2+} blips) are central in muscle contractility and excitation-contraction coupling (ECC) regulatory function. Ca^{2+} release events result from the opening of Ca^{2+} release channels (ryanodine receptors, $InsP_3$ receptors) localized in the intracellular Ca^{2+} stores (sarcoplasmic reticulum) membrane. The coordinated openings of these Ca^{2+} release channels form the building blocks for global Ca^{2+} transients, thereby opening up the possibility of fine-tuned regulation (local control theory of excitation-contraction coupling [2]) of contraction or other cellular functions. Different mechanisms are responsible for triggering local Ca^{2+} events, which strongly depend on the cell type.

In cardiomyocytes, *Ca^{2+} sparks*, local releases of Ca^{2+} from clustered Ryanodine receptors (RyRs) [4] and *Ca^{2+} puffs* [5] based on intracellularly synthesized inositol 1,4,5-trisphosphate ($InsP_3$) which activates the coordinated opening of clustered $InsP_3$ receptors ($InsP_3Rs$), are present. Functional crosstalk between RyRs and $InsP_3Rs$ has been observed in cardiomyocytes [6], which may have a significant regulatory function in cellular remodeling that accompanies various cardiac pathologies.

A detailed investigation of the underlying mechanisms of subcellular functional crosstalk requires the accurate detection, classification, and separation of Ca^{2+}sparks and Ca^{2+}puffs. While these events exhibit distinct spatiotemporal properties [6], the differences can be very subtle, leading to overlapping features [7], which poses challenges for their

* Corresponding author at: Department of Physiology, University of Bern, Buehlplatz 5, CH-3012 Bern, Switzerland.
E-mail address: marcel.egger@unibe.ch (M. Egger).

https://doi.org/10.1016/j.ceca.2024.102893
Received 22 December 2023; Received in revised form 24 March 2024; Accepted 23 April 2024
Available online 24 April 2024

classification.

Currently, pharmacological tools in combination with Ca^{2+}-sensitive dyes offer limited precision in distinguishing between Ca^{2+}sparks and Ca^{2+}puffs by using confocal full-frame imaging as the method of choice.

Here, we propose a deep learning model (DLM) that can efficiently locate and classify local Ca^{2+} release events in confocal full-frame image series. The approach is here exemplified in intracellular Ca^{2+} events in cardiomyocytes [3]. However, the results and conclusions presented below could benefit other research fields since the DLM can be trained and used on other cell types and experimental situations. For instance, local Ca^{2+} release mediated by RyRs and $InsP_3Rs$ occurs in smooth muscle cells [8] and hepatocytes [9]. The suggested methodology allows researchers to detect local Ca^{2+} release events automatically while simultaneously classifying them in a few minutes without human intervention.

2. Methods

Fig. 1 shows our approach using a trained U-Net [10], a deep learning (DL) architecture, which autonomously detects, localizes, and classifies Ca^{2+} release events. The model, after its initial training on a manually annotated dataset comprising rapid confocal full-frame Ca^{2+} imaging recordings, operates independently. The dataset focuses on three types of subcellular Ca^{2+} release events found in cardiomyocytes, namely Ca^{2+} sparks, Ca^{2+} puffs, and Ca^{2+} waves, all of which were manually segmented as described below.

In the following, we detail the dataset acquisition and annotation protocol and then describe the model training, inference, and evaluation process.

2.1. Dataset of annotated Ca^{2+} release events

The high-frequency full-frame confocal image series used to train the DLM were generated from recordings of cardiomyocytes isolated from either C57Bl/6 mice or the $InsP_3R$ type II overexpressing mouse model [11] using the fluorescent Ca^{2+} indicator fluo-3. Detailed information is provided in the supplementary materials (section S1).

Manual annotation of Ca^{2+} release events in fast full-frame confocal recordings is labor-intensive. Causes of this complexity include substantial noise, the presence of out-of-focus events [12], and possible crosstalk mechanisms between local Ca^{2+} release events [6]. To speed up the process, two experts in the field utilized a multistep semi-automatic approach to annotate Ca^{2+} release events in the recordings. The workflow is shown in Fig. 2A, and corresponding sample images from the dataset are given in Fig. 2B. First, a custom Fiji macro [13], described in detail in the supplementary materials (section S1.3), generated binary masks for each recording. We effectively identified potential events from the background by extracting the connected components at each frame of these masks. We then merged connected components with high spatial overlap in consecutive frames to represent each Ca^{2+} release event by a region of interest (ROI) spanning multiple frames. Manual correction of annotations was often necessary, especially when noise caused a missed event in a frame. Similarly, some ROIs had to be manually split if they contained more than one event. This process resulted in masks where each ROI corresponded to an individual Ca^{2+} release event.

The detected events were then manually classified into one of four classes: *Ca^{2+} spark*, *Ca^{2+}puff*, *Ca^{2+} wave*, and *undefined*. Whenever the ROIs representing two Ca^{2+}puffs were spatially contiguous, they were merged to represent a single event. The *undefined* class included events that could not be accurately assigned to any other category, typically for events located at the edge of the recording (spatially or temporally) or when the Ca^{2+} signal was intertwined with artifacts. Events marked as *undefined* were ignored during the training procedure. Finally, each event was assigned a unique integer identifier. Ultimately, the annotation process produced two masks for each recording: a *classification mask* indicating the class of each pixel and an *events mask* indicating the event's identifier associated with each pixel (Fig. 2A, right).

The two annotators analyzed and classified each Ca^{2+} release event independently. In cases of disagreement, another expert in the field was asked to review the disputed labels and make a final decision. Whenever this decision was not possible, the ROI would be labeled as undefined.

Ca^{2+} release events were annotated in 43 recordings, with 34 recorded from atrial cells and nine from to ventricular cells. They constitute the dataset for training and evaluating the proposed DLM. The duration of the recordings ranges from 500 to 1900 frames (~3400 ms to

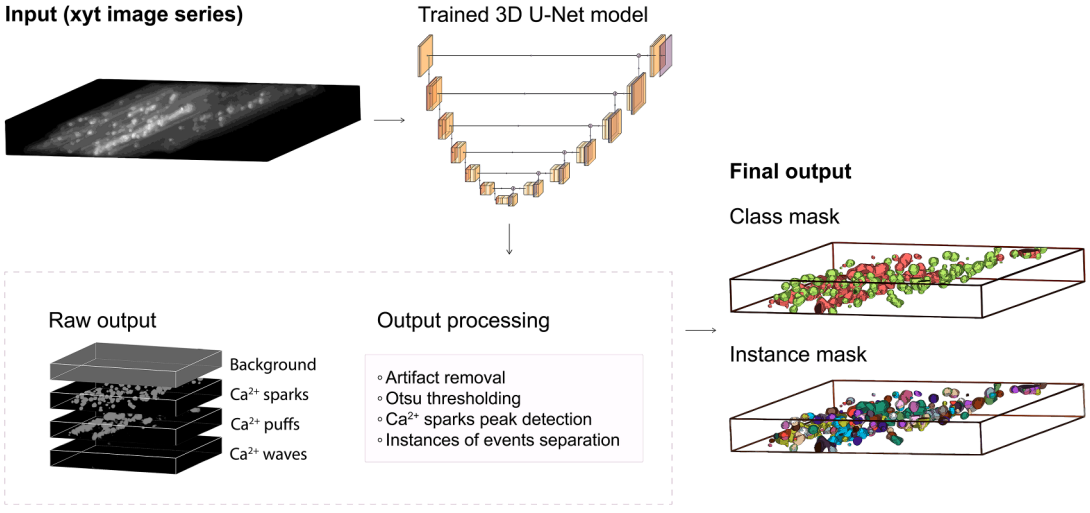


Fig. 1. Summarized workflow. Given a full-frame confocal Ca^{2+}fluorescence recording, the workflow produces two output movies. The first movie localizes and classifies various classes of Ca^{2+} release events, and the second enumerates individual event instances. The input movie is processed by the DLM, a complex algorithm consisting of several layers. The model's parameters are adjusted during training to address the problem (refer to Section 2.2.1 for additional details). The DLM produces pixel-wise probability distributions for each class (raw output). These distributions are post-processed through a sequence of steps. First, minor artifacts are removed, and a threshold is applied to obtain a binary segmentation mask. This mask is then combined with the output of the DLM to classify each pixel, resulting in a mask where the Ca^{2+} release events are classified (class mask). Then, the classification mask is further processed to identify peaks of Ca^{2+} sparks, enabling the separation of individual spark instances. Finally, the watershed algorithm is used to detect instances of Ca^{2+} puffs and Ca^{2+} waves. The resulting mask contains numbered regions corresponding to the different instances of the Ca^{2+} release events (instance mask).

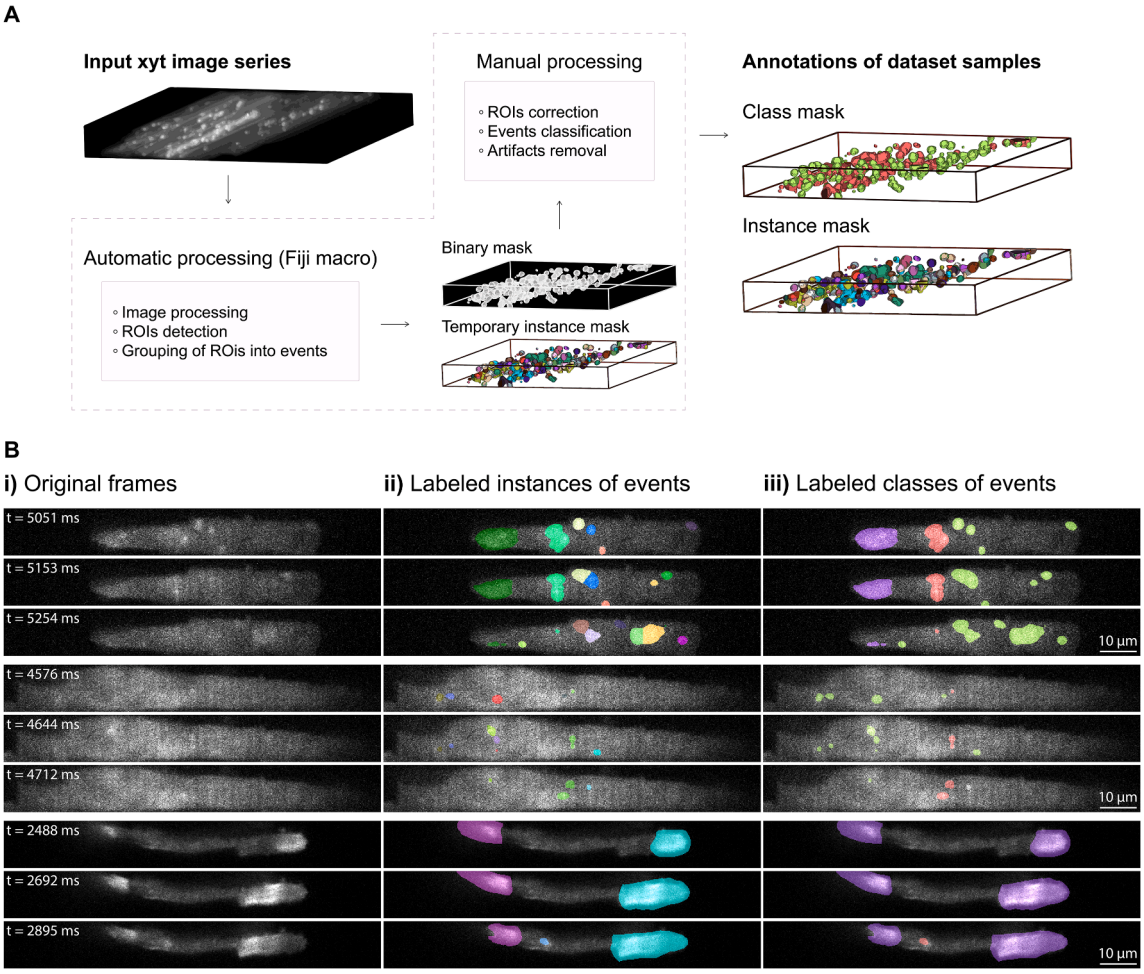


Fig. 2. Annotated full-frame image dataset of cardiomyocytes loaded with a Ca^{2+}-sensitive fluorescent dye. A) Workflow for the creation of the labels. The (semi-) automatic processing returns: 1. a mask representing the classified events' locations; 2. a mask where each event is separated and assigned a different integer (denoted by distinct colors in the figure). B) Sampled frames from the annotated dataset. i) Original frames from three recordings at three increasing time steps. ii) Annotated individual event instances resulting from the semi-automatic annotation approach. Each color represents a different Ca^{2+} release event. iii) Annotated frames after the manual classification of Ca^{2+}release events: green - Ca^{2+} sparks, red - Ca^{2+} puffs, purple - Ca^{2+} waves, and grey - unclassified event/artifact. As shown in the top example, two individual events of the same type can be contiguous.

~12,900 ms). Therefore, we annotated 35,443 frames in which we could detect approximately 1400 Ca^{2+} sparks, 300 Ca^{2+} puffs, and 25 Ca^{2+} waves. Fig. 2B shows examples of annotated Ca^{2+} release events with the resulting class and instance masks.

2.2. 3D U-Net architecture and model training

The annotated dataset of 43 recordings served to train and evaluate our automatic Ca^{2+} release event detection method. The fundamental component of this method is a DL network derived from the U-Net architecture [10]. The architecture was adapted to accept 3-dimensional inputs to suit our requirements, corresponding to two spatial dimensions and time (Fig. 1, [14]). Further details are given in the supplements S2.

2.2.1. Training procedure

The whole dataset was split with a ratio of 80 %/20 % for training and testing, respectively. Accordingly, 34 recordings were used for training and 9 for DLM testing. The U-Net model receives time segments of full-frame confocal Ca^{2+} recordings as input and generates a 4-dimensional output representing the probability distribution across four classes (background, Ca^{2+} spark, Ca^{2+} puff, Ca^{2+} wave) for each pixel.

Due to limited GPU memory, we adopted a sliding window strategy and extracted multiple overlapping time segments of 256 frames (~1740 milliseconds) from each training sequence with a step of 32 frames. Additional details about the sliding window approach are provided in the supplementary materials S2.1. During training, we applied data augmentation by randomly mirroring these segments along each spatial dimension (vertical and horizontal) [15]. We provide more information about data augmentation in the supplements S2.2.

Pixels of input images were normalized to the range [0, 1]. The normalization process was performed relative to the minimum value of the image and the maximum value achievable with a 16-bit pixel format (i.e., 65,535).

We used the Lovász-Softmax loss function [16], a continuous and differentiable surrogate for the Intersection over Union (IoU) score, which measures the overlap between two segmentation masks. The IoU score is calculated as the size of the intersection of the masks divided by the size of their union. This score ranges from 0 to 1, where 1 indicates perfect overlap (identical masks) and 0 denotes no overlap at all. A higher IoU score reflects better agreement between the compared regions. We excluded regions marked as undefined in the annotated dataset for loss computation and the first and last six frames of each input segment due to insufficient temporal context. Since the loss is computed for each class and later averaged over all classes, each type of event has the same impact on the learning procedure. This implicitly

handles the imbalance present in our dataset. The Adam optimizer [17] was used with a fixed learning rate of 10^{-4}. The model was trained for 100,000 iterations on batches of 4 time segments, requiring approximately 60 h of computation on a single NVIDIA GeForce RTX 3090 GPU. The code was implemented using Python 3.10 with the PyTorch framework version 2.0 [18].

2.2.2. Inference

The DLM model automatically detected intracellular Ca^{2+} events in two steps: segmentation and detection. The trained network produced a pixel-wise classification of the input recording during segmentation. The recordings were split into overlapping segments of 256 frames with a step of 32 frames. Predictions of time segments were merged to produce a probability map of the size of the input recording, except for the first and last six frames, which were ignored due to insufficient temporal context. Background pixels were determined by applying Otsu thresholding to these probability maps, a method that computes the threshold that minimizes intra-class intensity variance [19]. Then, each non-background pixel was assigned the highest probability class.

In the detection step, the obtained pixel-wise labels were used to detect the individual instances of Ca^{2+} release events (also simply called *instances* hereafter). First, individual Ca^{2+}puffs and Ca^{2+} waves were separated based on the connected components of their masks. Due to oscillations in the fluorescence signal, the ROIs of individual events may be missing in some frames, resulting in "holes" in the temporal dimension. Therefore, Ca^{2+}puff or Ca^{2+} wave instances were merged whenever there was a gap of 2 frames or less between two consecutive and spatially overlapping events. We then removed detections of Ca^{2+} puffs shorter than 35 ms (5 frames) and Ca^{2+} waves with a diameter smaller than 15 μm (75 pixels), as they typically correspond to artifacts.

The separation of individual Ca^{2+} sparks required additional steps. Specifically, Ca^{2+} sparks were separated using the 3D watershed algorithm, which takes a list of the peaks of the Ca^{2+} sparks as input. Peaks were identified as the local maxima of the original recordings masked by the binary mask of the Ca^{2+} spark class. Detected peaks were separated by a given minimal distance between them, defined as 1.8 μm (9 pixels) in the spatial dimensions and 20 ms (3 frames) in the time dimension. As with the other classes of events, spatio-temporally unreasonably "small" detections were removed in the last step. Specifically, we removed Ca^{2+} sparks ROIs shorter than 20 ms (3 frames) or with a diameter smaller than 0.6 μm (3 pixels) from segmentation masks, labelling their ROIs as background.

This approach can process a recording of 1000 frames, including loading from disk, in about 25 s when executed on a single NVIDIA GeForce RTX 3090.

The annotated dataset mentioned in Section 2.1. will be available for open access. Additionally, we provide DLM users with an **integrated graphical user interface** (GUI) .[1] This interface enables the simple loading of xyt-full frame images. In addition to classifying and splitting the detected ROIs into individual events, the GUI analyses the detected and classified events and reports specific parameters such as amplitude and FDHM. The code used to train the DLM, adaptable for other datasets, is also available online.[2] We would like to emphasize that by making the interface freely available, we are giving the interested scientific community the opportunity to test our proposed approach.

3. Results

The performance of our trained 3D U-Net model was evaluated on a test dataset of nine samples using three methods: pixel-based evaluation, instance-based evaluation (i.e., an evaluation based on the instances of individual events), and inter-rater variability evaluation, which is a comparison of the outputs of the DLM with human experts' opinions. Here, we illustrate the protocol followed for each evaluation type and the results we obtained for the dataset considered in this study.

3.1. Pixelwise evaluation of trained DLM

First, we evaluated our trained DLM by computing the average IoU scores for each class of events between the processed U-Net segmentation masks and the annotated class masks of the test dataset. We also computed the IoU scores for the binarized masks obtained after combining all Ca^{2+} release events as the foreground class. The results are presented in Table 1. The higher the IoU value, the more accurate the model used, i.e., the better the agreement between the predicted events and the manually annotated Ca^{2+} release events. The IoU score is susceptible to small changes in the dimension of the segmented regions, especially when they are very small. In our case, it results in low scores for local Ca^{2+} release events (Ca^{2+} sparks and Ca^{2+} puffs). For quantitative analysis, Fig. 3A compares the annotated masks and the predictions of our model for some sample frames.

3.2. Instance-based evaluation of trained DLM

To evaluate the DLM, the annotated instances of Ca^{2+} release events were matched with those predicted by the model. For each pair of annotated and predicted events, the intersection over minimum (IoMin) score was computed using the formula:

$$Score(Y,P) = \mathrm{IoMin}(Y,P) = \frac{|Y \cap P|}{\min(|Y|,|P|)},$$

where Y represents the binary mask of an annotated event, and P represents the binary mask of an event detected by the model. In essence, the IoMin score is a measure of overlap, similar to the IoU score. Specifically, IoMin divides the intersection area by the smaller of the two areas, rather than their union. This score ranges from 0 to 1, where 1 indicates that one of the mask entirely covers the other one and 0 denotes no overlap at all. This makes IoMin particularly effective in cases where one ROI area is smaller than the other.

Events were matched when their IoMin score was above 0.5. This threshold was chosen as it meaningfully represents when two overlapping ROIs represent the same event (see Fig. 4 for a graphical explanation).

To analyze the matching results, we report the distribution of matched detected events per annotated event (recall scores) and the distribution of matched annotated events per detected event (precision scores) in Table 2.

Our method achieved similar precision and recall scores for all types of Ca^{2+} events: nearly 60 % of all annotated events were detected, while from the set of predicted events, approximately 40 % were correct detections. Ca^{2+} waves were the only exception to this pattern, achieving perfect precision (100 %) for all the predicted events. It is important to mention that the numbers of events along the columns of Table 2 do not necessarily sum up to the total number of events, as some events might match several classes, as can be observed in Fig. 3B.

Table 3 summarizes the performance of the DLM. Notably, it achieved a minimum detection rate of 75 % for Ca^{2+} release events in the test dataset samples, identifying 234 out of 346 labeled events. Our

Table 1
IoU scores between manual annotations and predictions from our method averaged over the recordings of the test set.

Event class	IoU score
Ca^{2+} sparks	0.20
Ca^{2+} puffs	0.18
Ca^{2+} waves	0.28
Binarized segmentation	0.34

[1] https://github.com/r-janicek/xytCalciumSignalsDetection
[2] https://github.com/dottipr/sparks_project

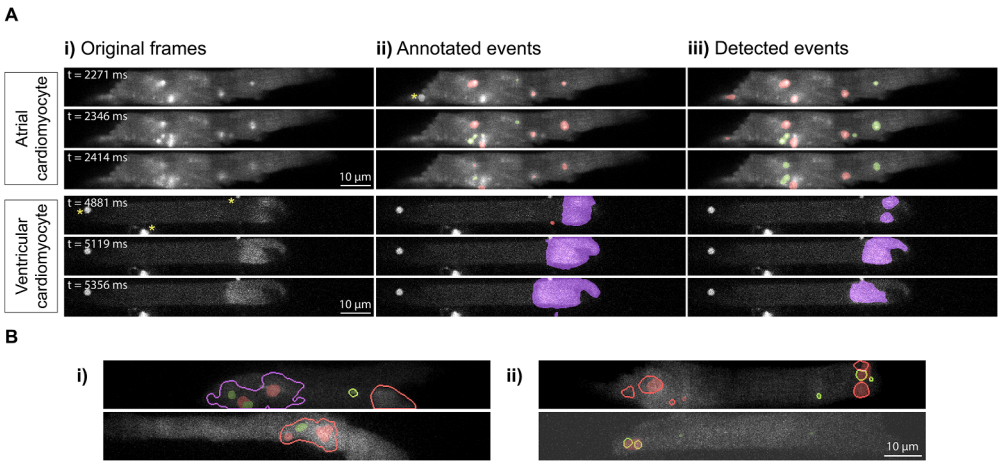


Fig. 3. Qualitative results of the proposed DLM. A) Comparison of annotations and DLM predictions on six selected frames extracted from two recordings of the test dataset: an atrial cell (top) and a ventricular cell (bottom). i) Original frames from two recordings at three different time steps. Asterisks on the first frame of the bottom cell denote artifacts resulting from dye loading, which were annotated as background in the training labels. Reported times correspond to the timing of the frame in the recording. (ii) Manual annotations (green: Ca^{2+} sparks; red: Ca^{2+} puffs; purple: Ca^{2+} waves). The first frame of the top cell contains a grey region (marked with an asterisk). It corresponds to an artifact or a Ca^{2+} release event that could not be classified. iii) Labels predicted by our method. B) Detected and labeled Ca^{2+} release events can overlap multiple events. Contours denote labeled events, while transparent colors denote detected events. Each color represents a different type of Ca^{2+} release event (green: Ca^{2+} sparks; red: Ca^{2+} puffs; purple: Ca^{2+} waves). i) On the top frame, a labeled Ca^{2+} wave is matched with two Ca^{2+} sparks and two Ca^{2+} puffs belonging to the model's detections. ii) On the top frame, two labeled Ca^{2+} sparks and two labeled Ca^{2+} puffs are matched with a unique Ca^{2+} puff. Similarly, in the bottom frame, two labeled Ca^{2+}sparks are matched with a Ca^{2+} puff.

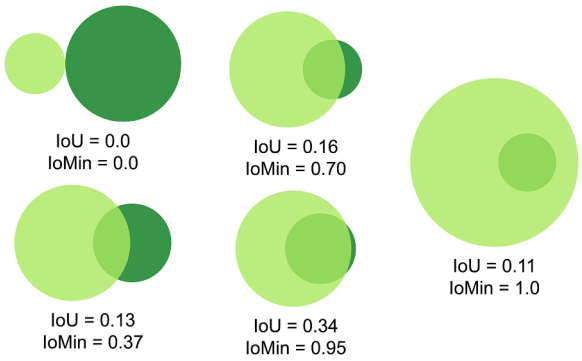


Fig. 4. Illustration of the reason for choosing the IoMin score over the IoU score. IoU heavily penalizes regions that overlap but have highly different sizes. The IoMin score is more suitable than the IoU score for our purpose, as it effectively captures when two events should be matched.

model consistently detects Ca^{2+} waves. Detailed definitions of all metrics are provided in the supplementary materials S5.

3.3. Assessment by experts

This analysis aims to determine if the segmentations produced by our model are discernible from those manually outlined by a human expert. Six experts in the field of Ca^{2+} signaling segmented the three types of Ca^{2+} release events in a sample of ten frames selected from the test dataset. We obtained the opinions of seven distinct observers: the six experts and the predictions of the U-Net model. Examples of the resulting segmented masks are shown in Fig. 5, and further information about the annotation procedure is given in the supplementary materials S3.

First, we estimated an agreement among all six experts on the selected frames. The statistical value that represents the reliability of the agreement among a group of observers was computed using Fleiss' kappa [20]. Results are shown in Table 4. We observed that the outcomes of image segmentation and classification of intracellular Ca^{2+} events exhibit considerable variability when performed by multiple human experts.

Then, we used the resulting annotations to compute the majority consensus among all other observers for each pixel. This procedure generated segmented frames representing the collective opinion of all other observers. It was performed by first including the DLM's opinion in the consensus computation and then excluding it. The agreement

Table 2
Matching of Ca^{2+} release events between DLM detections and annotations in the test dataset. Bold numbers highlight the correct matches. Note that the bold values in Table A correspond to recall (TP/(TP+FN)), and the bold values in Table B correspond to precision (TP/(TP+FP)), hence the title of the tables. Details about the computation of the values are provided in the supplementary materials S5. The sum of events in each class along the columns differs from the total number of events (annotated and detected) because an event may match more than one other event. In Table B, the total number of predicted events is reported, and the number of events that are later removed is indicated in parenthesis. The percentages in the other rows are computed based on the number of valid events.

A)Matches from annotated Ca^{2+} release events to predicted events (recall scores)

Class of annotated events	Ca^{2+} sparks	Ca^{2+} puffs	Ca^{2+} waves
Total number of annotated events	265	74	7
Matched with predicted Ca^{2+} sparks	**153 *(57.7 %)***	39 *(52.7 %)*	6 *(85.7 %)*
Matched with predicted Ca^{2+} puffs	24 *(9.1 %)*	**41 *(55.4 %)***	5 *(71.4 %)*
Matched with predicted Ca^{2+} waves	0 *(0 %)*	0 *(0 %)*	**4 *(57.1 %)***
Undetected events	91 *(34.3 %)*	19 *(25.7 %)*	1 *(14.3 %)*

B)Matches from predicted Ca^{2+} release events to annotated events (precision scores)

Class of predicted events	Ca^{2+} sparks	Ca^{2+} puffs	Ca^{2+} waves
Total number of events predicted by model	404 (17)	138 (7)	4 (0)
Matched with annotated Ca^{2+} sparks	**151 *(39.0 %)***	18 *(13.7 %)*	0 *(0 %)*
Matched with annotated Ca^{2+} puffs	69 *(17.9 %)*	**49 *(37.4 %)***	0 *(0 %)*
Matched with annotated Ca^{2+} waves	36 *(9.3 %)*	12 *(9.1 %)*	**4 *(100 %)***
Unmatched with any annotated events	132 *(34.1 %)*	57 *(43.5 %)*	0 *(0 %)*

Table 3
Model performance on test dataset by class and average across classes. The rate of detected events is the number of annotated events in each class detected by the DLM (including misclassified ones) divided by the number of annotated events in the same class. The rate of correct events is the number of correctly detected and classified events divided by the total number of events detected in the given class. E.g., for the row corresponding to Ca2+ sparks, the value indicates that out of all the Ca2+ sparks detected by our model (which account for 65.7 % of the total number of annotated Ca2+ sparks), 59.2 % were correctly classified. The F1-Score, which provides a balanced view of the classifier's accuracy, is the harmonic mean of precision and recall.

	% Detected	% Correct	F_1-Score
Ca^{2+} sparks	65.7 %	59.2 %	0.47
Ca^{2+} puffs	74.3 %	66.2 %	0.45
Ca^{2+} waves	85.7 %	100 %	0.73
Average	75.2 %	75.1 %	0.55

between all observers was summarized using Cohen's kappa [21]. The detailed protocol for agreement computation is given in the supplementary materials S3.

By computing the Kruskal-Wallis test on the seven groups of observers' kappa for both cases, we determined whether any statistically significant differences among them exist. We did not observe any significant differences in the agreement of the predictions of the DLM with the mean of the majority vote of the experts, neither in cases where the DLM's outcome was included ($p = 0.775$) nor excluded in the majority voting ($p = 0.863$). Fig. 6 illustrates the variability of the agreement values obtained within each group.

4. Discussion and limitations

It is important to acknowledge that Ca^{2+} release events can be represented by distinct, correct regions of interest (ROIs), and there is no canonical accurate method for labeling recordings. This assertion is bolstered by the significant variability observed among human observers. Regrettably, the IoU score's sensitivity to minor differences in small object shapes poses obstacles to robustly score agreement between ROIs detected with different methods (expert, group of experts, DLM) for this criterion since the event borders are not distinctly defined owing to anisotropic Ca^{2+} diffusion in cardiac myocytes. The U-Net frequently detects events with larger or smaller ROIs than the annotated ones proposed (take the Ca^{2+} wave present in the lower frames of Fig. 3A as an instance), resulting in "apparently" inaccurate performance of our model when evaluated by the IoU score. Fig. 4 offers a visual representation of how the IoU score is affected by differences in the areas of annotated and detected events. Therefore, assessing our model through event instance-based metrics is of interest. Nevertheless, as depicted in Fig. 3A, our model yields convincing qualitative results overall.

The event-wise analysis revealed that our model can identify more than 75 % of the annotated events while correctly classifying more than 75 % on average. Some misclassifications corresponded to complex Ca^{2+} release events that we did not annotate. Namely, our model could detect Ca^{2+} sparks on top of Ca^{2+} puffs that were labeled as Ca^{2+} puffs in our dataset. Sometimes, Ca^{2+} sparks and the underlying Ca^{2+} puffs were detected; sometimes, only the Ca^{2+}sparks on top were detected. For this reason, out of 74 annotated Ca^{2+} puffs, 39 were matched with Ca^{2+} sparks. Such Ca^{2+} release events were not annotated in our dataset

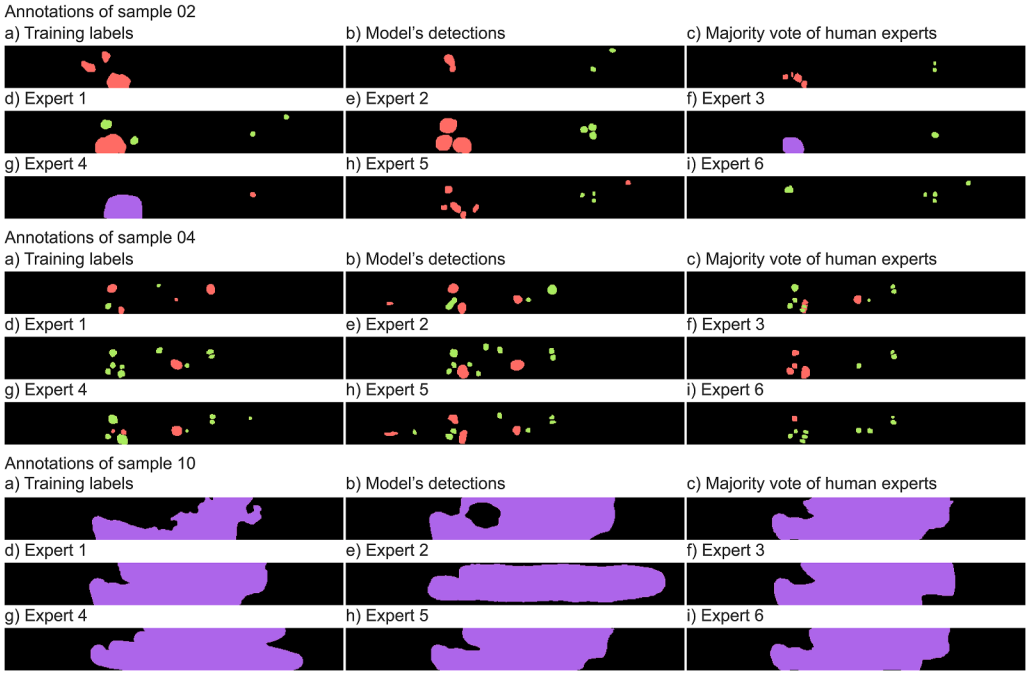


Fig. 5. Examples of segmented frames illustrating the high variability between human annotators. Each color corresponds to a different type of Ca^{2+}release event (green: Ca^{2+} sparks; red: Ca^{2+} puffs; purple: Ca^{2+} waves). For each of the three examples, the presented figures are a) segmented mask used for DLM training; b) segmentation provided by the DLM processed output; c) majority vote computed using the value that was selected by the largest number of experts per pixel; d)-i) masks segmented by the six experts. Sample 02 illustrates a fair agreement among experts (Fleiss' kappa is 0.263); Sample 04 illustrates a moderate agreement among experts (Fleiss' kappa is 0.499); Sample 10 illustrates a substantial agreement among experts (Fleiss' kappa is 0.776).

Table 4
Agreement among all six experts on the selected single frames extracted from the test dataset, assessed by Fleiss' kappa calculation.

Frame ID	01	02	03	04	05	06	07	08	09	10	Average
Group agreement (Fleiss' kappa)	0.400	0.263	0.393	0.499	0.411	0.210	0.140	0.635	0.817	0.776	0.454

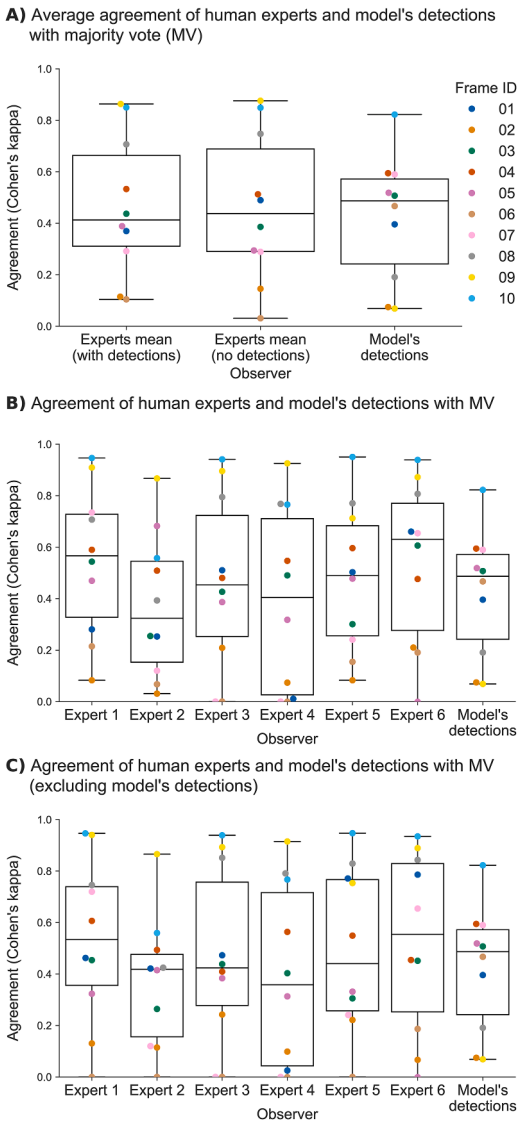


Fig. 6. Comparison of each expert with the majority vote. The agreement of the segmentation of the DLM (dark grey) with the mean of the majority vote of the experts (light grey) is in the same range as the agreement between human experts and their corresponding majority vote, including and excluding the segmented masks of the DLM. A) The agreement average on each frame is computed over all experts, both when the DLM opinion is excluded from the majority vote and when it is included. B) Each human expert (light grey) is compared with the majority vote of all other human experts, and the masks segmented by the DLM, while the masks segmented by the DLM (dark grey) are compared with all the human experts. C) Each human expert (light grey) is compared with the majority vote of all other human experts, and the DLM (dark grey) is compared with all the human experts. In all panels, distinct colors in the plots represent the mean kappa values of the different frames, as denoted in the legend on the right-hand side of Figure A). Statistical significance was assessed using the Kruskal-Wallis test.

because annotated ROIs were created by thresholding a denoised version of the original recording, which sometimes simplified and limited the granularity of our annotations. Specifically, our annotation methodology did not allow for events fully contained in other events.

Generally, the quality of annotations has a significant impact on the results obtained from DL approaches. We acknowledge that our annotations were mainly based on amplitude, and we were unable to include partly highly complex instances, such as the Ca^{2+} sparks on top of Ca^{2+} puffs mentioned earlier in this project. Consequently, the DLM's segmented masks are somewhat biased by our annotation procedure. Including additional parameters might produce more comprehensive results. Unfortunately, optimal parameters remain undetermined.

As mentioned earlier, a size discrepancy exists between labeled and detected events. Specifically, the former is generally larger due to the border of our annotations being determined by processing the recording with a threshold. As a result, annotated Ca^{2+} waves often include local Ca^{2+} release events that emerge during the wave's dissipating phase. Therefore, 36 detected Ca^{2+} sparks and 12 Ca^{2+} puffs were identified with annotated Ca^{2+} waves.

Our analysis showed that the temporal context plays an important role in recognizing Ca^{2+} release events, i.e., events that occurred very early in the auger recording are not detectable or are more difficult to detect than events that occurred at the end of the measurement, suggesting that a more extensive temporal context is required (Table 2A). The DLM also tends to categorize parts of already labeled Ca^{2+} waves as Ca^{2+} puffs, often occurring towards the end of the Ca^{2+} waves where their speed decreases [22]. In contrast, Ca^{2+} waves are accurately detected at their onset.

The data utilized in this study presents significant challenges due to various factors, including the effect of Ca^{2+} diffusion on border delineation. The diverse noise types in the data and its anisotropic nature pose difficulties for the 3D U-Net model. Temporal information plays a distinct role from spatial information, but the 3D U-Net's convolution processes each dimension uniformly.

We have shown that standard evaluation metrics, such as the IoU score, do not reflect the quality of the model's detections well. For this reason, we further evaluated our approach by asking six experts to annotate ten frames selected from the test dataset and compare their annotations with the detections of our model. Fig. 5 provides visual examples illustrating the differences among various annotations.

Table 4 shows that some frames present a relatively low agreement. These frames belong to recordings with a low number of small events, and delineating local Ca^{2+} release events is complex (e.g., sample 02 in Fig. 5). Conversely, we observed high agreement among raters on frames containing Ca^{2+} waves, as these are easier to identify than local Ca^{2+} release events (e.g., sample 10 in Fig. 5). Finally, the agreement is moderate on the frames with many local Ca^{2+} release events (e.g., sample 04 in Fig. 5). A reason for the low agreement on some frames and the misclassification of some local Ca^{2+} release events performed by our model could be that some events can also correspond to events that do not originate at the focal plane. When Ca^{2+} diffuses into the focal plane, it may give rise to events with a distorted signal [7].

Overall, our evaluation demonstrates the comparability of the performance of our DLM with human experts. Indeed, our inter-observer analysis revealed that the detections provided by our model lie within the same range of variability as human experts. This means that a differentiation between the segmented frames produced by human experts and those generated by our model is not feasible. Moreover, the proposed DLM offers practical advantages compared to manual annotation. Indeed, while the annotation of the frames by the experts took one hour per sample on average, the model can efficiently process a 1000-frame video in approximately 25 s, significantly reducing the time required for analysis.

The fact that DL has shown in recent years promising achievements across many fields, including medical and biological applications [23–25], supported the idea of trying a DL-based methodology on rapid full-frame confocal imaging data. Several methods for identifying local Ca^{2+} release events in line-scan images are available [26–29]. Some of these approaches already apply DL [28] and machine learning [30] approaches to enable the detection of Ca^{2+} release events. However, they do not handle full-frame confocal imaging data or distinguish between different local Ca^{2+} release event types. Additionally, most tools based on full-frame images require manual and time-consuming steps. For instance, although Juicer software [31] allows pixel-by-pixel event classification on full-frame confocal imaging, it requires several days to analyze whole cells. Finally, CaCLEAN [32] and iSpark [33]

automatically detect Ca^{2+} release sites in full-frame confocal imaging, but unlike our model, they do not perform classification. Similarly, the method proposed in [34] localizes Ca^{2+} sparks using statistical testing to discriminate events (i.e., Ca^{2+} sparks) from noise and other type of events, however, it requires the choice of several parameters. A fast and fully automatic DL-based approach for localizing and classifying local Ca^{2+} release events in full-frame confocal imaging is thus currently only available with our new DLM based approach reported here.

5. Conclusion

Local intracellular Ca^{2+} release events are present in almost all excitable cell types and are involved in various cellular regulatory processes and pathways. Analyzing these local Ca^{2+} release events is a prerequisite for a fundamental understanding of the local subcellular processes and their regulatory function, particularly under pathophysiological conditions.

This study successfully demonstrated the effectiveness and efficiency of a DLM in detecting and classifying intracellular Ca^{2+} events, exemplified in the context of cardiomyocytes, eliminating human intervention and bias. The DLM can, in principle, be adapted and applied to all full-frame confocal data and datasets representing Ca^{2+} signals. Previously analyzed data could be used to train DLM datasets collected from other cell types and in different contexts.

In conclusion, the proposed DLM offers the advantage of rapid detection and classification of Ca^{2+} release events, enabling more efficient data processing than previous methods. The DLM approach can detect more than 75 % of the Ca^{2+} release events independently of the noise characteristics in the original recordings. Notably, the model can locate most annotated Ca^{2+} release events and identify some instances not recorded by manual annotations within 5 min for the complete test dataset running on a single NVIDIA GeForce RTX 3090 GPU.

The result of the DLM analysis is comparable with the manual outcome obtained by field-experts analysis, which was previously considered the gold standard for Ca^{2+} event classification. Even if users choose to verify all events detected by the model manually, the analysis can be completed within a reasonable timeframe. Furthermore, the reproducibility of the DLM's results is assured due to the deterministic nature of the trained 3D U-Net architecture.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors did not use generative AI or AI-assisted technologies in the development of this manuscript.

Author agreement

I, Marcel Egger submitting this manuscript on behalf of myself and my co-authors. We hereby submit the manuscript titled "A Deep Learning-Based Approach for Efficient Detection and Classification of Local Ca^{2+} Release Events in Full-Frame Confocal Imaging" for publication in Cell Calcium.

This statement is to certify that all Authors have seen and approved the manuscript being submitted. We warrant that the article is the Authors' original work. We warrant that the article has not received prior publication and is not under consideration for publication elsewhere. On behalf of all Co-Authors, the corresponding Author shall bear full responsibility for the submission. We also confirm that all the research meets the ethical guidelines, including adherence to the legal requirements of the study country. We attest to the fact that all Authors listed on the title page have contributed significantly to the work, have read the manuscript, attest to the validity and legitimacy of the data and its interpretation, and agree to its submission the Cell Calcium.

CRediT authorship contribution statement

Prisca Dotti: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Miguel Fernandez-Tenorio:** Supervision, Writing – review & editing. **Radoslav Janicek:** Writing – review & editing, Supervision, Software, Methodology. **Pablo Márquez-Neila:** Writing – review & editing, Methodology. **Marcel Wullschleger:** Formal analysis. **Raphael Sznitman:** Writing – review & editing, Supervision, Methodology. **Marcel Egger:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Data availability

All data used for the analysis will be made available through the public data repository Zenodo.org (DOI: https://doi.org/10.5281/zenodo.10391727).

Funding

This work was supported by the Swiss National Science Foundation (310030_185211) and Novartis Res. Foundation to M.E.

Acknowledgments

We would like to thank Christian Soeller and Ernst Niggli for their helpful comments on the manuscript, and Natalia Shirokova for her valuable contribution to this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ceca.2024.102893.

References

[1] M.J. Berridge, M.D. Bootman, H.L. Roderick, Calcium signalling: dynamics, homeostasis and remodelling, Nat. Rev. Mol. Cell Biol. 4 (2003) 517–529, https://doi.org/10.1038/nrm1155.
[2] M.D. Stern, Theory of excitation-contraction coupling in cardiac muscle, Biophys. J. 63 (1992) 497–517.
[3] D.M. Bers, Cardiac excitation–contraction coupling, Nature 415 (2002) 198–205, https://doi.org/10.1038/415198a.
[4] H. Cheng, W.J. Lederer, M.B. Cannell, Calcium sparks: elementary events underlying excitation-contraction coupling in heart muscle, Science 262 (1993) 740–744, https://doi.org/10.1126/science.8235594.
[5] I. Parker, Y. Yao, Regenerative release of calcium from functionally discrete subcellular stores by inositol trisphosphate, Proc: Biol. Sci. 246 (1991) 269–274.
[6] M. Wullschleger, J. Blanch, M. Egger, Functional local crosstalk of inositol 1,4,5-trisphosphate receptor- and ryanodine receptor-dependent Ca^{2+} release in atrial cardiomyocytes, Cardiovasc. Res. 113 (2017) 542–552, https://doi.org/10.1093/cvr/cvx020.
[7] E. Niggli, N. Shirokova, A guide to sparkology: the taxonomy of elementary cellular Ca^{2+} signaling events, Cell Calcium 42 (2007) 379–387, https://doi.org/10.1016/j.ceca.2007.02.010.
[8] D.C. Hill-Eubanks, M.E. Werner, T.J. Heppner, M.T. Nelson, Calcium signaling in smooth muscle, Cold Spring Harb. Perspect. Biol. 3 (2011) a004549, https://doi.org/10.1101/cshperspect.a004549.
[9] M.J. Amaya, M.H. Nathanson, Calcium signaling in the liver, Compr. Physiol. 3 (2013) 515–539, https://doi.org/10.1002/cphy.c120013.
[10] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.
[11] H. Nakayama, I. Bodi, M. Maillet, et al., The IP3 receptor regulates cardiac hypertrophy in response to select stimuli, Circ. Res. 107 (2010) 659–666, https://doi.org/10.1161/CIRCRESAHA.110.220038.

[12] H. Cheng, W.J. Lederer, Calcium Sparks, Physiol. Rev. 88 (2008) 1491–1545, https://doi.org/10.1152/physrev.00030.2007.
[13] J. Schindelin, I. Arganda-Carreras, E. Frise, et al., Fiji: an open-source platform for biological-image analysis, Nat. Methods 9 (2012) 676–682, https://doi.org/10.1038/nmeth.2019.
[14] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, (2016). http://arxiv.org/abs/1606.06650 (Accessed May 24, 2023).
[15] C. Shorten, T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, J. Big. Data 6 (2019) 60, https://doi.org/10.1186/s40537-019-0197-0.
[16] M. Berman, A.R. Triki, M.B. Blaschko, The Lovasz-Softmax Loss: a tractable surrogate for the optimization of the intersection-over-union measure in Neural networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 4413–4421, https://doi.org/10.1109/CVPR.2018.00464.
[17] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, (2017). https://doi.org/10.48550/arXiv.1412.6980.
[18] A. Paszke, S. Gross, F. Massa, et al., PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, Curran Associates, Inc., 2019, in: https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html. accessed May 24, 2023.
[19] N. Otsu, A Threshold Selection Method from Gray-Level Histograms, IEEE Trans. Syst. Man Cybern. 9 (1979) 62–66, https://doi.org/10.1109/TSMC.1979.4310076.
[20] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (1971) 378–382, https://doi.org/10.1037/h0031619.
[21] M. McHugh, Interrater reliability: the kappa statistic, Biochemia Medica : Časopis Hrvatskoga Društva Medicinskih Biokemičara /HDMB 22 (2012) 276–282, https://doi.org/10.11613/BM.2012.031.
[22] H. Cheng, M.R. Lederer, W.J. Lederer, M.B. Cannell, Calcium sparks and $[Ca^{2+}]_i$ waves in cardiac myocytes, Am. J. Physiol.-Cell Physiol. 270 (1996) C148–C159, https://doi.org/10.1152/ajpcell.1996.270.1.C148.
[23] C. Ounkomol, S. Seshamani, M.M. Maleckar, F. Collman, G.R. Johnson, Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy, Nat. Methods 15 (2018) 917–920, https://doi.org/10.1038/s41592-018-0111-2.
[24] T. Falk, D. Mai, R. Bensch, et al., U-Net: deep learning for cell counting, detection, and morphometry, Nat. Methods 16 (2019) 67–70, https://doi.org/10.1038/s41592-018-0261-2.
[25] C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, Cellpose: a generalist algorithm for cellular segmentation, Nat. Methods 18 (2021) 100–106, https://doi.org/10.1038/s41592-020-01018-x.
[26] E. Picht, A.V. Zima, L.A. Blatter, D.M. Bers, SparkMaster: automated calcium spark analysis with ImageJ, Am. J. Physiol. Cell Physiol. 293 (2007) C1073–C1081, https://doi.org/10.1152/ajpcell.00586.2006.
[27] H. Cheng, L.S. Song, N. Shirokova, et al., Amplitude distribution of calcium sparks in confocal images: theory and studies with an automatic detection method, Biophys. J. 76 (1999) 606–617.
[28] S. Yang, R. Li, J. Chen, Z. Li, Z. Huang, W. Xie, Calcium spark detection and event-based classification of single cardiomyocyte using deep learning, Front. Physiol. 12 (2021). https://www.frontiersin.org/articles/10.3389/fphys.2021.770051. accessed May 24, 2023.
[29] J. Tomek, M. Nieves-Cintron, M.F. Navedo, C.Y. Ko, D.M. Bers, SparkMaster 2: a new software for automatic analysis of calcium spark data, Circ. Res. 133 (2023) 450–462, https://doi.org/10.1161/CIRCRESAHA.123.322847.
[30] W.A. Leigh, G. Del Valle, S.A. Kamran, et al., A high throughput machine-learning driven analysis of Ca2+ spatio-temporal maps, Cell Calcium 91 (2020) 102260, https://doi.org/10.1016/j.ceca.2020.102260.
[31] A. Illaste, M. Wullschleger, M. Fernandez-Tenorio, E. Niggli, M. Egger, Automatic detection and classification of Ca2+ release events in line- and frame-scan images, Biophys. J. 116 (2019) 383–394, https://doi.org/10.1016/j.bpj.2018.12.013.
[32] Q. Tian, L. Kaestner, L. Schröder, J. Guo, P. Lipp, An adaptation of astronomical image processing enables characterization and functional 3D mapping of individual sites of excitation-contraction coupling in rat cardiac muscle, Elife 6 (2017) e30425, https://doi.org/10.7554/eLife.30425.
[33] Q. Tian, L. Schröder, Y. Schwarz, et al., Large scale, unbiased analysis of elementary calcium signaling events in cardiac myocytes, J. Mol. Cell. Cardiol. 135 (2019) 79–89, https://doi.org/10.1016/j.yjmcc.2019.08.004.
[34] T. Bányász, Y. Chen-Izu, C.W. Balke, L.T. Izu, A new approach to the detection and statistical classification of Ca^{2+} sparks, Biophys. J. 92 (2007) 4458–4465, https://doi.org/10.1529/biophysj.106.103069.