Docent application

H: Selected publications

Olof Mogren

November 2025

H. Selected publications

The Accuracy Cost of Weakness: A Theoretical Analysis of Fixed-Segment Weak Labeling for Events in Time

John Martinsson *john.martinsson@ri.se*
Computer Science
RISE Research Institutes of Sweden
Centre for Mathematical Sciences
Lund University

Tuomas Virtanen *tuomas.virtanen@tuni.fi*
Signal Processing Research Centre
Tampere University

Maria Sandsten *maria.sandsten@matstat.lu.se*
Centre for Mathematical Sciences
Lund University

Olof Mogren *olof.mogren@ri.se*
RISE Research Institutes of Sweden
Swedish Centre for Impacts of Climate Extremes (climes)
Climate AI Nordics

Reviewed on OpenReview: *https://openreview.net/forum?id=tTw8wXBQ18*

Abstract

Accurate labels are critical for deriving robust machine learning models. Labels are used to train supervised learning models and to evaluate most machine learning paradigms. In this paper, we model the accuracy and cost of a common weak labeling process where annotators assign presence or absence labels to fixed-length data segments for a given event class. The annotator labels a segment as "present" if it sufficiently covers an event from that class, e.g., a birdsong sound event in audio data. We analyze how the segment length affects the label accuracy and the required number of annotations, and compare this fixed-length labeling approach with an oracle method that uses the true event activations to construct the segments. Furthermore, we quantify the gap between these methods and verify that in most realistic scenarios the oracle method is better than the fixed-length labeling method in both accuracy and cost. Our findings provide a theoretical justification for adaptive weak labeling strategies that mimic the oracle process, and a foundation for optimizing weak labeling processes in sequence labeling tasks.

1 Introduction

In supervised machine learning, labeled datasets are required for training and evaluation. During evaluation, the accuracy of the labels determine the quality of the analysis. However, in practice, labels often contain noise that varies with the input sample and label type. Noisy training labels present a persistent challenge in machine learning (Liang et al., 2009; Song et al., 2022). Deep learning models, in particular, are prone to overfitting noisy labels, raising questions about the nature of generalization (Zhang et al., 2021). Regularization techniques such as dropout (Srivastava et al., 2014), data augmentation (Shorten & Khoshgoftaar,

2019), and weight decay (Krogh & Hertz, 1991) mitigate overfitting but fail to eliminate the performance gap between training on noisy versus clean labels (Song et al., 2022).

Beyond the well-documented challenges posed by noisy training labels, inaccurate evaluation labels present a significant, yet often overlooked, obstacle to reliable machine learning. When evaluation metrics are computed against noisy ground truth, the apparent "best" performing model might simply be the one that most closely reproduces the noise present in the evaluation set, rather than exhibiting superior generalization capabilities. This very issue, where noisy evaluation labels can lead to the rejection of models that have learned the true clean label distribution, is a central concern addressed by Görnitz et al. (2014). This can lead to the selection of suboptimal models that perform well on the flawed evaluation data but generalize poorly to unseen, cleaner data or data from real-world applications. Consequently, performance benchmarks can be inflated and misleading, hindering meaningful comparisons between different approaches. Therefore, understanding the characteristics of label noise, not just in the training data but also in the evaluation data, is crucial for developing and selecting models that are truly effective and robust.

Labels are typically obtained through human annotation, a process that involves significant time and financial investment, particularly for complex data like audio or time-series signals. In this work, we consider a form of weak labeling where the annotator assigns presence or absence labels to predefined data segments. This offers a practical and cost-effective approach for annotating large audio datasets (Martin-Morato & Mesaros, 2023). To reduce cost, weak labels avoid specifying precise boundaries within the data segments, focusing instead on general presence or absence of the target class. However, this simplification introduces noise into the labels, especially for data with time-varying characteristics, such as audio signals, where events can occur intermittently within the labeled segment (Turpault et al., 2021). Understanding and mitigating this noise is critical to effectively leverage weak labels in downstream applications (Kumar & Raj, 2016).

The noise in weak labels can be categorized into two types: class label noise (mislabeling event presence or absence in a segment) and segment label noise (mislabeling due to misaligned segment boundaries). While class label noise has been extensively studied (Song et al., 2022; Zhang et al., 2021), the effects of segment label noise remain underexplored. This type of noise significantly affects tasks such as sound event detection (Hershey et al., 2021; Turpault et al., 2021; Shah et al., 2018) and medical image segmentation (Yao et al., 2023). Strategies like pseudo-labeling (Dinkel et al., 2022), robust loss functions (Fonseca et al., 2019), and adaptive pooling operators (McFee et al., 2018) aim to address challenges when training on weak labels. However, fully understanding the impact of weak labels requires quantifying their accuracy (Shah et al., 2018; Turpault et al., 2021).

While adaptive annotation methods have been explored in some domains, fixed-segment (FIX) annotation remains the de facto standard for large-scale sound event datasets due to its simplicity and scalability. Foundational datasets such as AudioSet (Gemmeke et al., 2017a), CHIME (Foster et al., 2015b), OpenMIC-2018 (Humphrey et al., 2018a), YBSS-200 (Singh & Joshi, 2019), and SONYC (Bello et al., 2019), as well as more recent datasets such as VGG Sound (Chen et al., 2020), MATS (Morato & Mesaros, 2021), and MAESTRO Real (Morato et al., 2023) continue to rely on fixed segments. A practical motivation is that for many sound sources, precise boundaries are inherently ambiguous: a passing car or fading background noise lacks sharp onsets and offsets, and audio scenes frequently contain overlapping sources. These challenges make FIX annotation a practical default even when adaptive approaches are conceptually appealing.

In this paper we use the term weak labeling exclusively to refer to temporal ambiguity introduced by segment-level human annotation. We do not use the term in the sense of pseudo-labels or automatically generated annotations.

Current methods typically estimate label noise rates *after* collecting labels (Song et al., 2022), employing techniques like noise transition matrices (Li et al., 2021) or cross-validation (Chen et al., 2019). In contrast, predicting label noise rates *before* data collection remains largely unexplored. This is particularly challenging when the noise stems from human annotators, as it is difficult to formalize. In cases involving partially automated processes, however, the noise introduced by the automated component can often be modeled under specific assumptions.

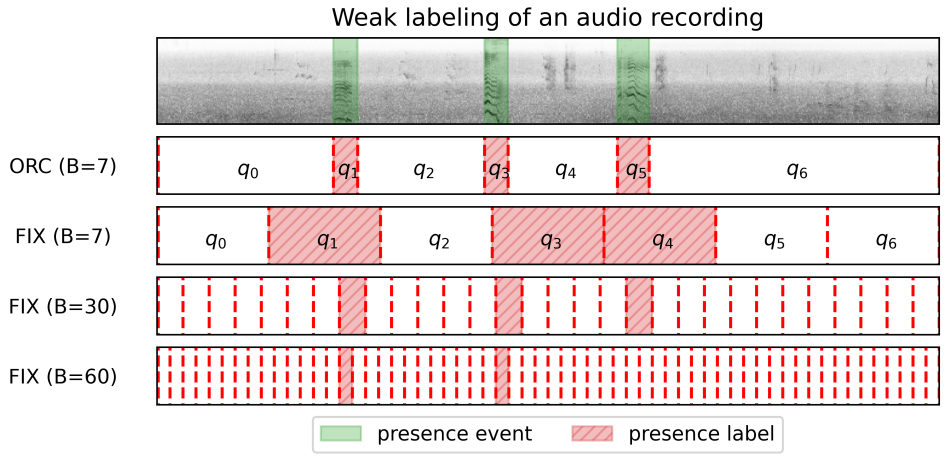


Figure 1: Resulting presence (red) and absence (white) labels from ORC and FIX weak labeling for an audio recording with three presence events (green). ORC weak labeling assigns labels to ground truth segments, achieving perfect label accuracy with $B = 7$ labels. FIX weak labeling, shown for different segment lengths ($B = 7$, $B = 30$, $B = 60$), introduces segment label noise as segments misalign with events. Longer segments reduce annotation cost but increase noise, while shorter segments align better but require more annotations. Note that too short segments ($B = 60$) may lead to the annotator missing the presence of the event because it does not cover a large enough fraction of it.

In this work, we model the automated component of a commonly used weak labeling method for segmentation tasks: fixed-length weak labeling (FIX). We quantify the segment label noise of this process, and study the expected label accuracy. This method, commonly employed in sound event detection, involves annotators providing presence or absence labels for fixed-length segments of the data (automated component), rather than specifying precise event boundaries. By simplifying the labeling process, FIX weak labeling reduces annotation effort but introduces segment label noise when segments misalign with the actual onsets and offsets of events. To benchmark this approach, we compare it to an oracle weak labeling method, ORC weak labeling, which assigns presence or absence labels to segments derived using the true onsets and offsets of the events.

Figure 1 illustrates the trade-offs between annotation cost and label accuracy for the ORC and FIX weak labeling methods. ORC weak labeling achieves perfect label accuracy by aligning the segments with the ground truth presence events (green) using a minimal number of annotated segments. In contrast, FIX weak labeling shows varying accuracy depending on segment length: shorter segments improve alignment ($B = 30$) but require more annotations, while longer segments ($B = 7$) reduce cost at the expense of accuracy. In addition, too short segments ($B = 60$) can lead to the annotator missing event presence. These trade-offs are central to understanding how FIX weak labeling can be used effectively. By analyzing the FIX weak labeling method, we provide a theoretical framework to guide data collection efforts.

Our analysis shows that FIX weak labeling systematically introduces segment noise, and we provide closed-form expressions for expected label accuracy and the optimal segment length. These results establish a quantitative baseline for annotation design, highlight the limits of non-adaptive strategies, and motivate the development of adaptive methods that approximate oracle labeling.

In summary, our contributions include:

- Closed-form expressions for label accuracy and annotation cost in FIX and ORC weak labeling, made tractable by assuming an annotator model and a simplified data distribution.
- A simulation study demonstrating that our theoretical framework generalizes to more complex data distributions and serves as an upper bound for the accuracy of FIX weak labeling.

- A theoretical foundation for developing adaptive weak labeling methods that better approximate ORC weak labeling, such as (Martinsson et al., 2024) for sound event detection and (Kim et al., 2023) for image segmentation.

Our analysis focuses on one-dimensional data, and the assumptions are justified by common characteristics in bioacoustic sound events. These time-localized, non-stationary animal vocalizations often require annotators to hear significant portions of the sound to assign accurate presence labels. Note, however, that while our framework is tailored to this domain, the principles extend to annotation of events in time in other data that shares these characteristics.

2 Problem Setting

The analysis is framed within a multi-pass binary labeling setting. Here, an annotator assigns binary labels (presence or absence) to data segments based on the occurrence of specific sound events. The annotator model abstracts how an annotator interacts with data by labeling segments, without requiring precise knowledge of event boundaries. While inspired by time-localized and non-stationary sound events, this framework is generalizable to any time series with similar characteristics.

It's important to emphasize that, in this weak labeling setting, the concept of overlapping events is not explicitly modeled. Overlapping events from the same class are treated as a single, longer presence event, because presence/absence labels cannot differentiate between individual event instances. For instance, in an audio recording with two birds calling simultaneously, this weak labeling framework simplifies the overlap into a single 'present' event. While this simplification is necessary when studying weak labeling in this setting, it fundamentally restricts our ability to resolve polyphony (the identification of multiple overlapping sound events). We leave the exploration of annotator models capable of providing richer labels to future work; this is beyond the scope of the current study.

For events of different classes occurring simultaneously, annotation is typically carried out in a multi-pass setup: each class is annotated independently, often by different annotators. Overlaps across classes are therefore not mutually exclusive—both events can be marked present within the same temporal window. The results from this paper are still valid when viewed for each class separately.

2.1 The Assumed Data Distribution

A sound event e is defined by its start time $a_e \in \mathbb{R}$, end time $b_e \in \mathbb{R}$, and class $c_e \in \mathcal{C}$, denoted as $e = (a_e, b_e, c_e)$. Audio recordings are assumed to have finite length T, and we assume that the events are uniformly distributed locally in time (see Section 4.2 and Appendix A.4 for more details).

2.2 The Assumed Annotator Model

For a given sound event class $c \in \mathcal{C}$, the annotator decides the presence or absence of an event e of class c in a data segment $q = (a_q, b_q)$, where $d_q = b_q - a_q$ is the fixed-length of the segment. We will refer to q as a query segment because it is queried for a presence or absence label. Let $l_q \in \{0, 1\}$ denote the weak label indicated by the annotator for query segment q, where $l_q = 1$ indicates presence of an event of class c in q and $l_q = 0$ indicates absence of that event class in q. Detecting the presence of an event requires observing a sufficient fraction of the event within the query segment, formalized as follows:

Definition 1. The *event fraction* is the fraction of the total event duration $d_e = b_e - a_e$ that overlaps with the query segment q,

$$h(e, q) = \frac{|e \cap q|}{d_e}, \tag{1}$$

where $e \cap q$ is the intersection of (a_e, b_e) and (a_q, b_q).

Definition 2. The *presence criterion* $\gamma \in (0, 1]$ is the minimum event fraction required for the annotator to detect the presence of e in q,

$$h(e, q) \geq \gamma. \tag{2}$$

The annotator assigns a presence label ($l_q = 1$) to q if there is sufficient overlap with any presence event e of class c ($h(e, q) \geq \gamma$); otherwise, it assigns an absence label ($l_q = 0$).

The parameter γ reflects the annotator's sensitivity: lower γ values indicate sensitivity to smaller event fractions, while higher values require larger fractions. This model of perceptual ability is particularly suited for non-stationary events (e.g., a specific birdsong) where recognizing a relative portion of the event's structure is key, as opposed to stationary sounds (e.g., an engine hum) which might be identified after a fixed absolute duration.

This framework captures variability in annotator behavior. For example, detecting "human speech" or "bird song" may only require hearing a small fraction of the event (γ closer to 0), while recognizing specific phrases or bird species might demand a near-complete observation (γ closer to 1). The value of γ thus depends on the annotator and the complexity of the event class. This model provides a flexible yet precise way to simulate annotator behavior and quantify their labeling performance. However, it is important to note that this model is deterministic, focusing on temporal alignment between events and the query segment. In practice, human annotation often involves stochastic factors, such as variability in perception and judgment, which are not explicitly modeled here.

2.3 Label Accuracy

Label accuracy measures the alignment between annotator-provided labels and ground truth labels:

Definition 3. The label accuracy is defined as

$$F(e, q, \gamma) = \begin{cases} \frac{|e \cap q|}{d_q}, & \text{if } l_q = 1, \\ \frac{d_q - |e \cap q|}{d_q}, & \text{if } l_q = 0. \end{cases} \tag{3}$$

For instance, consider a 3-second query segment ($d_q = 3$) that overlaps exactly one second ($|e \cap q| = 1$) with a 2-second sound event ($d_e = 2$) of the class bird song ($c =$ "bird song"). The annotator assigns a presence label ($l_q = 1$) with label accuracy $\frac{|e \cap q|}{d_q} = \frac{1}{3}$ if half or less of the event needs to be in the query segment ($\gamma \leq 0.5$). Contrary, the annotator assigns an absence label ($l_q = 0$) with label accuracy $\frac{d_q - |e \cap q|}{d_q} = \frac{3-1}{3} = \frac{2}{3}$ if more than half of the event ($\gamma > 0.5$) needs to be in the query segment. This formulation isolates the segment label noise ($1 - F(e, q, \gamma)$) introduced by the automated component (fixed-length segments) of the FIX weak labeling method.

3 The Label Accuracy and Cost of ORC Weak Labeling

Let us start with the ORC weak labeling method. This method uses a priori information about the event start and end times and is therefore not available in practice, but should be seen as an upper bound on what can be achieved with weak labeling. The start and end times of the true presence and absence events are used to construct the query segments:

$$\mathbb{Q}_{\text{ORC}} = \{(a_0, b_0), (a_1, b_1), \ldots, (a_{B_{\text{ORC}}-1}, b_{B_{\text{ORC}}-1})\} = \{q_0, \ldots, q_{B_{\text{ORC}}-1}\}, \tag{4}$$

where (a_i, b_i) is the ith ground truth presence or absence event. The annotator indicates presence or absence for each of these segments, which by construction results in the ground truth annotations, illustrated in Figure 2. In the example, there are three target events (green), and four absence events, which means that $B_{\text{ORC}} = 7$. In general $B_{\text{ORC}} \in \{2M - 1, 2M + 1\}$, where M denotes the number of presence events. The number of absence events can be fewer than $2M + 1$ if the recording starts or ends with a presence event, however, for simplicity and without losing generality, we will consider $B_{\text{ORC}} = 2M + 1$ as the minimum number of query segments needed for ORC to derive the ground truth. From an annotation cost perspective, this is the most cautious choice, and it is also the most likely outcome. The query accuracy is 1 for each query segment since by construction the fraction of correctly labeled data in each query segment will be 1 when given the correct presence or absence labels.

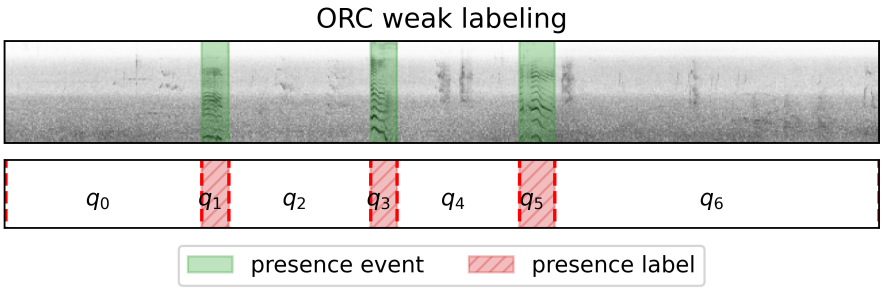


Figure 2: ORC weak labeling of an audio recording with three target events ($M = 3$) shown in green and four absence events. The $B_{\text{ORC}} = 7$, query segments $q_0, \ldots, q_6$ are derived from the ground truth segmentation of the data, and therefore the label accuracy will by definition be 1.

In summary, the ORC weak labeling method produces annotations with label accuracy 1, using the minimum number of query segments needed to achieve this. We use this as a reference on what can be achieved for weak labeling data.

4 The Label Accuracy and Cost of FIX Weak Labeling

The outline of this section is as follows. In Section 4.1 we define the FIX labeling method. In Section 4.2 we derive a closed-form expression for the expected label accuracy of a query segment given that it overlaps with a single event of deterministic event length. We note that it is only in the cases of overlap between a query segment and an event that a presence label can occur under the assumed annotator model, and that the expectation in label accuracy over these cases therefore can be viewed as the expected presence label accuracy. For the remainder of the paper we will simply write expected label accuracy when referring to the expectation over the overlapping cases, unless explicitly stated otherwise.

In the same section we derive the optimal query length with respect to the expected label accuracy, the maximum expected label accuracy and the number of query segments needed (proxy for annotation cost). In Section 4.3 we explain how the expression for expected label accuracy can be used in the case of a single event of stochastic length, and in Section 4.4 we explain under which conditions this can be used when multiple events can occur. Finally, we derive a closed form expression for the expected label accuracy of an audio recording with multiple events of stochastic length in Section 4.5, and provide an alternative interpretation of the theory in Section 4.6.

4.1 The FIX Weak Labeling Method

The FIX weak labeling method, commonly used in practice, splits the audio recording into fixed and equal length query segments, and then an annotator is asked to provide either a presence or absence label for each of the query segments. Let B_{FIX} denote the number of query segments used, then the query segments for an audio recording of length T are defined as

$$\mathbb{Q}_{\text{FIX}} = \{(a_0, b_0), (a_1, b_1), \ldots, (a_{B_{\text{FIX}}-1}, b_{B_{\text{FIX}}-1})\} = \{q_0, \ldots, q_{B_{\text{FIX}}-1}\}, \tag{5}$$

where the start and end timings of each query segment is $q_i = (a_i, b_i) = (id_q, (i+1)d_q)$ and the fixed query segment length is $d_q = T/B_{\text{FIX}}$. We illustrate this in Figure 3, where the presence criterion for the annotator is $\gamma = 0.5$. There are three presence events and four absence events, and using only $B_{\text{FIX}} = 7$ query segments results in annotations with an average label accuracy that is lower than 1.

We want to find an expression for the expected label accuracy for a given data distribution and query segment length. In addition, we want to understand the query length that maximize the expected label accuracy.

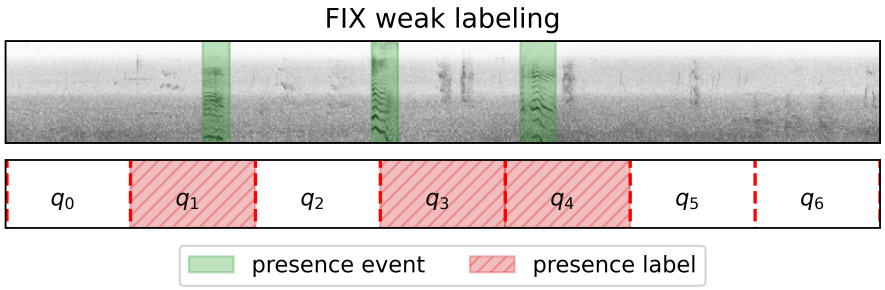


Figure 3: Illustration of the FIX weak labeling method. The audio recording contains presence events (green). The FIX method divides the recording into fixed-length query segments (e.g., q_0 to q_6). Note how the alignment between segments and presence events affects the accuracy of presence labels (red hatched).

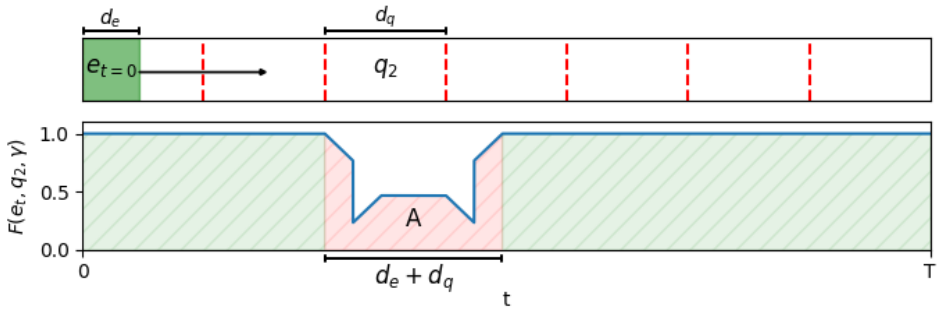


Figure 4: *Top panel:* A single event (e_t) of length d_e can occur at various end times (t) within the recording of length T. *Bottom panel:* The resulting label accuracy for query segment q_2 (arbitrarily chosen for illustration) of length d_q as a function of the event's end time (t). Overlap between the event and the query segment leads to segment label noise and a reduced label accuracy, which in this case occur when $t \in [a_2, a_2+d_e+d_q]$ where a_2 is the start time of q_2. The red hatched area (A) represents the cumulative label accuracy during these overlapping scenarios. The figure illustrates the evolution of the accuracy function $F(e_t, q, \gamma)$ as the event end time sweeps across the segment. The apparent central alignment at 0.5 is specific to this example and should not be interpreted as a general property of FIX labeling.

4.2 The Expected Label Accuracy of a Query Segment given Event Overlap

To derive a tractable closed-form expression, we analyze a simplified data distribution where each recording of length T contains a single event of deterministic length d_e. This idealized case is the simplest possible annotation scenario, and the resulting accuracy can therefore be interpreted as a theoretical upper bound: any added complexity, such as multiple events or variable event lengths, introduces additional opportunities for error.

The setup is illustrated in the upper panel of Figure 4, where a single event e_t of length d_e can occur at any time $t \in [0, T]$ (indicated by the arrow). The bottom panel of Figure 4 shows the label accuracy for a specific query segment (q_2) as the end time (t) of the event varies. The area (A) highlighted in hatched red indicates the label accuracy in the cases of overlap between the query segment and the event, and the area in hatched green indicate the label accuracy in the cases of no overlap, which is by the definition of the annotator model is always 1. Crucially, while this figure illustrates the accuracy for query segment q_2, the shape of this accuracy function remains the same for other query segments; only its position along the x-axis would change.

To simplify the mathematical analysis, without loss of generality, we can fix the query segment to start at time 0, $q = (0, d_q)$, and represent the event with its ending time t as $e_t = (t - d_e, t)$. In this way, $t \in [0, d_e + d_q]$ describes all possible overlap occurrences. That is, when $t = 0$ the event ends at the start of the query segment, and when $t = d_e + d_q$ the event starts at the end of the query segment. To formalize this, we can express the expected label accuracy in case of overlap by integrating over all possible event end times (t) where overlap occurs:

$$\mathbb{E}_{\mathrm{t}\sim p}\left[F(e_{\mathrm{t}}, q, \gamma)\right] = \int_0^{d_e+d_q} F(e_t, q, \gamma)p(t)\mathrm{d}t \tag{6}$$

$$= \frac{1}{d_e + d_q}\int_0^{d_e+d_q} F(e_t, q, \gamma)\mathrm{d}t \tag{7}$$

$$= \frac{A}{d_e + d_q}, \tag{8}$$

where $\mathrm{t} \sim p$ denotes a random variable t distributed according to a distribution p, and $p(t)$ denotes the probability of realization t. We assume that the distribution of the relative offsets t between events and overlapping query segments is uniformly distributed (empirically verified in Appendix A.4). There are two sources of variation that makes this plausible: (i) the start time of the recording varies depending on when the recording session was started, and (ii) the start time of the event varies depending on when the sound source emits the event. Note that this assumption is likely to not hold if $d_q \gg d_e$, but that leads to very weak labels which is not wanted in practice. Using this assumption we get $p(\mathrm{t}) = 1/(d_e + d_q)$, and by observing that the integral $\int_0^{d_e+d_q} F(e_t, q, \gamma)\mathrm{d}t$ describes the hatched red area denoted A in Figure 4 we arrive at the final expression in Eq. 8.

Remember that absence labels can occur when there is no overlap (always correct) and when there is overlap but the presence criterion is not fulfilled, and presence labels can only occur when there is overlap and the presence criterion is fulfilled. Therefore, inaccurate labels only occur in the case of overlap. The expected label accuracy in the case of overlap therefore describes the accuracy of the labels when segment label noise can occur, which happens around the boundaries of the true event.

In Appendix A.1 we show how to express A in terms of the event length d_e, the query segment length d_q and the presence criterion γ under the assumption that the annotator presence criterion can be fulfilled ($d_q \geq \gamma d_e$), and that it can not be fulfilled ($d_q < \gamma d_e$). Finally, we arrive at the following four main theorems:

Theorem 1. The expected label accuracy in case of overlap between a query segment q of length d_q and a single event e of deterministic length d_e is

$$f(d_q) = \mathbb{E}_{\mathrm{t}\sim p}\left[F(e_{\mathrm{t}}, q, \gamma)\right] = \begin{cases} \frac{d_e\left(2\gamma d_q - 2\gamma^2 d_e + d_q\right)}{d_q(d_e+d_q)}, & \text{if } d_q \geq \gamma d_e, \\ \frac{d_q}{d_e+d_q}, & \text{if } d_q < \gamma d_e, \end{cases} \tag{9}$$

when the presence criterion for the annotator is γ.

Proof. See Appendix A.1 for the proof. We show how to express the area A in Eq. 8 in terms of d_e, d_q and γ for the two assumptions: $d_q \geq \gamma d_e$, and $d_q < \gamma d_e$. □

Theorem 2. The query length that maximizes the expected label accuracy in case of overlap for a given event length d_e is

$$d_q^* = d_e\gamma\frac{2\gamma + \sqrt{4\gamma^2 + 4\gamma + 2}}{2\gamma + 1}. \tag{10}$$

Proof. See Appendix A.2 for the proof. We compute the derivative of $f(d_q)$ with respect to d_q, and show that d_q^* is the maximum. □

Theorem 3. The maximum expected label accuracy in case of overlap between a query segment of length d_q and an event of length d_e when $d_q \geq \gamma d_e$ is

$$f^*(\gamma) = f(d_q^*) = 2\gamma\left(2\gamma + 1 - \sqrt{4\gamma^2 + 4\gamma + 2}\right) + 1. \tag{11}$$

Proof. See Appendix A.3 for the proof. We substitute d_q for d_q^* in Eq. 9. □

Theorem 4. The number of queries B^*_{FIX} (cost) that are needed by FIX to maximize the expected label accuracy in case of overlap for an audio recording of length T when $d_e = 1$ is

$$B^*_{\text{FIX}} = \frac{T}{d_q^*}. \tag{12}$$

Proof. $T/B^*_{\text{FIX}} = d_q^*$, which by Theorem 2 leads to maximum label accuracy. □

In summary, Theorem 1 gives us an expression $f(d_q)$ for the expected label accuracy when query segments of length d_q are used to detect events of length d_e and the presence criterion for the annotator is γ. We use this to find the query segment length d_q^* that maximize the expected label accuracy, leading to Theorem 2. Theorem 2 show the query segment length d_q^* that maximizes expected label accuracy for a given event length and annotator criterion. Further, by inserting d_q^* into Theorem 1, $f^*(\gamma) = f(d_q^*)$, we get Theorem 3, which is the maximum achievable expected label accuracy for a given annotator criterion γ. We have omitted the case $d_q < \gamma d_e$ when deriving $f^*(\gamma)$, since maximizing the expected label accuracy in the case when the annotator presence criterion can not be fulfilled is not very interesting, since we can not get presence labels. Note that $f^*(\gamma)$ is a function of only γ, meaning that the maximum expected label accuracy is independent of the target event length when considering a single deterministic event. Finally, Theorem 4 show that an annotator needs to weakly label B^*_{FIX} query segments for each audio recording to achieve the maximum label accuracy in expectation, which can be seen as a proxy for annotation cost.

There is arguably no simpler audio data distribution to annotate than when recordings only contain a single event of deterministic length (except for when no event occurs at all). We can therefore treat $f^*(\gamma)$ as an upper bound on the maximum expected label accuracy for any audio distribution. We demonstrate this empirically in the results in Section 6. However, in practice audio recordings often contain events that vary both in length and number. Let us therefore consider how the derived theory can be useful also in these cases.

4.3 Stochastic Event Length

Events may vary in length according to some event length distribution. Let $p(d_e)$ denote the probability of the outcome that an event has length d_e, and let $d_e \sim p(d_e)$ denote that d_e is a sample from that distribution. The expected label accuracy over a distribution of event lengths for a given γ and query segment length d_q can then be computed as

$$\mathbb{E}_{d_e \sim p(d_e)}\left[f(d_q)\right] = \int_0^\infty f(d_q) p(d_e) \mathrm{d}d_e. \tag{13}$$

While we do not provide a closed form solution for this, we can solve the integral in Eq. 13 by numerical integration. Note that d_q^* in Theorem 2 depends on the single event length d_e, and to find it for a distribution we would need to solve Eq. 13 for a range of d_q and find the one that leads to the best label accuracy. However, for some event length distributions, setting d_e to the average of the distribution turns out to be a good heuristic. We perform a simulation study in Section 6.1.2 to support these claims.

4.4 Multiple Events

There may be multiple (M) events present in a given audio recording. In Figure 5 we show the label accuracy for all possible occurrences of a query segment q_t in a recording with two events ($M = 2$). Note that we have put the subscript t on the query segment (q_t) instead of the event as in the prior analysis. This formulation is entirely equivalent, but when talking about multiple events it is more intuitive to consider them as fixed in time for a given recording, and that the query segments occur relative them at random. There are now two regions where overlap occurs, one around e_1 and one around e_2. On average we get $2A/2(d_e + d_q) = A/(d_e + d_q) = f(d_q)$. That is, the theory we derived for the single event case explains the multiple event case.

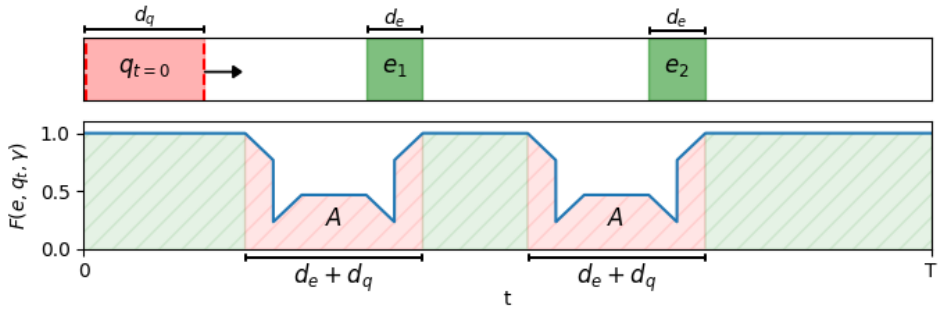


Figure 5: *Top panel:* Two events ($M = 2$) of length d_e that are fixed in time within a recording of length T, and a query segment $q_t = (-d_q + t, t)$. *Bottom panel:* The resulting label accuracy of q_t for $t \in [0, T - d_q]$, simulating that the q_t can appear anywhere at random in time in relation to the events. As before, when there is overlap between the query segment and an event the label accuracy is below 1, otherwise it is always 1.

However, for this to hold we need to assume that for any event the closest other event is least d_q away in time. In Figure 5 this holds since the start of e_2 is at least d_q away from the end of e_1. If this assumption holds then the expected label accuracy for multiple events is $f(d_q)$. The assumption is plausible if events are sparse in relation to d_q. Note that $d_q^* \in (0, d_e \frac{2+\sqrt{10}}{3}]$ for $\gamma \in (0, 1]$ according to Theorem 2. That is, when considering the optimal query length d_q^* this assumption translates to that events should be no closer than approximately $1.72 d_e$ for $\gamma = 1$, $0.81 d_e$ for $\gamma = 0.5$, and 0 for $\gamma \to 0$. We perform a simulation study in Section 6.1.3 to see the effect of breaking this assumption, and we leave it to future work to derive the expected label accuracy in case of overlap for multiple events.

4.5 The Expected Label Accuracy of an Audio Recording

We now know the expected label accuracy of a query segment given event overlap, and how to use this for a stochastic event lengths and multiple events. We can use this to derive an expression for the expected label accuracy of and audio recording of finite length (T) that has multiple (M) stochastic event lengths ($d_e \sim p(d_e)$).

Theorem 5. The expected label accuracy for an audio recording of length T, with M events of stochastic event length $d_e \sim p(d_e)$ that are spaced at least d_q apart is

$$\mathbb{E}_{d_e \sim p(d_e)} \left[-\frac{2M d_e^2 \gamma^2}{T d_q} + \frac{2M d_e \gamma}{T} - \frac{M d_q}{T} + 1 \right]. \tag{14}$$

Proof. We will do this proof by picture. In Figure 5 we have two events ($M = 2$), in general for M events the accumulated label accuracy in the cases of overlap is MA (the sum of the hatched red areas), the total amount of overlapping cases is $M(d_e + d_q)$ and the total amount of non-overlapping cases is therefore $T - M(d_e + d_q)$ for an audio recording of length T. In the case of no overlap, the label accuracy is always 1, which means that the accumulated label accuracy in the case of no overlap (sum of the green hatched areas) is $T - M(d_e + d_q)$. Normalizing for the entire duration of the recording we arrive at

$$\frac{AM + T - M(d_e + d_q)}{T} = -\frac{2M d_e^2 \gamma^2}{T d_q} + \frac{2M d_e \gamma}{T} - \frac{M d_q}{T} + 1, \tag{15}$$

and as before we can simply compute an expectation over the event length distribution. □

Theorem 5 tells us the expected label accuracy under FIX weak labeling with query segment length d_q for an audio recording of length T, with M events of stochastic event length $d_e \sim p(d_e)$. If we want to account

for class label noise, where the annotator gives the wrong label with probability ρ, this can be included by simply scaling the whole expression in Eq. 14 by $(1-\rho)$. That is, the expected label accuracy for the cases of overlap allows us to express a variety of things about the expected label accuracy of an audio recording.

However, note that we have T in the denominator of all terms except the term that is 1, meaning that if we let T approach ∞, then the expected label accuracy approaches 1. That is, considering the accuracy of both absence and presence labels equally can lead to hiding the effect that we want to understand in this paper, which is the effect of d_q on the accuracy of the presence labels. We could derive a balanced accuracy in a similar way as above, but instead we choose to continue our analysis looking only at the expected label accuracy in the case of overlap.

4.6 Expected Label Accuracy given Overlap when $d_q = \delta d_e$

As a result of the proof for Theorem 3 in Appendix A.3 we get an alternative dimensionless interpretation of the expected label accuracy when the query segment length is expressed as a factor of the event $d_q = \delta d_e$,

$$f(\delta d_e) = \frac{(2\gamma+1)\delta - 2\gamma^2}{\gamma(1+\gamma)}, \tag{16}$$

and an expression for the ratio that maximizes it

$$\delta^* = \frac{d_q^*}{d_e} = \gamma \frac{2\gamma + \sqrt{2\gamma^2 + 2\gamma + 1}}{2\gamma + 1}. \tag{17}$$

This alternative formulation illustrates that it is the ratio $\delta = d_q/d_e$ that affects the expected label accuracy of a single event, and not the absolute lengths d_q and d_e. Further, we can use this interpretation to rewrite Theorem 5 as

$$\mathbb{E}_{\delta \sim p(\delta)} \left[\frac{Md\delta\,(-\delta + 2\gamma) - 2Md\gamma^2 + T\delta}{T\delta} \right], \tag{18}$$

where δ denotes a random variable with probability distribution $p(\delta)$.

5 Simulating the Label Accuracy of FIX Weak Labeling

To validate the theory, we simulated FIX labeling of various audio recording distributions and compared the average simulated label quality with the theoretical results from Section 4.2. The code used for these simulations is released openly[1].

We generated 1000 audio recordings of length $T = 100$ seconds for each configuration. The number of events, M, and the event length distributions varied across simulations, as detailed below:

- **Single Event with Deterministic Length:** We simulated recordings with $M = 1$ event of deterministic length $d_e = 1$ second.
- **Single Event with Stochastic Length from Normal Distributions:** We drew event lengths from two normal distributions with the same mean but different variances ($\mathcal{N}(3, 0.1)$ and $\mathcal{N}(3, 1)$), and from two normal distributions with different means but the same variance ($\mathcal{N}(0.5, 0.1)$ and $\mathcal{N}(5, 0.1)$). For these simulations, $M = 1$.
- **Single Event with Stochastic Length from Gamma Distributions:** We sample event lengths from two gamma distributions (offset by 0.5 seconds due to computation cost) with different shape parameters but the same scale parameter (Gamma$(0.8, 1) + 0.5$ and Gamma$(0.2, 1) + 0.5$) with $M = 1$.
- **Single Event with Stochastic Length from Real Length Sample:** We used the event length distributions for dog barks and baby cries from the NIGENS dataset (Trowitzsch et al., 2019) with $M = 1$.

[1] `https://github.com/johnmartinsson/the-accuracy-cost-of-weakness`

- **Multiple Events with Deterministic Length:** We simulated recordings with multiple events ($M = 30$ and $M = 50$) where each event had a deterministic length of $d_e = 1$ second.

For recordings with stochastic event lengths or multiple events, the length of each of the M events was sampled from the specified distribution. Each sampled event was then placed randomly within the recording. The start time a_e of each event was drawn uniformly at random from $[0, T - d_e]$. If multiple events were present, overlapping events were merged into one presence event. For each generated audio recording, we simulated FIX labeling using different annotator presence criteria $\gamma \in [0.01, 0.99]$ and a range of query segment lengths d_q. The query segment lengths were linearly spaced between a small fraction of the minimum event length observed in the distribution and a value several times the maximum observed event length.

We then computed the average label accuracy over the query segments that overlaps with an event in each recording. For each query segment q we check if the annotator presence criterion ($h(e, q) \geq \gamma$) is fulfilled for any event $e \in E$, where E is the set of all events that overlap with q. If this is true for any of the events then q is given a presence label ($l_q = 1$) otherwise it is given an absence label ($l_q = 0$). The label accuracy is then computed in a similar way as in Eq. 3, but since we can now have multiple events overlapping with the same query segment, we need to consider the union of all overlapping events $\cup_{e \in E} e$ when computing the label accuracy of assigning label l_q to that query segment. The total amount of overlap becomes $|(\cup_{e \in E} e) \cap q|$ instead of $|e \cap q|$. However, when $M = 1$ this is equivalent to Eq. 3 ($|(\cup_{e \in E} e) \cap q| = |e \cap q|$), since $|E| = 1$.

In this way, we simulated the effect of breaking the assumption that events are spaced at least d_q apart, and could better understand the effect this had when compared to the derived theory. Finally, for each considered γ, we empirically determined the maximum average label accuracy across all tested query lengths and the corresponding optimal query length. These empirical results were then compared to the theoretical predictions.

6 Results

In this section we present the results of the simulated annotation process, and show how these connect to the derived theory. We start by looking at the expected label accuracy and the query segment length that maximize the expected label accuracy for FIX and ORC weak labeling, and then we relate this to the annotation cost.

6.1 Expected Label Accuracy given Overlap

We evaluate how different annotator presence criteria (γ) influence the achievable label accuracy given overlap under FIX weak labeling. We first examine the case of a single event with a deterministic length, then extend our simulation study to stochastic event lengths, and finally to multiple events occurring within the same recording.

6.1.1 Single Event with Deterministic Length

The simulated results are derived using the simulation setup described in section 5, with $M = 1$ (a single event) and $d_e = 1$ (deterministic length). In Figure 6, we show the maximum expected label accuracy given overlap (left) and the corresponding query length that maximize the label accuracy (right) for different γ. $f^*(\gamma)$ is the maximum expected label accuracy achievable with annotator presence criterion γ for the considered event length. We can see that the simulated average label accuracy closely follows the expected label accuracy, and that the corresponding segment length leading to this maximum is the same in theory and simulation.

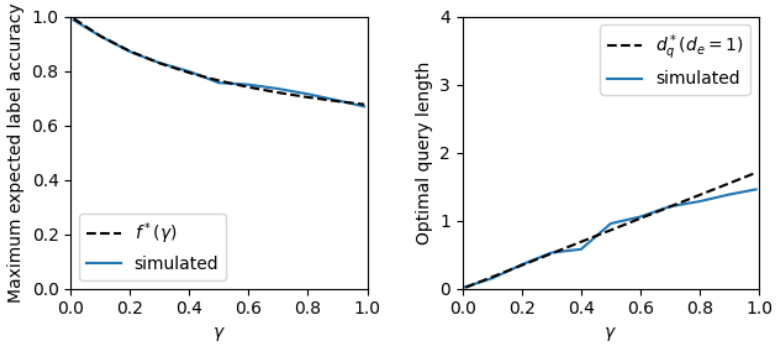


Figure 6: In the left panel we show the maximum expected label accuracy, $f^*(\gamma)$, for different γ, and the average maximum label accuracy from the simulations. In the right panel we show the query length that leads to this maximum label accuracy in theory, for $d_e = 1$, and in simulation. The theory follows the simulations well.

In Figure 6 we see that if the annotator needs to hear more than 50% of the sound event to detect presence ($\gamma = 0.5$) then the highest achievable label accuracy is $f^*(0.5) \approx 0.76$. This means that on average there is around 34% segment label noise around the presence labels. We also see that the query length that gives the maximum label accuracy is $d_q^* \approx 0.81$. The gap to the ORC weak labeling method which always gives a label accuracy of 1, is large especially for large γ. In general, we can see how the maximum label accuracy deteriorates with a growing γ, and which query segment length to choose to maximize label accuracy in expectation.

6.1.2 Single Event with Stochastic Length

We now consider stochastic event lengths. We do this to better understand the effect of the event length distribution on the maximum expected label accuracy and the optimal query length. We solve the integral in Eq. 13 by numerical integration over different event length distributions, and compare with the theory derived for a single deterministic event length and simulations. In each figure we present the derived theoretical rules $f^*(\gamma)$ and d_q^* for the simplified event length distribution, the results from integration of Eq. 13 with different event length distributions $p(d_e)$ (numerical), and the simulated results using the procedure described in section 5 (simulated) where event lengths are sampled from different distributions. Note that, since d_q^* is derived for a deterministic event length d_e, and require a choice of this value, we set d_e to the average event length (μ) for each distribution in these experiments as a heuristic. We then present the maximum expected label accuracy for different γ (left in figures) and the query segment length that maximizes the expected label accuracy (middle in figures), and the histogram for the considered event length distributions (right in figures).

In Figure 7 and Figure 8 we see that the mean and variance of the normal distribution have a small (if any) effect on the maximum expected label accuracy, but the mean does affect which query segment length that maximizes the expected label accuracy. We also see that d_q^* follows the simulated and numerical optimal query length well for all considered normal distributions, when d_e is set to the average event length (μ) for the considered event length distribution. The average event length can be used as a heuristic value if we only know the average and not the true distribution to integrate over.

In Figure 9 we can see that a gamma distribution does affect the maximum expected label accuracy, and that simply setting d_e to the average event length of the distribution leads to underestimating the optimal query length. Since it is not possible to optimize for both short and long events at the same time using FIX weak labeling, this type of distribution is quite challenging.

In Figure 10 we validate the theory against a real sample of event lengths from either baby cries or dog barks. Numerical integration between the derived expression and the histogram predicts the simulations well.

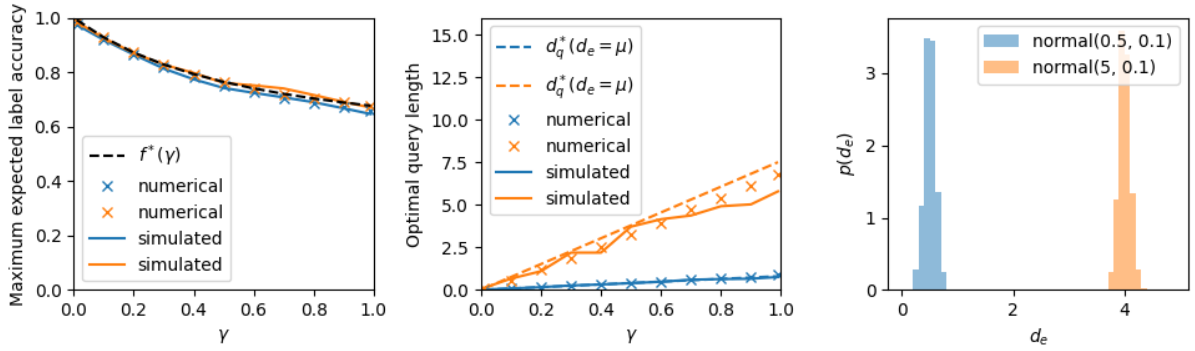


Figure 7: We validate the theory for stochastic event lengths drawn from two normal distributions with different means, but the same variance. We show the expected label accuracy (left panel), the optimal query length (middle panel), and the considered event length distributions (right panel).

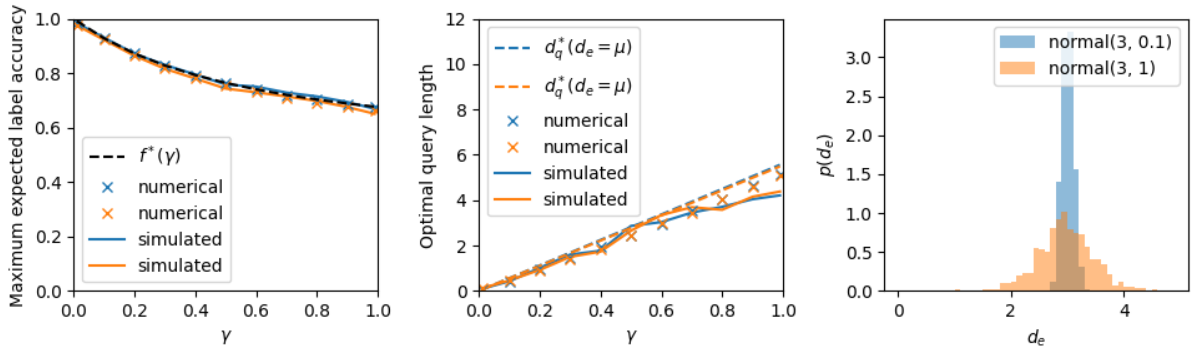


Figure 8: We validate the theory for stochastic event lengths drawn from two normal distributions with different variance, but the same mean. We show the expected label accuracy (left panel), the optimal query length (middle panel), and the considered event length distributions (right panel).

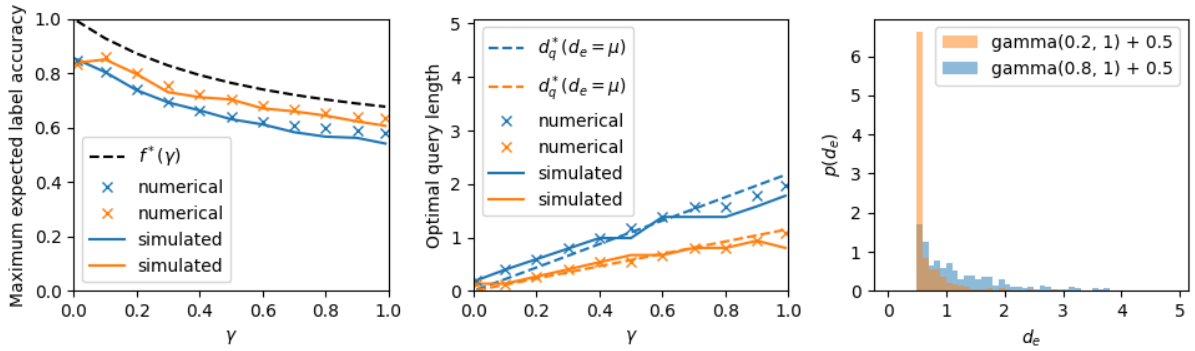


Figure 9: We validate the theory for stochastic event lengths drawn from two gamma distributions with different shape parameters, but the same scale parameter. We show the expected label accuracy (left panel), the optimal query length (middle panel), and the considered event length distributions (right panel).

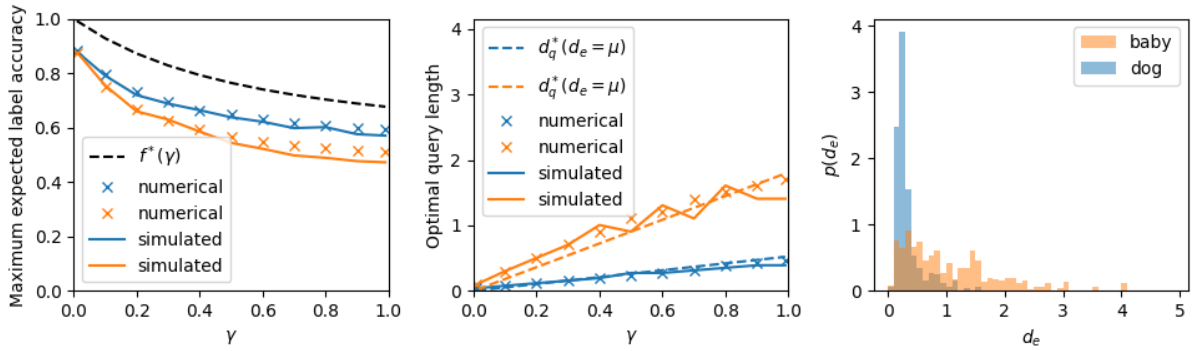


Figure 10: Barking dog and crying baby event length distributions from the NIGENS dataset (Trowitzsch et al., 2019). These annotations have been made with a strong guarantee for high quality onsets and offsets.

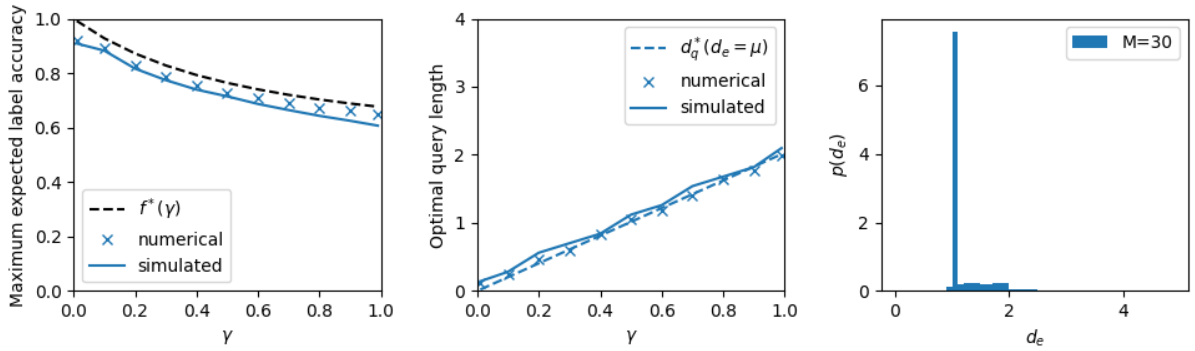


Figure 11: We validate the theory for multiple events of length $d_e = 1$. We show the expected label accuracy (left panel), the optimal query length (middle panel), and the considered event length distributions (right panel). Note that presence events longer than 1 can occur if two or more events overlap. We sample 30 events with event length $d_e = 1$ occur at random for each audio recording in this simulation.

6.1.3 Multiple Events with Stochastic Length

In these simulations we allow multiple events to occur in the same recording ($M > 1$). In Figure 11 we show the results of sampling 30 events of length $d_e = 1$ for each audio recording. This does have an effect on the expected maximum label accuracy and the corresponding query length, though the impact is relatively modest. In Figure 12 we show the results of sampling 50 events of length $d_e = 1$ for each audio recording. This is an extreme case, where the event density of the recording is very high. These results demonstrate that even under high event densities, the simulated maximum accuracy follows the theoretical predictions closely. This confirms that the single-event theory provides a robust upper bound even when assumptions about event sparsity are strongly violated.

6.2 Annotation Cost for Maximum Expected Label Accuracy given Overlap

Achieving maximum expected label accuracy comes at a cost, and understanding this cost trade-off is essential for practical annotation efforts. The cost model we employ accounts for both the time spent listening to audio and the effort required to label presence or absence events.

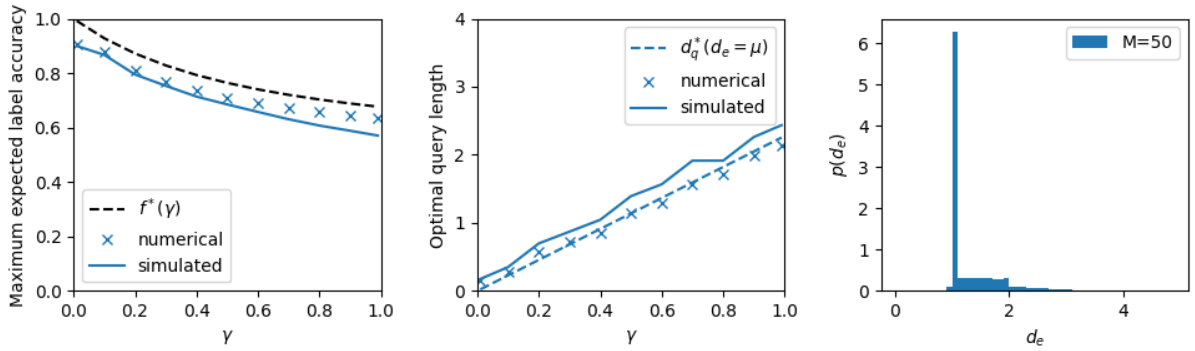


Figure 12: We validate the theory for multiple events of length $d_e = 1$. We show the expected label accuracy (left panel), the optimal query length (middle panel), and the considered event length distributions (right panel). Note that presence events longer than 1 can occur if two or more events overlap. We sample 50 events with event length $d_e = 1$ occur at random for each audio recording in this simulation.

6.2.1 Formalizing the Cost Model

The derived theory for the optimal query length allows us to analyze the cost of achieving maximum expected label accuracy under different annotator models for FIX weak labeling. We assume that the whole audio recording of length T is listened to. The key difference in cost between the FIX and ORC weak labeling method is the number of segments (B) that need to be given a presence or absence label. We formalize a cost model as:

$$C(T, B) = (1 - r)T + rB, \tag{19}$$

where $1 - r$ represents the cost of listening to one second of audio (cost per second), and r represents the cost of answering a query (cost per query). The term $(1 - r)T$ therefore represents the cost of listening to T seconds of audio, and the term rB the cost of assigning B presence or absence labels. Using this cost model, we calculate the cost of annotating an audio recording of length T with M sound events of length $d_e = 1$ using either FIX or ORC weak labeling. For FIX, the number of queries that maximize expected label accuracy is given by $B^*_{\text{FIX}} = T/d_q^*$ (see Theorem 4). For ORC, achieving an expected label accuracy of 1 requires at least $B^*_{\text{ORC}} = 2M + 1$ queries.

In practice, we do not know the number of events M. To explore potential overestimation of M when, for example, using a weak labeling process that tries to mimic ORC weak labeling, we model B_{ORC} as a multiple of the necessary number of queries: $B_{\text{ORC}} = sB^*_{\text{ORC}}$, where $s \in \{1, 2, 4, 8\}$ represents the degree of overestimation. This approach captures scenarios where the number of events are either precisely estimated ($s = 1$) or significantly overestimated ($s = 8$) during the annotation process. In practice, B_{ORC} could be set based on a bound on M. For example, by estimating a maximum expected number of sound events in a recording, M_{max}, based on knowledge of typical event density, or characteristics of the audio recording. We assume that overestimation by more than a factor of 8 is unlikely. The relative cost between FIX and ORC weak labeling can then be computed as:

$$\frac{C_{\text{FIX}}}{C_{\text{ORC}}} = \frac{C(T, B^*_{\text{FIX}})}{C(T, B_{\text{ORC}})}, \tag{20}$$

where a ratio larger than 1 indicates that FIX is more costly than ORC, and a ratio smaller than 1 indicates that FIX is less costly than ORC.

6.2.2 Effect of annotator criteria (γ) and cost ratio (r).

Figure 13 (left) shows the relative cost for varying annotator criteria $\gamma \in [0.1, 1]$. As $\gamma \to 0.1$, the cost of FIX increases sharply, reflecting the need for an infinitely large number of queries to achieve an expected label accuracy of 1. In practice, achieving perfect accuracy with FIX is infeasible due to the associated cost.

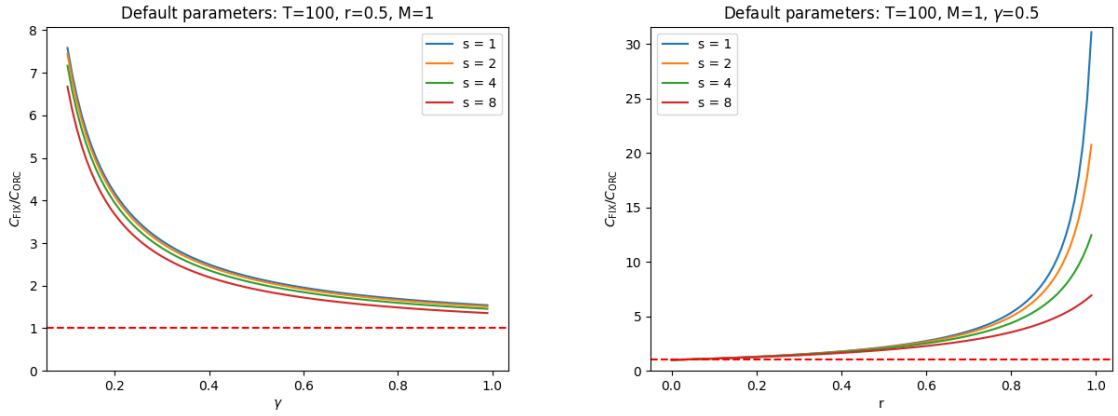


Figure 13: The relative cost of FIX and ORC for varying annotator criteria γ (left), and cost ratios r (right). The default parameters are: $T = 100$, $r = 0.5$, $M = 1$ and $\gamma = 0.5$. We simulate overestimating the number of needed queries $B_{\text{ORC}} = s(2M + 1)$ by a factor of s for $s \in \{1, 2, 4, 8\}$ to see how this affects the relative cost. The cost of FIX is greater than the cost of ORC above the dashed red line where the cost ratio is 1.

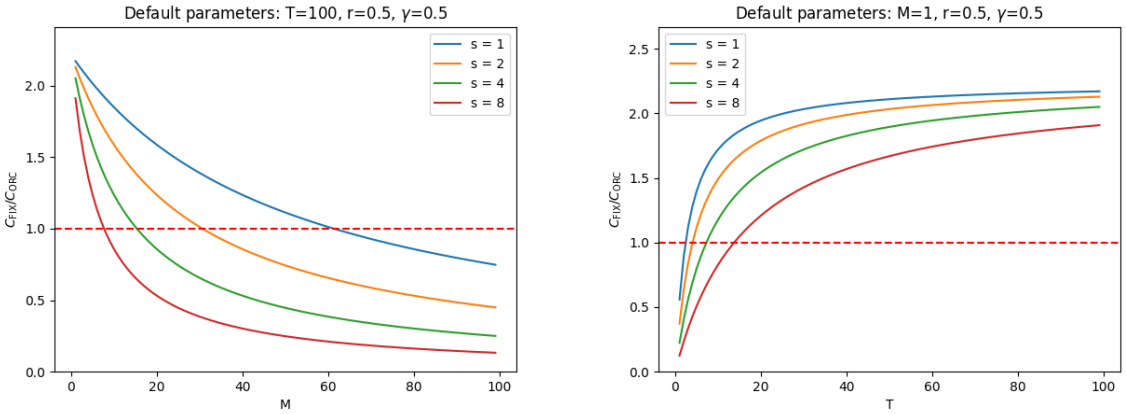


Figure 14: The relative cost of FIX and ORC for varying number of sound events M (left) and recording lengths T (right). The default parameters are: $T = 100$, $r = 0.5$, $M = 1$ and $\gamma = 0.5$. We simulate overestimating the number of needed queries $B_{\text{ORC}} = s(2M + 1)$ by a factor of s for $s \in \{1, 2, 4, 8\}$ to see how this affects the relative cost. The cost of FIX is greater than the cost of ORC above the dashed red line where the cost ratio is 1.

For higher γ, the cost of FIX becomes more comparable to ORC. However, combining this with Theorem 3 reveals that FIX can either match ORC in cost but with lower expected accuracy or achieve similar accuracy at a much higher cost.

The right panel of Figure 13 examines the impact of the cost ratio r. Across all tested values, ORC remains less costly than FIX in the default setting ($T = 100$, $r = 0.5$, $\gamma = 0.5$, $M = 1$). This confirms that the relative cost advantage of ORC is robust to changes in r.

6.2.3 Effect of number of events (M) and recording length (T).

Figure 14 explores the impact of M and T on the relative cost. In the left panel, we see that for $s = 1$, ORC is less costly than FIX when the number of events is below 60. However, as s increases to 8, FIX becomes

less costly when at most 10 events are present. These results indicate that the relative cost depends heavily on the density of sound events in the recording and the estimated annotation budget for ORC.

In the right panel, varying T shows a similar trend. For shorter recordings (high event density), ORC loses its cost advantage. However, it's important to note that the maximum achievable expected label accuracy with FIX under default settings ($\gamma = 0.5$) is $f^*(0.5) \approx 0.76$, whereas ORC achieves 1.0. In such cases, the additional cost of ORC may be justified by the significantly higher label quality.

While these results indicate that the relative cost depends on the sound event density, we should remember that we are considering weak labeling of presence events. This implies that all M events in this analysis are treated as non-overlapping, as the annotation task does not consider temporal overlaps for this analysis. The scenario of $M > 60$ non-overlapping events of length 1 in a recording of length $T = 100$ is therefore unlikely in practice. Similarly, estimating 10 events as 80 (modeled by $s = 8$) for an audio recording of length $T = 100$ represents a substantial overestimation and seems improbable given the capabilities of modern sound event detection tools.

7 Related Work

This work introduces a framework for characterizing segmentation label noise in FIX weak labeling, a largely unexplored area. Below, we review studies addressing noisy labels and approaches to mitigate their effects, with a focus on weak labeling in audio and related domains.

7.1 Understanding Noisy Labels

Noisy labels are a partial description of the target model, influencing its performance. Early work by Liang et al. (2009) introduced the concept of *measurements* for conditional exponential families, encompassing labels and constraints for model learning with minimal human input—a goal shared by this work.

In deep learning, the ability of models to overfit noisy labels has prompted studies into the relationship between noise rate and generalization (Zhang et al., 2021; Chen et al., 2019). Research on class label noise often assumes a noise transition matrix (Li et al., 2021) but rarely considers spatially correlated errors like those arising in segmentation tasks (Yao et al., 2023). For audio, Hershey et al. (2021) demonstrated that training on strongly labeled data yields better results than weakly labeled data, highlighting the need for precise labels, particularly in evaluation. In multi-modal tasks, such as audio-visual video parsing, a key challenge is *modality-specific label noise*, where a video-level tag may apply to the audio stream but not the visual, or vice versa (Cheng et al., 2022; Zhou et al., 2024). Our work focuses specifically on characterizing the *segment label noise* that arises from the temporal misalignment between fixed-length query segments and true event boundaries.

7.2 Mitigating Noisy Labels

Several strategies address noisy labels, including regularization techniques like dropout (Srivastava et al., 2014), data augmentation (Shorten & Khoshgoftaar, 2019), and specialized loss functions (Fonseca et al., 2019). For weakly labeled audio, Dinkel et al. (2022) proposed a pseudo-labeling approach, iteratively refining labels to improve training performance. Despite these advances, most methods focus on training labels and offer limited insights into noisy evaluation labels, underscoring the need for frameworks that quantify label noise, such as the one proposed in this work.

7.3 Strong vs. Weak Labeling

Strong labeling, where the annotator provides the event boundaries and the class label, while often precise, is resource-intensive and subject to annotator variability (Mesaros et al., 2017). In bioacoustics, experts use spectrograms for efficient annotation (Cartwright et al., 2017), but the reliance on specialists limits scalability. Weak labeling, by contrast, simplifies the annotation task, which is especially important for crowd-sourced annotations, enabling broader data collection (Martin-Morato & Mesaros, 2023). However, segment label noise, especially at event boundaries, remains a significant challenge.

Dataset	Task	Fixed Length
CHIME (Foster et al., 2015a)	Single-pass multi-label	4 seconds
AudioSet (Gemmeke et al., 2017b)	Single-pass multi-label	10 seconds
MAESTRO Real (Martin-Morato & Mesaros, 2023)	Single-pass multi-label	10 seconds
OpenMIC-2018 (Humphrey et al., 2018b)	Multi-pass binary-label	10 seconds

Table 1: Large-scale audio datasets using variations of FIX weak labeling.

Large-scale audio datasets employing FIX weak labeling are summarized in Table 1. Two common annotation tasks are single-pass multi-label and multi-pass binary-label annotation (Cartwright et al., 2019). Single-pass multi-label annotation asks annotators to recognize the presence of multiple event classes during a single pass through the data. In contrast, multi-pass binary-label annotation asks annotators to detect the presence or absence of a single event class at a time through multiple passes through the data.

Cartwright et al. (2019) studied the trade-offs between these tasks and found that binary labeling is preferable when high recall is required. For example, AudioSet (Gemmeke et al., 2017b) employs single-pass multi-label annotation with non-overlapping 10-second segments, which limits temporal resolution. Conversely, MAESTRO Real (Martin-Morato & Mesaros, 2023) uses overlapping 10-second segments with a 9-second overlap, increasing the accuracy of the derived labels.

The choice of segment length and overlap significantly impacts the utility of weak labeling. For example, while overlapping segments increase label accuracy (Martin-Morato & Mesaros, 2023), they still fail to distinguish events occurring close in time. Current work aims to better understand the effect of different choices of the segment length for FIX weak labeling.

7.4 Contributions of This Work

Existing research focuses predominantly on class label noise or assumes noise independence. This work extends these efforts by characterizing segment label noise specific to FIX weak labeling, providing a foundation for improving both training and evaluation processes in weakly labeled datasets.

8 Discussion

FIX labeling has been employed in many works, with varying degrees of complexity. Theorem 2 provides a useful rule of thumb for selecting the best segmentation length for a given event length, and Eq. 13 provides a way to use this theorem to analyze stochastic event length distributions. Our results suggest that, in most cases, knowing the average event length provides a good estimate, but understanding the (approximate) distribution of event lengths improves the analysis.

Implications for Practical Annotation

Our theoretical analysis positions FIX weak labeling as a baseline strategy: it is simple and scalable but inherently limited by segment label noise. By quantifying this limitation, our work provides the necessary theoretical justification for moving towards adaptive methods that aim to approximate the oracle process.

The analysis highlights the trade-offs in label accuracy and annotation cost between FIX and ORC weak labeling. While FIX can be less costly under specific conditions (e.g., high event density), these conditions are unlikely to occur in real-world annotation tasks. Furthermore, even in cases where FIX is less costly, its significantly lower label accuracy ($f^*(0.5) \approx 0.76$ vs. 1.0 for ORC) can negate its cost advantage. This gap represents the "accuracy cost of weakness" that is inherent to any non-adaptive weak labeling strategy. This gap only increases when d_q is chosen sub-optimally, which is often the case in practice due to budget considerations. Given the rarity of extreme event densities and the importance of high-quality labels, ORC is likely the better theoretical choice for most annotation tasks.

However, ORC weak labeling is not available in practice since it uses the true event boundaries. This provides a clear theoretical justification for developing adaptive weak labeling methods, which aim to approximate

the ORC process. For instance, methods that use active learning or change-point detection to define query boundaries (Martinsson et al., 2024; Kim et al., 2023) are practical attempts to bridge the gap between FIX and ORC. Martinsson et al. (2024) empirically evaluates an adaptive change-point detection method (A-CPD) and compares that to FIX weak labeling and ORC weak labeling, showing the benefit of an adaptive weak labeling method for annotation of sound events. Our work provides the tools to quantify the maximum potential accuracy gain for such methods over a simple FIX baseline, offering a principled way to evaluate the trade-off between the complexity of an adaptive strategy and its achievable accuracy. Future research should focus on mitigating the potential biases when modeling ORC weak labeling (e.g., annotation errors, overfitting to sparse events) while retaining its theoretical advantages.

Implications for Model Evaluation

Despite the extensive focus on noisy training labels, evaluation labels are often implicitly assumed to be perfect. As emphasized in the introduction, inaccurate evaluation labels present a significant challenge. When noise is present in both training and evaluation data, we risk selecting models that merely replicate the evaluation noise, potentially overlooking those with superior generalization abilities. This echoes the central concern highlighted by Görnitz et al. (2014). We can use Theorem 3 to understand the properties of the best performing model when the evaluation data contains FIX weak labels. For example, for $\gamma = 0.5$ the annotations will at most have an expected label accuracy of $f^*(0.5) \approx 0.76$. The "best" performing model will therefore be a model that mimics this specific noise profile. Our theory thus provides a better understanding of the target that models are optimizing for when evaluated on weakly labeled data.

This is also relevant for standard sound event detection (SED) evaluation metrics, such as the segment-based F_1 score (Mesaros et al., 2016), which divide audio into fixed-length segments. When using ground truth labels for evaluation, we effectively have an annotator with $\gamma \to 0$. The expected label accuracy then becomes $f(d_q) = d_e/(d_e + d_q)$, where d_q is the segment length. This formula shows that a small d_q minimizes segment label noise, but choosing a very small segment length negates the desired effect of mitigating temporal imprecision in the ground truth and also increases computational cost. The theory presented here can help inform such trade-offs.

Theoretical Properties and Validation

The expression for expected label accuracy derived in this paper applies to the simplest scenario, where only a single event with deterministic length is present. In all of our results, we observe that $f^*(\gamma)$ is greater than or equal to the expected and average label accuracy that FIX weak labeling achieve for more complex distributions. This suggests that $f^*(\gamma)$ can be considered an upper bound for a given annotation process. However, a formal proof showing that adding more events or introducing event length variability leads to a harder distribution to annotate is beyond the scope of this paper.

To connect this theoretical framework to a real-world setting, we conducted an empirical analysis using the weakly and strongly labeled versions of the AudioSet dataset, as detailed in Appendix A.5. By treating the 10-second weak labels of AudioSet as the output of a FIX process, we calculated the empirical label accuracy against the corresponding strong labels. Our theoretical model, when applied to the event length distribution of the "Animal" class, accurately predicted this empirical accuracy for a presence criterion of $\gamma \approx 0.26$. This serves as an empirical validation of our framework on a large-scale dataset.

Generalization and Future Directions

While our work is grounded in audio event detection, the core principles are broadly generalizable because the mathematics depends only on two fundamental quantities: the event duration d_e, and the query segment length d_q. These quantities can be directly mapped to other domains, e.g., video action spotting, electrocardiography, seismology, or high-frequency trading. Consequently, our core results (Theorems 1-3) transfer directly to these domains, provided three conditions hold: (i) the events uniformly distributed locally in time, making the uniform relative offset a reasonable model (as empirically verified in Appendix A.4), (ii) the annotation process relies on observing a minimum fraction γ of an event, and (iii) the events are suf-

ficiently sparse. This framework can also be extended to higher dimensions, such as analyzing the weak labeling of rectangles in images or cubes in point clouds.

Finally, if the same presence criterion γ is applicable for all event classes, Theorem 1 applies to the joint event length distribution. However, real-world presence criteria for different event classes may vary, requiring more complex models. Future empirical studies on annotator behavior could help refine this model and improve its practical applicability.

9 Conclusions

This study introduces a novel theoretical framework for understanding the trade-offs between label accuracy and annotation cost in weak labeling methods, particularly focusing on sound event detection where weak labeling is often employed to reduce annotation costs. We specifically compared fixed-length (FIX) and oracle (ORC) approaches.

We have demonstrated that FIX weak labeling, while cost-effective in specific scenarios, is inherently limited by segment label noise. The expressions we derived theoretically provide actionable insights into optimizing segment length for maximizing expected label accuracy under FIX. However, these results also underscore the fundamental trade-offs: shorter segments improve alignment with event boundaries but significantly increase annotation cost, while longer segments reduce cost at the expense of accuracy. In addition, how short these segments can be chosen depends on the ability of the annotator to detect presence of fractions of the events. In contrast, ORC labeling achieves perfect accuracy but can incur higher costs if events are very dense and the number of events are overestimated.

Our findings have several practical implications:

- **Annotation Strategy:** FIX weak labeling remains a robust, scalable choice for many practical applications. However, when high label accuracy is essential, ORC weak labeling—or adaptive methods approximating it—should be prioritized.
- **Adaptive Techniques:** Theoretical justification for adaptive weak labeling methods, e.g., methods based on active learning or iterative refinement, that mimic ORC weak labeling, which suggests promising avenues for improving annotation efficiency without compromising accuracy.
- **Evaluation Criteria:** Our analysis highlights the potential biases introduced by segment-level label noise in evaluating sound event detection models. Therefore, carefully aligning evaluation criteria with the intended model properties is critical.

Future research should address several limitations and extensions identified in our study. Developing practical approaches that reliably mimic ORC weak labeling by estimating the query segments without introducing a lot of unwanted bias in the labels remains an open challenge. Additionally, extending this framework to multi-dimensional data and multiple presence classes could broaden its applicability to other domains, such as medical imaging and point clouds.

In conclusion, the insights presented in this work offer a foundation for optimizing weak labeling processes, balancing cost and accuracy to meet the needs of diverse machine learning applications. By refining annotation strategies and leveraging adaptive methods, researchers can enhance the quality of labeled datasets. This, in turn, will drive advancements in supervised learning across domains, building upon the foundational understanding presented in this work.

Acknowledgments

The authors would like to express their gratitude to the following individuals and organizations for their contributions and support:

- **Annamaria Mesaros**: Provided insightful questions and feedback on an early draft of this work.

- **Magnus Oskarsson**: Contributed a proof sketch for Theorem 2 and gave feedback on an early draft.
- **Edvin Listo Zec**: Assisted with the proof of Theorem 2.
- This work was supported by the Swedish Foundation for Strategic Research (SSF; FID20-0028) and Sweden's Innovation Agency (2023-01486).

References

Juan P. Bello, Claudio Silva, Oded Nov, R. Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: a system for monitoring, analyzing, and mitigating urban noise pollution. *Commun. ACM*, 62(2):68–77, January 2019. ISSN 0001-0782. doi: 10.1145/3224204. URL `https://doi.org/10.1145/3224204`.

Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):–246, 2017. ISSN 25730142. doi: 10.1145/3134664.

Mark Cartwright, Graham Dove, Ana Elisa Méndez Méndez, Juan P. Bello, and Oded Nov. Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists. *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–11, 2019. doi: 10.1145/3290605.3300522.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1062–1070. PMLR, May 2019. URL `https://proceedings.mlr.press/v97/chen19g.html`. ISSN: 2640-3498.

Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pp. 431–448, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19829-8. doi: 10.1007/978-3-031-19830-4_25. URL `https://doi.org/10.1007/978-3-031-19830-4_25`.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, and Yujun Wang. Pseudo Strong Labels for Large Scale Weakly Supervised Audio Tagging. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2022-May:336–340, 2022. ISSN 15206149. doi: 10.1109/ICASSP43922.2022.9746431. arXiv: 2204.13430 Publisher: IEEE ISBN: 9781665405409.

Eduardo Fonseca, Frederic Font, and Xavier Serra. Model-Agnostic Approaches To Handling Noisy Labels When Training Sound Event Classifiers. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 16–20, October 2019. doi: 10.1109/WASPAA.2019.8937249. URL `https://ieeexplore.ieee.org/document/8937249`. ISSN: 1947-1629.

Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D. Plumbley. Chime-home: A dataset for sound source recognition in a domestic environment. *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*, pp. 1–5, 2015a. doi: 10.1109/WASPAA.2015.7336899.

Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D. Plumbley. Chime-home: A dataset for sound source recognition in a domestic environment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015b. doi: 10.1109/WASPAA.2015.7336899.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017a. doi: 10.1109/ICASSP.2017.7952261.

Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 776–780, 2017b. ISSN 15206149. doi: 10.1109/ICASSP.2017.7952261.

Nico Görnitz, Anne Porbadnigk, Alexander Binder, Claudia Sannelli, Mikio Braun, Klaus-Robert Mueller, and Marius Kloft. Learning and Evaluation in Presence of Non-i.i.d. Label Noise. In Samuel Kaski and Jukka Corander (eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 293–302, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL https://proceedings.mlr.press/v33/gornitz14.html.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017. URL https://arxiv.org/abs/1609.09430.

Shawn Hershey, Daniel P.W. Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R. Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 366–370, 2021. ISSN 15206149. doi: 10.1109/ICASSP39728.2021.9414579.

Eric Humphrey, Simon Durand, and Brian McFee. Openmic-2018: An open data-set for multiple instrument recognition. In *ISMIR*, pp. 438–444, 2018a.

Eric J. Humphrey, Simon Durand, and Brian McFee. OpenMIC-2018: An open dataset for multiple instrument recognition. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, pp. 438–444, 2018b.

Hoyoung Kim, Minhyeon Oh, Sehyun Hwang, Suha Kwak, and Jungseul Ok. Adaptive superpixel for active learning in semantic segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 943–953, 2023. doi: 10.1109/ICCV51070.2023.00093.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R.P. Lippmann (eds.), *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf.

Anurag Kumar and Bhiksha Raj. Audio event detection using weakly labeled data. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pp. 1038–1047, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450336031. doi: 10.1145/2964284.2964310. URL https://doi.org/10.1145/2964284.2964310.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably End-to-end Label-noise Learning without Anchor Points. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6403–6413. PMLR, July 2021. URL https://proceedings.mlr.press/v139/li21l.html. ISSN: 2640-3498.

Percy Liang, Michael I. Jordan, and Dan Klein. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 641–648, Montreal Quebec Canada, June 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553457. URL https://dl.acm.org/doi/10.1145/1553374.1553457.

Irene Martin-Morato and Annamaria Mesaros. Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31:902–914, 2023. ISSN 23299304. doi: 10.1109/TASLP.2022.3233468.

John Martinsson, Olof Mogren, Maria Sandsten, and Tuomas Virtanen. From Weak to Strong Sound Event Labels using Adaptive Change-Point Detection and Active Learning. In *EUSIPCO 2024 - 32nd European Signal Processing Conference*, 2024. URL http://arxiv.org/abs/2403.08525.

Brian McFee, Justin Salamon, and Juan Pablo Bello. Adaptive Pooling Operators for Weakly Labeled Sound Event Detection. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(11):2180–2193, November 2018. ISSN 2329-9290. doi: 10.1109/TASLP.2018.2858559. URL https://doi.org/10.1109/TASLP.2018.2858559.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences (Switzerland)*, 6(6), 2016. ISSN 20763417. doi: 10.3390/app6060162.

Annamaria Mesaros, Toni Heittola, and Dan Ellis. Datasets and evaluation. In Tuomas Virtanen, Mark Plumbley, and Dan Ellis (eds.), *Computational Analysis of Sound Scenes and Events*, chapter 6, pp. 147–179. Springer Cham, 2017.

Irene Martin Morato and Annamaria Mesaros. Mats - multi-annotator tagged soundscapes, May 2021. URL https://doi.org/10.5281/zenodo.4774960.

Irene Martin Morato, Manu Harju, and Annamaria Mesaros. Maestro real - multi-annotator estimated strong labels, 2023.

Ankit Shah, Anurag Kumar, Alexander G. Hauptmann, and Bhiksha Raj. A Closer Look at Weak Label Learning for Audio Events. pp. 1–10, 2018. URL http://arxiv.org/abs/1804.09288.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

Janvijay Singh and Raviraj Joshi. Background sound classification in speech audio segments. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–6, 2019. doi: 10.1109/SPED.2019.8906597.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from Noisy Labels with Deep Neural Networks: A Survey, March 2022. URL http://arxiv.org/abs/2007.08199. arXiv:2007.08199.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Ivo Trowitzsch, Jalil Taghia, Youssef Kashef, and Klaus Obermayer. The NIGENS General Sound Events Database. pp. 1–5, 2019. URL http://arxiv.org/abs/1902.08314.

Nicolas Turpault, Romain Serizel, Emmanuel Vincent, Nicolas Turpault, Romain Serizel, and Emmanuel Vincent. Analysis of weak labels for sound event tagging. 2021. URL https://hal.inria.fr/hal-03203692.

Jiachen Yao, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna, and Chao Chen. Learning to segment from noisy annotations: A spatial correction approach. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Qc_OopMEBnC.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, March 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3446776. URL https://dl.acm.org/doi/10.1145/3446776.

Jinyuan Zhou, Dading Guo, Yuhang Zhong, Jize Liu, Yu Yang, Yan Wang, and Ming-Hsuan Tan. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, 132:5308–5329, nov 2024. doi: 10.1007/s11263-024-02142-3. URL `https://doi.org/10.1007/s11263-024-02142-3`.

A Appendix

We do not include all simplifications of expressions in the proofs, but we do provide the code for a symbolic mathematics solver (SymPy) at GitHub[2], where all results can be verified. The notebook named "symbolic_verification_of_analysis.ipynb" can be used to verify the analysis.

A.1 Proof of Theorem 1

We will derive an expression for the expected query segment accuracy given overlap with a single event in terms of d_e, d_q, and γ, under all possible assumptions which will prove Theorem 1.

Proof. We need to consider two main assumptions. The first assumption is that the presence criterion for the annotator can be fulfilled, that is, $d_q \geq \gamma d_e$, and the second assumption is that the annotator presence criterion can not be fulfilled, that is, $d_q < \gamma d_e$. This happens if the query segment length is so short that it can never cover a large enough fraction of the event of interest to make presence detection feasible.

Assumption 1. The annotator presence criterion can be fulfilled ($d_q \geq \gamma d_e$).

Under this assumption there are two possible cases for the relation between d_q and d_e, either the event length is longer or equal to the query segment length, $d_e \geq d_q$ (case i), or the event length is shorter than the query segment length, $d_e < d_q$ (case ii). In Figure 15, we plot the query segment accuracy, $F(e_t, q, \gamma)$, for $t \in [0, d_e + d_q]$ for case (i) on the left, and case (ii) on the right. We describe in more detail in Appendix A.1.1 how the query segment accuracy behaves as a function of different amounts of overlap between the query segment and the event. Briefly, what we see in Figure 15 is that initially there is arbitrarily little overlap ($t_0^{(i)}$ and $t_0^{(ii)}$), an absence label is given to the query segment and the accuracy is therefore 1. Then the accuracy decrease linearly with the amount of overlap until the presence criterion is fulfilled and a presence label is given ($t_1^{(i)}$ and $t_1^{(ii)}$). After that, the accuracy linearly increase with the amount of overlap between the event and query segment until we reach a ceiling for the accuracy when either the whole query segment is inside the event ($t_2^{(i)}$) or the query segment covers the whole event ($t_2^{(ii)}$). Finally, the overlap between the query segment and the event starts to decrease again ($t_3^{(i)}$ and $t_3^{(ii)}$), and everything is symmetrical.

We continue by dropping the case superscripts show in the figure for $A_1, \ldots, A_3$ and $t_0, \ldots, t_5$, and only provide the full proof for case (i), but the proof for case (ii) is similar. In both cases the area A in Eq. 8 can be divided into five distinct parts:

$$A = 2A_1 + 2A_2 + A_3, \tag{21}$$

where A_1 and A_2 are counted twice due to symmetry.

[2] `https://github.com/johnmartinsson/the-accuracy-cost-of-weakness`

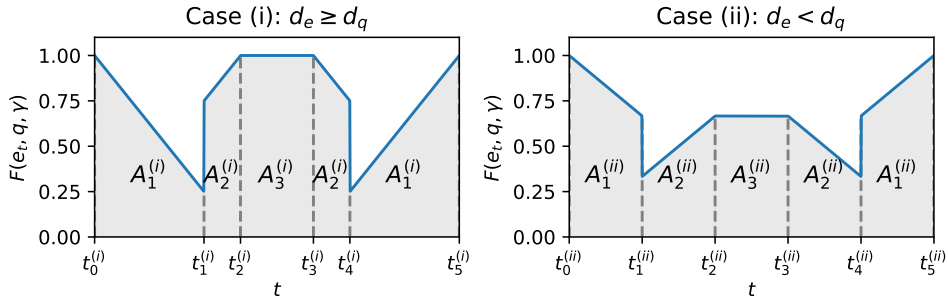


Figure 15: Assuming $d_q \geq d_e\gamma$, we plot the query segment accuracy, $F(e_t, q, \gamma)$, for $t \in [0, d_e + d_q]$, where $t_0 = 0$ and $t_5 = d_e + d_q$. Case (i) where $d_e \geq d_q$ is shown in the left panel, and case (ii) where $d_e < d_q$ is shown in the right panel.

The variables $t_0, t_1, \ldots, t_5$, represent the different states t of overlap where the discontinuities of $F(e_t, q, \gamma)$ occur, and using these we can express the areas as the following integrals:

$$A_1 = \int_{t_0}^{t_1} F(e_t, q, \gamma)\mathrm{d}t = \int_{t_4}^{t_5} F(e_t, q, \gamma)\mathrm{d}t, \tag{22}$$

and

$$A_2 = \int_{t_1}^{t_2} F(e_t, q, \gamma)\mathrm{d}t = \int_{t_3}^{t_4} F(e_t, q, \gamma)\mathrm{d}t, \tag{23}$$

due to symmetry, and

$$A_3 = \int_{t_2}^{t_3} F(e_q, q, \gamma)\mathrm{d}t. \tag{24}$$

We use that the query segment accuracy $F(e_t, q, \gamma)$ is linear in each interval, which means that the areas can be expressed as

$$A_1 = \frac{F(e_{t_0}, q, \gamma) + F(e_{t_1-}, q, \gamma)}{2}(t_1 - t_0), \tag{25}$$

$$A_2 = \frac{F(e_{t_1+}, q, \gamma) + F(e_{t_2}, q, \gamma)}{2}(t_2 - t_1), \tag{26}$$

and

$$A_3 = \frac{F(e_{t_2}, q, \gamma) + F(e_{t_3}, q, \gamma)}{2}(t_3 - t_2), \tag{27}$$

where t^- indicate that we approach the discontinuity at t from below and t^+ from above. We now only need to express $t_0, \ldots, t_3$ and $F(e_{t_0}, q, \gamma), \ldots, F(e_{t_3}, q, \gamma)$ in terms of d_e, d_q and γ to conclude the proof. For brevity, these have been provided in Table 2. See section A.1.1 for details on how to express these in terms of d_q, d_e and γ.

We provide the steps for case (i), and leave the derivation for case (ii) to the reader. We substitute the expressions for case (i), provided in Table 2, into equations Eq. 25-27, and the resulting expressions for the areas $A_1^{(i)}$, $A_2^{(i)}$, and $A_3^{(i)}$ into Eq. 21 which give

Case (i), $d_e \geq d_q$		Case (ii), $d_e < d_q$	
$t_0^{(i)} = 0$	$F(e_{t_0}^{(i)}, q, \gamma) = 1$	$t_0^{(ii)} = 0$	$F(e_{t_0}^{(ii)}, q, \gamma) = 1$
$t_1^{(i)} = \gamma d_e$	$F(e_{t_1^-}^{(i)}, q, \gamma) = \frac{d_q - \gamma d_e}{d_q}$	$t_1^{(ii)} = \gamma d_e$	$F(e_{t_1^-}^{(ii)}, q, \gamma) = \frac{d_q - \gamma d_e}{d_q}$
$t_2^{(i)} = d_q$	$F(e_{t_1^+}^{(i)}, q, \gamma) = \frac{\gamma d_e}{d_q}$	$t_2^{(ii)} = d_e$	$F(e_{t_1^+}^{(ii)}, q, \gamma) = \frac{\gamma d_e}{d_q}$
$t_3^{(i)} = d_e$	$F(e_{t_2}^{(i)}, q, \gamma) = 1$	$t_3^{(ii)} = d_q$	$F(e_{t_2}^{(ii)}, q, \gamma) = \frac{d_e}{d_q}$
	$F(e_{t_3}^{(i)}, q, \gamma) = 1$		$F(e_{t_3}^{(ii)}, q, \gamma) = \frac{d_e}{d_q}$

Table 2: A summary of the derived expressions for $t_0, \ldots, t_3$ and $F(e_{t_0}, q, \gamma), \ldots, F(e_{t_3}, q, \gamma)$ for each case. $F(e_{t_1^-}, q, \gamma)$ and $F(e_{t_1^+}, q, \gamma)$ denotes the limits when approaching t_1 from below and above respectively.

$$
\begin{aligned}
A^{(i)} &= \frac{2}{2}(1 + \frac{d_q - \gamma d_e}{d_q})\gamma d_e + \frac{2}{2}(1 + \frac{\gamma d_e}{d_q})(d_q - \gamma d_e) + (d_e - d_q) \\
&= (2d_q - \gamma d_e)\frac{\gamma d_e}{d_q} + (d_q + \gamma d_e)(d_q - \gamma d_e)\frac{1}{d_q} + (d_e - d_q) \\
&= \frac{1}{d_q}(2\gamma d_q d_e - \gamma^2 d_e^2 + \cancel{d_q^2} - \gamma^2 d_e^2 + d_e d_q - \cancel{d_q^2}) \\
&= \frac{1}{d_q}(2\gamma d_q d_e - 2\gamma^2 d_e^2 + d_e d_q) \\
&= \frac{d_e}{d_q}(2\gamma d_q - 2\gamma^2 d_e + d_q).
\end{aligned}
$$

Finally, by substituting A for $A^{(i)}$ in Eq. 8 we arrive at

$$
\frac{A^{(i)}}{d_e + d_q} = \frac{d_e(2\gamma d_q - 2\gamma^2 d_e + d_q)}{d_q(d_e + d_q)} \tag{28}
$$

which shows that Eq. 9 holds for case (i) under the assumption that $d_q \geq \gamma d_e$. Similarly, this also holds for case (ii).

Assumption 2. The annotator presence criterion can not be fulfilled ($d_q < \gamma d_e$).

When the presence criterion can not be fulfilled we never get any presence labels, this means that the fraction of the query segment that overlaps with an event is always incorrectly given an absence label. When the query segment completely overlaps with an event the query segment accuracy will be 0 (seen between t_1 and t_2 in Figure 16).

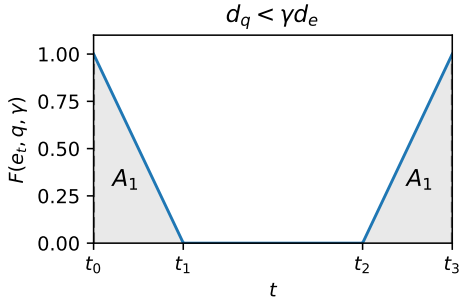


Figure 16: Assuming that $d_q < \gamma d_e$, we plot the query segment accuracy, $F(e_t, q, \gamma)$, for $t \in [0, d_e + d_q]$, where $t_0 = 0$ and $t_3 = d_e + d_q$.

The area A_1 is counted twice due to symmetry. The discontinuity at t_1 occurs for the smallest $t \in [0, d_e + d_q]$ for which $F(e_t, q, \gamma) = 0$, which happens for the smallest t for which the whole query segment overlaps with the event $|e \cap q| = d_q$ at $t = d_q$. We therefore have that $t_1 - t_0 = t_3 - t_2 = d_q$.

When there is no overlap between the query segment and the event giving a presence label is always correct, thus $F(e_{t_0}, q, \gamma) = 1$. However, giving an absence label to a query segment that completely overlaps with an event gives the query segment accuracy 0, thus $F(e_{t_1}, q, \gamma) = 0$. The total area under the curve is therefore $2A_1 = d_q$ and by normalizing with $t_3 - t_0 = d_e + d_q$, we get $d_q/(d_e + d_q)$, which proves the $d_q < \gamma d_e$ case of Eq. 9, and concludes the proof.

□

A.1.1 Details on the expressions in Table 2

This section provides a detailed explanation of the values presented in Table 2. For each case (i) and (ii), we will define the specific time points t_0, t_1, t_2, t_3 where the query segment accuracy function $F(e_t, q, \gamma)$ changes, and explain the corresponding value of the function at these points based on the overlap between the event e_t and the query segment q. The states t_4 and t_5 are analogous to t_1 and t_0, respectively, and therefore not illustrated. The difference is that the amount of overlap between the query segment and event decreases (instead of increases) when approaching these states.

Case (i): $d_e \geq d_q$

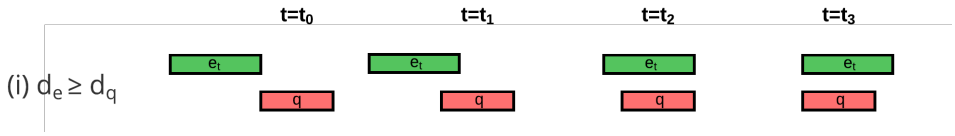


Figure 17: An illustration of how the sound event e_t and the query segment q overlap at the four distinct states $t = t_0, \ldots, t_3$ for case (i) where $d_e \geq d_q$.

- $t_0^{(i)}$: At $t_0^{(i)} = 0$, the end of the event e_t aligns perfectly with the beginning of the query segment q. This means there is no overlap between the event and the query segment ($|e_{t_0^{(i)}} \cap q| = 0$). Therefore, assuming the annotator absence criterion applies, the query segment accuracy is $F(e_{t_0^{(i)}}, q, \gamma) = \frac{d_q - |e_{t_0^{(i)}} \cap q|}{d_q} = \frac{d_q - 0}{d_q} = 1$.
- $t_1^{(i)}$: The time $t_1^{(i)} = \gamma d_e$ represents the point where the annotator presence criterion is first met. Before this point ($t < t_1^{(i)}$), the overlap $|e_t \cap q|$ is less than γd_e, and the query segment accuracy is given by $F(e_t, q, \gamma) = \frac{d_q - |e_t \cap q|}{d_q}$. As t approaches $t_1^{(i)}$ from the left, $|e_t \cap q|$ approaches γd_e, hence $\lim_{t \to t_1^-} F(e_t, q, \gamma) = \frac{d_q - \gamma d_e}{d_q}$. At $t = t_1^{(i)}$, the presence criterion is met, and the accuracy function switches to $F(e_t, q, \gamma) = \frac{|e_t \cap q|}{d_q}$. As t approaches $t_1^{(i)}$ from the right, $|e_t \cap q|$ is slightly greater than γd_e, and $\lim_{t \to t_1^+} F(e_t, q, \gamma) = \frac{\gamma d_e}{d_q}$. This transition is visually represented in Figure 17 at time $t = t_1$.
- $t_2^{(i)}$: At $t_2^{(i)} = d_q$, the entire query segment q is fully contained within the event e_t. This means the overlap is maximal: $|e_{t_2^{(i)}} \cap q| = d_q$. Since the presence criterion is met, the query segment accuracy is $F(e_{t_2^{(i)}}, q, \gamma) = \frac{|e_{t_2^{(i)}} \cap q|}{d_q} = \frac{d_q}{d_q} = 1$. This behavior is visually represented in Figure 17 at time $t = t_2$, where the green box representing the event fully covers the red box representing the query segment.
- $t_3^{(i)}$: At $t_3^{(i)} = d_e$, the entire query segment q still fully overlaps with the event e_t. Similar to t_2, the overlap is $|e_{t_3^{(i)}} \cap q| = d_q$, and therefore $F(e_{t_3^{(i)}}, q, \gamma) = \frac{|e_{t_3^{(i)}} \cap q|}{d_q} = \frac{d_q}{d_q} = 1$. This is depicted in Figure 17 at time $t = t_3$.

Case (ii): $d_e < d_q$

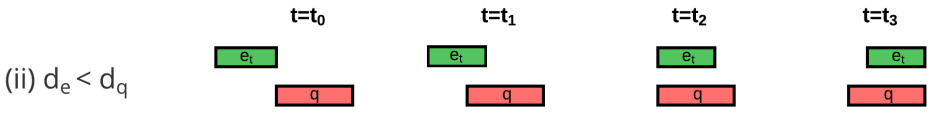


Figure 18: An illustration of how the sound event e_t and the query segment q overlap at the four distinct states $t = t_0, \ldots, t_3$ for case (ii) where $d_e < d_q$.

- $t_0^{(ii)}$: At $t_0^{(ii)} = 0$, the end of the event e_t aligns perfectly with the beginning of the query segment q. There is no overlap ($|e_{t_0^{(ii)}} \cap q| = 0$). Assuming the annotator absence criterion applies, the query segment accuracy is $F(e_{t_0^{(ii)}}, q, \gamma) = \frac{d_q - |e_{t_0^{(ii)}} \cap q|}{d_q} = \frac{d_q - 0}{d_q} = 1$.
- $t_1^{(ii)}$: The time $t_1^{(ii)} = \gamma d_e$ again marks the point where the annotator presence criterion is first met. Before this ($t < t_1^{(ii)}$), the overlap $|e_t \cap q| < \gamma d_e$, and $F(e_t, q, \gamma) = \frac{d_q - |e_t \cap q|}{d_q}$. Approaching $t_1^{(ii)}$ from

the left, $|e_t \cap q| \to \gamma d_e$, thus $\lim_{t \to t_1^-} F(e_t, q, \gamma) = \frac{d_q - \gamma d_e}{d_q}$. At $t = t_1^{(ii)}$, the criterion is met, and the function becomes $F(e_t, q, \gamma) = \frac{|e_t \cap q|}{d_q}$. Approaching from the right, $|e_t \cap q|$ is slightly greater than γd_e, so $\lim_{t \to t_1^+} F(e_t, q, \gamma) = \frac{\gamma d_e}{d_q}$. This transition is shown in Figure 18 at $t = t_1$.

- $t_2^{(ii)}$: At $t_2^{(ii)} = d_e$, the beginning of the event e_t aligns with the beginning of the query segment q. At this point, the overlap is maximal, as the entire event is contained within the query segment: $|e_{t_2^{(ii)}} \cap q| = d_e$. Since the presence criterion is met, the query segment accuracy is $F(e_{t_2^{(ii)}}, q, \gamma) = \frac{|e_{t_2^{(ii)}} \cap q|}{d_q} = \frac{d_e}{d_q}$. This situation is illustrated in Figure 18 at $t = t_2$.
- $t_3^{(ii)}$: At $t_3^{(ii)} = d_q$, the end of the event e_t aligns with the end of the query segment q. Similar to $t_2^{(ii)}$, the entire event is contained within the query segment, so the overlap is $|e_{t_3^{(ii)}} \cap q| = d_e$. Consequently, the query segment accuracy is $F(e_{t_3^{(ii)}}, q, \gamma) = \frac{|e_{t_3^{(ii)}} \cap q|}{d_q} = \frac{d_e}{d_q}$. This corresponds to the state depicted in Figure 18 at $t = t_3$.

Understanding these key time points and the corresponding query segment accuracy values is crucial for calculating the area under the curve, which represents the expected query segment accuracy.

A.2 Proof of Theorem 2

Proof. We start by finding a unique critical point d_q^* which makes $f'(d_q^*) = 0$ when $d_q \geq \gamma d_e$. We then show that d_q^* is a global maximum by analyzing the boundaries of $f(d_q)$ on its' domain when $d_q \geq \gamma d_e$. We show that $f(d_q^*) \geq f(\gamma d_e)$ and that $f(d_q^*) \geq \lim_{d_q \to \infty} f(d_q)$. Since d_q^* is a unique critical point we conclude that it must be a global maximum of the function $f(d_q)$ when $d_q \geq \gamma d_e$. Lastly, we show that $f(d_q^*) \geq f(\gamma d_e) \geq f(d_q)$ when $d_q < \gamma d_e$ which proves that d_q^* is a global maximum of the function $f(d_q)$ for $d_q > 0$.

1. Finding the unique critical point d_q^*.

To find the critical points, we need to compute the derivative of $f(d_q)$ with respect to d_q and set it to zero. Let $N(d_q) = d_e(-2d_e\gamma^2 + 2d_q\gamma + d_q)$ and $D(d_q) = d_q(d_e + d_q)$. Then $f(d_q) = \frac{N(d_q)}{D(d_q)}$. Using the quotient rule, the derivative is given by:

$$f'(d_q) = \frac{N'(d_q)D(d_q) - N(d_q)D'(d_q)}{[D(d_q)]^2}$$

First, we find the derivatives of the numerator and the denominator:

$$\begin{aligned} N'(d_q) &= \frac{d}{dd_q}[d_e(-2d_e\gamma^2 + 2d_q\gamma + d_q)] \\ &= d_e(0 + 2\gamma + 1) \\ &= d_e(2\gamma + 1) \end{aligned}$$

$$\begin{aligned} D(d_q) &= d_q(d_e + d_q) = d_e d_q + d_q^2 \\ D'(d_q) &= \frac{d}{dd_q}[d_e d_q + d_q^2] \\ &= d_e + 2d_q \end{aligned}$$

Now, we plug these into the quotient rule formula:

$$f'(d_q) = \frac{[d_e(2\gamma + 1)][d_q(d_e + d_q)] - [d_e(-2d_e\gamma^2 + 2d_q\gamma + d_q)][d_e + 2d_q]}{[d_q(d_e + d_q)]^2}$$

To find the critical points, we set $f'(d_q) = 0$, which means the numerator must be zero:

$$[d_e(2\gamma+1)][d_q(d_e+d_q)] - [d_e(-2d_e\gamma^2 + 2d_q\gamma + d_q)][d_e + 2d_q] = 0$$

Since $d_e > 0$, we can divide by d_e:

$$(2\gamma+1)d_q(d_e+d_q) - (-2d_e\gamma^2 + 2d_q\gamma + d_q)(d_e + 2d_q) = 0$$

Expanding the terms:

$$\begin{aligned}(2\gamma+1)(d_e d_q + d_q^2) - (-2d_e^2\gamma^2 - 4d_e d_q\gamma^2 + 2d_e d_q\gamma + 4d_q^2\gamma + d_e d_q + 2d_q^2) = 0 \\ 2\gamma d_e d_q + 2\gamma d_q^2 + d_e d_q + d_q^2 - (-2d_e^2\gamma^2 - 4d_e d_q\gamma^2 + 2d_e d_q\gamma + 4d_q^2\gamma + d_e d_q + 2d_q^2) = 0\end{aligned}$$

Collecting and rearranging the terms to form a quadratic equation in d_q:

$$\begin{aligned}(2\gamma + 1 - 4\gamma - 2)d_q^2 + (2\gamma + 1 + 4\gamma^2 - 2\gamma - 1)d_e d_q + 2d_e^2\gamma^2 = 0 \\ (-2\gamma - 1)d_q^2 + (4\gamma^2)d_e d_q + 2d_e^2\gamma^2 = 0 \\ (2\gamma+1)d_q^2 - 4\gamma^2 d_e d_q - 2d_e^2\gamma^2 = 0\end{aligned}$$

Using the quadratic formula $d_q = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where $a = 2\gamma + 1$, $b = -4d_e\gamma^2$, $c = -2d_e^2\gamma^2$:

$$\begin{aligned}d_q &= \frac{4d_e\gamma^2 \pm \sqrt{(-4d_e\gamma^2)^2 - 4(2\gamma+1)(-2d_e^2\gamma^2)}}{2(2\gamma+1)} \\ &= \frac{4d_e\gamma^2 \pm \sqrt{16d_e^2\gamma^4 + 8(2\gamma+1)d_e^2\gamma^2}}{4\gamma+2} \\ &= \frac{4d_e\gamma^2 \pm \sqrt{16d_e^2\gamma^4 + 16d_e^2\gamma^3 + 8d_e^2\gamma^2}}{4\gamma+2} \\ &= \frac{4d_e\gamma^2 \pm \sqrt{8d_e^2\gamma^2(2\gamma^2 + 2\gamma + 1)}}{4\gamma+2} \\ &= \frac{4d_e\gamma^2 \pm 2d_e|\gamma|\sqrt{4\gamma^2 + 4\gamma + 2}}{2(2\gamma+1)}\end{aligned}$$

Since $\gamma > 0$, we have $|\gamma| = \gamma$:

$$\begin{aligned}d_q &= \frac{4d_e\gamma^2 \pm 2d_e\gamma\sqrt{4\gamma^2 + 4\gamma + 2}}{2(2\gamma+1)} \\ &= \frac{2d_e\gamma^2 \pm d_e\gamma\sqrt{4\gamma^2 + 4\gamma + 2}}{(2\gamma+1)} \\ &= d_e\gamma\frac{2\gamma \pm \sqrt{4\gamma^2 + 4\gamma + 2}}{2\gamma+1}\end{aligned}$$

We note that $\sqrt{4\gamma^2 + 4\gamma + 2} = 2\sqrt{\gamma^2 + \gamma + 0.5} > 2\sqrt{\gamma^2} = 2\gamma$, which means that we need to choose the positive sign for $d_q > 0$ to be true. The value of d_q that makes the derivative zero is therefore uniquely defined by:

$$d_q = d_e\,\gamma\frac{2\,\gamma + \sqrt{4\,\gamma^2 + 4\,\gamma + 2}}{2\,\gamma + 1} \geq d_e\gamma,$$

where the last inequality holds because $\sqrt{4\gamma^2 + 4\gamma + 2} = 2\sqrt{\gamma^2 + \gamma + 0.5} \geq 1$.

2. Analyze the function at the boundaries of its' domain.

To understand why this critical point corresponds to a maximum, we analyze the function $f(d_q)$ as d_q at the boundaries of its domain.

2a. $f(d_q)$ **when** $d_q = \gamma d_e$ **(**$d_q \geq \gamma d_e$**).**

$$\begin{aligned} f(\gamma d_e) &= \frac{d_e\left(2(\gamma d_e)\gamma - 2d_e\gamma^2 + (\gamma d_e)\right)}{(\gamma d_e)\,(d_e + \gamma d_e)} \\ &= \frac{d_e\left(2d_e\gamma^2 - 2d_e\gamma^2 + \gamma d_e\right)}{(\gamma d_e)\,(d_e + \gamma d_e)} \\ &= \frac{d_e\,(\gamma d_e)}{(\gamma d_e)\,(d_e + \gamma d_e)} \\ &= \frac{d_e^2\gamma}{(\gamma d_e)d_e(1+\gamma)} \\ &= \frac{1}{1+\gamma}. \end{aligned}$$

2b. $f(d_q)$ **as** $d_q \to \infty$ **(**$d_q \geq \gamma d_e$**).**

We want to evaluate the limit of $f(d_q)$ as d_q approaches infinity:

$$\lim_{d_q \to \infty} f(d_q) = \lim_{d_q \to \infty} \frac{d_e(-2d_e\gamma^2 + (2\gamma+1)d_q)}{d_e d_q + d_q^2}$$

Divide the numerator and the denominator by the highest power of d_q in the denominator, which is d_q^2:

$$\lim_{d_q \to \infty} f(d_q) = \lim_{d_q \to \infty} \frac{d_e\left(-\frac{2d_e\gamma^2}{d_q^2} + \frac{2\gamma+1}{d_q}\right)}{\frac{d_e}{d_q} + 1}$$

As $d_q \to \infty$, the terms $\frac{2d_e\gamma^2}{d_q^2}$, $\frac{2\gamma+1}{d_q}$, and $\frac{d_e}{d_q}$ all approach 0. Thus,

$$\lim_{d_q \to \infty} f(d_q) = \frac{d_e(0+0)}{0+1} = 0$$

This means that as d_q becomes very large, the function $f(d_q)$ approaches 0.

2c. Showing that $f(d_q^*) \geq f(\gamma d_e)$**.**

We want to show that $f(d_q^*) \geq f(\gamma d_e)$. Or equivalently, that $f(d_q^*) - f(\gamma d_e) \geq 0$. From Theorem 3 we know that $f(d_q^*) = 2\gamma\left(2\gamma + 1 - \sqrt{4\gamma^2 + 4\gamma + 2}\right) + 1$, and from 2a we know that $f(\gamma d_e) = \frac{1}{1+\gamma}$. After substitution and some algebraic manipulation, we get

$$\gamma\left(\frac{4\gamma^2 + 6\gamma + 3}{1+\gamma} - 2\sqrt{4\gamma^2 + 4\gamma + 2}\right) \geq 0.$$

Since $\gamma > 0$, it suffices to show that

$$\frac{4\gamma^2 + 6\gamma + 3}{1+\gamma} \geq 2\sqrt{4\gamma^2 + 4\gamma + 2}.$$

Squaring both sides of the above inequality and simplifying, we obtain the equivalent inequality

$$\left(\frac{4\gamma^2 + 6\gamma + 3}{1+\gamma}\right)^2 \geq 4(4\gamma^2 + 4\gamma + 2).$$

After further algebraic manipulations (which we leave to the reader), we arrive at the inequality

$$(2\gamma + 1)^2 \geq 0.$$

Since $(2\gamma + 1)^2 \geq 0$ holds for all γ, and the previous steps are all equivalences, we conclude that

$$f(d_q^*) - f(\gamma d_e) \geq 0$$

for $0 < \gamma \leq 1$, and therefore,

$$f(d_q^*) \geq f(\gamma d_e).$$

2d. Showing that $f(d_q^*) \geq \lim_{d_q \to \infty} f(d_q)$.

We combine the results from 2a-2c to get

$$\begin{aligned} f(d_q^*) &\geq f(\gamma d_e) \\ &= \frac{1}{1+\gamma} \\ &\geq 0 \\ &= \lim_{d_q \to \infty} f(d_q). \end{aligned}$$

2e. $f(d_q)$ as $d_q \to (\gamma d_e)^-$ ($d_q < \gamma d_e$).

Since we are approaching γd_e from the left, we have that $f(d_q) = d_q/(d_e + d_q)$. This function is continuous for $d_q < \gamma d_e$, so the limit is given by the direct substitution:

$$\begin{aligned} \lim_{d_q \to (\gamma d_e)^-} \frac{d_q}{d_e + d_q} &= \frac{\gamma d_e}{d_e + \gamma d_e} \\ &= \frac{\gamma d_e}{d_e(1+\gamma)} \\ &= \frac{\gamma}{1+\gamma} \end{aligned}$$

2f. Showing that $f(\gamma d_e) \geq f(d_q)$ when $d_q < \gamma d_e$. We start by noting that $f(\gamma d_e) = \frac{1}{1+\gamma} \geq \frac{\gamma}{1+\gamma} = \lim_{d_q \to (\gamma d_e)^-}$. Now it is sufficient to show that $f(d_q) = d_q/(d_q + d_e)$ is strictly decreasing for decreasing d_q, which we do by computing the derivative of $f(d_q)$ with respect to d_q using the quotient rule:

$$\begin{aligned} f'(d_q) &= \frac{(d_q + \gamma)(1) - d_q(1)}{(d_q + \gamma)^2} \\ &= \frac{d_q + \gamma - d_q}{(d_q + \gamma)^2} \\ &= \frac{\gamma}{(d_q + \gamma)^2}. \end{aligned}$$

Since $\gamma > 0$ and $(d_q + \gamma)^2 > 0$ for all $d_q > 0$, we have $f'(d_q) > 0$ for all $d_q > 0$. This implies that the function $f(d_q)$ is strictly increasing on the interval $(0, \infty)$. Therefore, if $0 < c \leq b$, it must be the case that $f(c) \leq f(b)$. Moreover, since $c < b$, $f(c) < f(b)$. Thus, for any $b > 0$, $f(b) > f(c)$ for all $0 < c \leq b$. Now let $0 < d_q = c \leq \gamma d_e = b$.

3. Combining everything (2a-2f)

We have derived a unique critical point $d_q^* \geq \gamma d_e$ by setting the first derivative of $f(d_q)$ to zero. We have then shown that $f(d_q^*)$ is greater than or equal to $f(d_q)$ at the limits of its' domain when $d_q \geq \gamma d_e$. Finally, we show that $f(d_q^*) \geq f(\gamma d_e) \geq f(d_q)$ when $d_q < \gamma d_e$. Therefore, the value of d_q that is the global maximum of $f(d_q)$ when $d_q > 0$ is:

$$\boxed{d_q^* = d_e \, \gamma \frac{2\,\gamma + \sqrt{4\,\gamma^2 + 4\,\gamma + 2}}{2\,\gamma + 1}}$$

□

A.3 Proof of Theorem 3

Proof. From Theorem 2 we have that

$$d_q^* \;=\; \frac{d_e\,\gamma\left(2\,\gamma + \sqrt{4\,\gamma^2 + 4\,\gamma + 2}\right)}{2\,\gamma + 1}$$

maximizes the function

$$f(d_q) \;=\; \frac{d_e\big(-2\,d_e\,\gamma^2 \;+\; (2\,\gamma + 1)\,d_q\big)}{d_q\,\big(d_e + d_q\big)}.$$

We wish to show that the maximum label accuracy given overlap, $f^*(\gamma) = f\big(d_q^*\big)$, is

$$2\,\gamma\Big(2\,\gamma + 1 \;-\; \sqrt{4\,\gamma^2 + 4\,\gamma + 2}\Big) \;+\; 1.$$

1. Express $f(d_q)$ in terms of a dimensionless variable.

Define

$$\delta \;=\; \frac{d_q}{d_e}.$$

Then

$$d_q \;=\; \delta\,d_e, \quad d_e + d_q \;=\; d_e\,(1 + \delta),$$

and

$$f(d_q) \;=\; f(\delta\,d_e) \;=\; \frac{d_e\big(-2\,d_e\,\gamma^2 + (2\,\gamma + 1)\,\delta\,d_e\big)}{(\delta\,d_e)\,\big(d_e + \delta\,d_e\big)} = \frac{-2\,\gamma^2 + (2\,\gamma + 1)\,\delta}{\delta\,\big(1 + \delta\big)}.$$

We can therefore write

$$f(\delta) \;=\; \frac{-2\,\gamma^2 \;+\; (2\,\gamma + 1)\,\delta}{\delta\,(1 + \delta)}.$$

2. Identify the optimal dimensionless query length δ^*.

From Theorem 2, we know that

$$d_q^* \;=\; \frac{d_e\,\gamma\,\big(2\,\gamma \;+\; \sqrt{4\,\gamma^2 + 4\,\gamma + 2}\big)}{2\,\gamma + 1}.$$

Dividing both sides by d_e gives

$$\delta^* \;=\; \frac{d_q^*}{d_e} \;=\; \gamma\,\frac{2\,\gamma \;+\; \sqrt{4\,\gamma^2 + 4\,\gamma + 2}}{2\,\gamma + 1}.$$

We need to show that

$$f\big(\delta^*\big) \;=\; 2\,\gamma\Big(2\,\gamma + 1 - \sqrt{4\,\gamma^2 + 4\,\gamma + 2}\Big) \;+\; 1.$$

3. Compute $f(\delta^*)$ explicitly.

Let

$$N(\delta) \;=\; -2\,\gamma^2 \;+\; (2\,\gamma + 1)\,\delta, \quad D(\delta) \;=\; \delta\,(1 + \delta).$$

Then $f(\delta) = \frac{N(\delta)}{D(\delta)}$.

1. *Numerator at δ^*.*

$$N(\delta^*) = -2\gamma^2 \;+\; (2\gamma+1)\,\delta^* = -2\gamma^2 \;+\; (2\gamma+1)\Big[\gamma\,\frac{2\gamma+\sqrt{4\gamma^2+4\gamma+2}}{2\gamma+1}\Big].$$

Inside the brackets, $(2\gamma+1)$ cancels:

$$N(\delta^*) = -2\gamma^2+\gamma\left(2\gamma+\sqrt{4\gamma^2+4\gamma+2}\right) = -2\gamma^2+2\gamma^2+\gamma\,\sqrt{4\gamma^2+4\gamma+2} = \gamma\,\sqrt{4\gamma^2+4\gamma+2}.$$

2. *Denominator at δ^*.*

$$D(\delta) \;=\; \delta\,(1+\delta).$$

Hence,

$$D(\delta^*) = \delta^*\left(1+\delta^*\right) = \Big[\gamma\,\frac{2\gamma+\sqrt{4\gamma^2+4\gamma+2}}{2\gamma+1}\Big]\Big[1 \;+\; \gamma\,\frac{2\gamma+\sqrt{4\gamma^2+4\gamma+2}}{2\gamma+1}\Big].$$

The second bracket becomes a single fraction:

$$1+\gamma\,\frac{2\gamma+\sqrt{4\gamma^2+4\gamma+2}}{2\gamma+1} = \frac{(2\gamma+1) \;+\; \gamma\left(2\gamma+\sqrt{4\gamma^2+4\gamma+2}\right)}{2\gamma+1}.$$

Combining, we get

$$D(\delta^*) = \gamma\,\frac{2\gamma+\sqrt{4\gamma^2+4\gamma+2}}{2\gamma+1} \;\times\; \frac{(2\gamma+1)+2\gamma^2+\gamma\,\sqrt{4\gamma^2+4\gamma+2}}{2\gamma+1}.$$

So

$$D(\delta^*) = \gamma\,\frac{(2\gamma+\sqrt{4\gamma^2+4\gamma+2})\left(2\gamma+1+2\gamma^2+\gamma\,\sqrt{4\gamma^2+4\gamma+2}\right)}{(2\gamma+1)^2}.$$

3. *Form the ratio.* Thus,

$$f(\delta^*) = \frac{N(\delta^*)}{D(\delta^*)} = \frac{\gamma\,\sqrt{4\gamma^2+4\gamma+2}}{\gamma\,\frac{(2\gamma+\sqrt{4\gamma^2+4\gamma+2})\left(2\gamma+1+2\gamma^2+\gamma\,\sqrt{4\gamma^2+4\gamma+2}\right)}{(2\gamma+1)^2}}.$$

Cancel the common factor γ, invert the denominator and multiply:

$$f(\delta^*) = \frac{\sqrt{4\gamma^2+4\gamma+2}\,(2\gamma+1)^2}{(2\gamma+\sqrt{4\gamma^2+4\gamma+2})\left(2\gamma+1+2\gamma^2+\gamma\,\sqrt{4\gamma^2+4\gamma+2}\right)}.$$

You can verify by direct expansion (or by a symbolic algebra tool which we provide in the supplementary material) that

$$\frac{\sqrt{4\gamma^2+4\gamma+2}\,(2\gamma+1)^2}{(2\gamma+\sqrt{4\gamma^2+4\gamma+2})\left(2\gamma+1+2\gamma^2+\gamma\,\sqrt{4\gamma^2+4\gamma+2}\right)} = 2\gamma\left(2\gamma+1-\sqrt{4\gamma^2+4\gamma+2}\right) \;+\; 1.$$

Thus

$$f(\delta^*) \;=\; 2\gamma\left(2\gamma+1-\sqrt{4\gamma^2+4\gamma+2}\right)+1,$$

which proves that

$$f^*(\gamma) = f(d_q^*) = 2\gamma\left(2\gamma+1-\sqrt{4\gamma^2+4\gamma+2}\right) \;+\; 1.$$

Hence, Eq. 11 holds, completing the proof. □

A.4 Empirical Analysis of Uniform Relative Offset Distribution Between Events and Overlapping Segments

We empirically verify that the relative offset between events and overlapping query segments can be modeled well by a uniform distribution. An event is denoted by $e = (a_e, b_e, c_e)$, where a_e is the start time, b_e is the end time, and c_e is the class, where $c_e \in \{\text{"dogs"}, \text{"baby"}\}$. Similarly, a query segment is denoted by $q = (a_q, b_q)$. We define the event distribution using the labeled start times of different sound event classes from the NIGENS Trowitzsch et al. (2019) dataset, but we fix the event length d_e to the median event length of the respective sound class to respect the deterministic event length assumption. We then verify that the relative offset between the events and the overlapping segments, defined as $b_q - a_e$, is uniform over the range $[0, d_e + d_q]$. To simulate realistic scenarios that maintain a reasonable label accuracy, we let $d_q \in \{\frac{d_e}{10}, d_e, 10d_e\}$. That is, the query segment is not larger than 10 times the median event length. Note that if $d_q \gg d_e$ then the uniform relative offset assumption is not expected to hold, but that also means that the label accuracy will be very low which is not wanted in practice. We present the results in Figure 19. For both sound event classes the distribution looks flat for all three choices of d_q, meaning that it can be modeled well by a uniform distribution, verifying that this is a plausible assumption in practical scenarios.

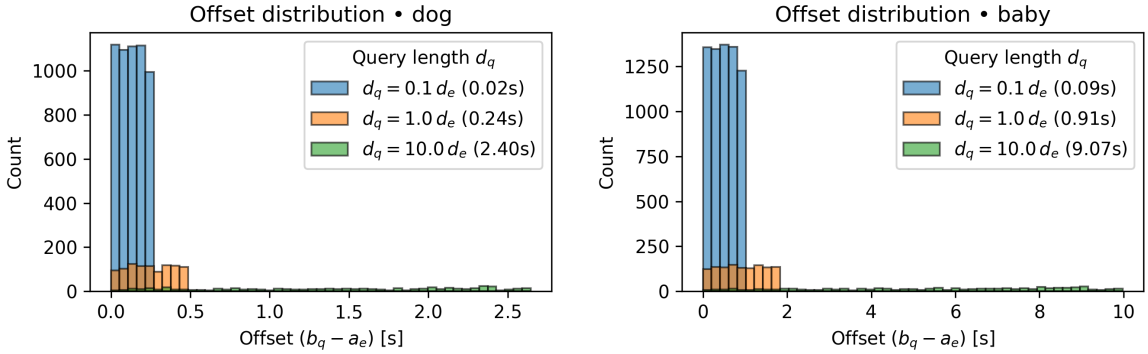


Figure 19: The distribution of relative offsets between events and overlapping query segments for dog (left) and baby (right) sound event classes from the NIGENS Trowitzsch et al. (2019) dataset. We use the annotated relative start times of the events, and fix the event length d_e to the median event length for the respective event class. The distribution looks flat and can be modeled well with a uniform distribution.

A.5 Empirical Analysis of Theory

In an attempt to empirically validate the theory we have compared the weakly labeled version of AudioSet Hershey et al. (2017) with the strongly labeled version Hershey et al. (2021). The weakly labeled version of AudioSet uses 10 second segments, corresponding to $d_q = 10$. A subset of AudioSet has been strongly labeled by indicating start and end times of the weakly labeled events. For each strongly labeled event, we compare the strong label to the weak label to compute the accuracy. That is, if the weakly labeled segment of length d_q indicates the presence of an event, and the corresponding strong label for that event has length d_e, then the accuracy is $\frac{d_e}{d_q}$ for that event. We compute this accuracy for all sound events corresponding to the "Animal" class, and take the average. The theoretical accuracy for a given annotator criterion γ is derived by taking the numerical average over the event lengths for the "Animal" class to estimate the numerical integration over the event length distribution presented Eq. (13).

The results are shown in Figure 20. Note that the query segment is not chosen to maximize label accuracy as in previous analysis. Since $d_q = 10$ it is longer than most 'Animal' events in AudioSet and restricts the maximum event length that can be annotated to $d_e \leq d_q$ as can be seen in the right panel of Figure 20. In the left panel of Figure 20 we see that the label accuracy of the weakly labeled version of AudioSet falls within the range predicted by the theory for $\gamma \in (0, 1]$. Assuming that the theory is correct would indicate that $\gamma = 0.26$ is the presence criterion that best models the weak labeling process of AudioSet. While these

results do not reject the proposed theory we would need to empirical estimate γ based on real annotators to properly validate it. This is considered as out of scope for this paper, but would be interesting future work.

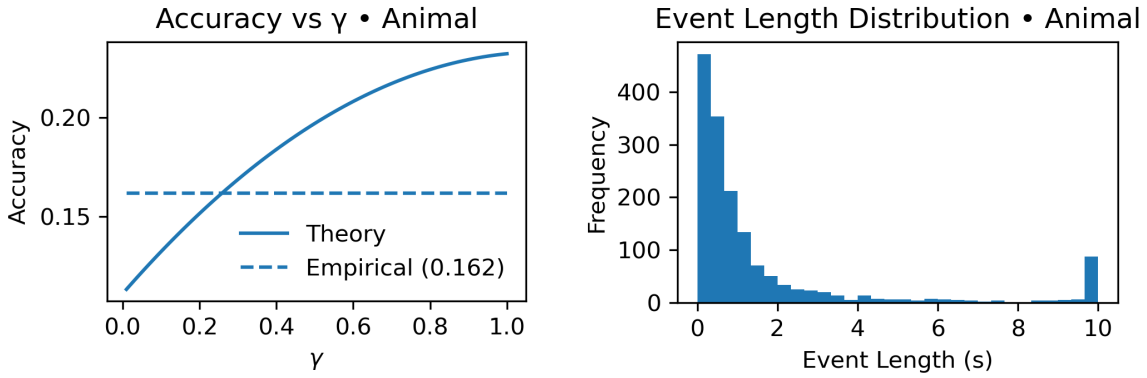


Figure 20: The theoretical prediction of the label accuracy for different γ (left) when averaging over the event length distribution (right) for the animal events in the strongly labeled subset of AudioSet. The empirical accuracy (dashed blue line) indicates the label accuracy that was derived by comparing the weakly labeled version of AudioSet with the strongly labeled version. The empirical label accuracy falls within the range predicted by the theory.

Aggregation Strategies for Efficient Annotation of Bioacoustic Sound Events Using Active Learning

Richard Lindholm[§,1], Oscar Marklund[§,1]
1. Centre for Mathematical Sciences, Faculty of Engineering
Lund University, Sweden

Olof Mogren[2,3], John Martinsson[1,2,3]
2. RISE Research Institutes of Sweden
3. Climate AI Nordics

***Abstract*—The vast amounts of audio data collected in Sound Event Detection (SED) applications require efficient annotation strategies to enable supervised learning. Manual labeling is expensive and time-consuming, making Active Learning (AL) a promising approach for reducing annotation effort. We introduce *Top K Entropy*, a novel uncertainty aggregation strategy for AL that prioritizes the most uncertain segments within an audio recording, instead of averaging uncertainty across all segments. This approach enables the selection of entire recordings for annotation, improving efficiency in sparse data scenarios. We compare *Top K Entropy* to random sampling and *Mean Entropy*, and show that fewer labels can lead to the same model performance, particularly in datasets with sparse sound events. Evaluations are conducted on audio mixtures of sound recordings from parks with meerkat, dog, and baby crying sound events, representing real-world bioacoustic monitoring scenarios. Using *Top K Entropy* for active learning, we can achieve comparable performance to training on the fully labeled dataset with only 8% of the labels. *Top K Entropy* outperforms *Mean Entropy*, suggesting that it is best to let the most uncertain segments represent the uncertainty of an audio file. The findings highlight the potential of AL for scalable annotation in audio and time-series applications, including bioacoustics.**

***Index Terms*—Active Learning, Sound Event Detection, Bioacoustics, Annotation Efficiency, Uncertainty Sampling, Biodiversity Monitoring**

I. Introduction

The growing biodiversity crisis demands scalable and efficient monitoring techniques to enhance conservation efforts. Bioacoustics, the study of sound in biological contexts, has emerged as a powerful tool for biodiversity monitoring, offering a non-invasive and cost-effective means to collect rich ecological data over large spatial and temporal scales [1], [2]. Analyzing animal vocalizations, environmental sounds, and anthropogenic noise from audio recordings can provide crucial insights into species presence, population dynamics, and ecosystem health. However, the vast amounts of audio data generated by bioacoustic monitoring programs create major challenges in data processing, annotation, and analysis.

Sound Event Detection (SED) plays a vital role in automating the analysis of bioacoustic data, aiming to identify and classify sound events of interest, such as animal vocalizations, within continuous audio streams [1]. Supervised deep learning models have achieved remarkable success in SED [3], yet their performance hinges on access to large, accurately labeled datasets. The manual annotation of bioacoustic recordings—marking sound events and assigning class labels—is costly, labor-intensive, and constitutes a major bottleneck in developing effective SED models.

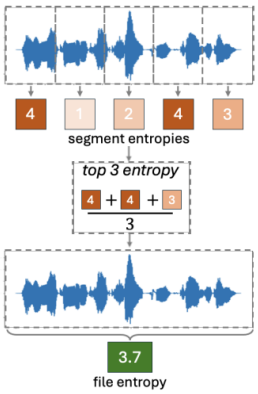


Fig. 1. Top K Entropy uncertainty aggregation selects the top K segment entropies (here K=3) obtained from the segments of the file. The resulting uncertainty for the file is the average of the selected segment entropies.

Active Learning (AL) is a promising strategy to mitigate the annotation bottleneck. By selecting the most informative data points for annotation, AL minimizes labeling effort while maintaining high model performance [4]. While extensively studied in image classification and natural language processing, AL for bioacoustic SED remains underexplored. Existing studies [5], [6] have demonstrated its potential but often focus on frame- or segment-based selection rather than the practical scenario where entire audio files must be annotated.

Traditional AL frameworks typically assume independent data points and construct query batches dynamically by combining uncertainty-based selection with diversification strategies [4], [7]–[9]. In contrast, AL for SED operates on segments of audio recordings, which are queried for annotation [5]. However, in practice, annotation is often performed at the file level, creating a mismatch between segment-based selection and real-world labeling processes. Instead of constructing query batches from independent segments, the challenge shifts to selecting entire recordings by ranking predefined batches of audio segments. This requires new uncertainty aggregation strategies that can effectively translate segment-level uncertainty into file-level scores, making them suitable for batch

§ Equal contribution.

selection at the recording level.

In this paper, we propose *Top K Entropy*, a novel uncertainty aggregation strategy for active learning in bioacoustic SED. Instead of treating all segments equally, this approach focuses on the most uncertain segments within each file (see figure 1). Our method is evaluated against other aggregation strategies using a bioacoustic soundscape dataset. The results show that *Top K Entropy* achieves competitive performance with up to 92% fewer labeled examples. This demonstrates its practical value for large-scale bioacoustic monitoring.

The rest of the paper is organized as follows: Section II details our active learning framework, including the SED model and the studied aggregation strategies. Section III presents experimental results, highlighting annotation efficiency and generalization. Section IV discusses the broader implications, and Section V concludes the paper.

II. Uncertainty Aggregation for Active Learning in SED

Our active learning framework is built around a pool-based approach, iteratively refining a segment-based SED model by strategically incorporating new batches of annotated data. Our methodology consists of three key components: (1) a segment-based SED model, (2) uncertainty aggregation strategies for selecting entire audio files for annotation, and (3) diversification techniques to improve batch diversity.

A. Segment-based Sound Event Detection Model

The SED model employed in this study is designed to process short segments in audio files, allowing fine-grained temporal resolution and efficient feature extraction. For feature extraction, we use the YAMNet architecture, a deep convolutional neural network pre-trained on a large-scale audio dataset called AudioSet [10]. YAMNet is known for its ability to capture robust and general-purpose audio representations, making it well suited for transfer learning in bioacoustic domains.

Specifically, each audio file is divided into segments of 0.12 seconds duration, with an overlap of 50%. For each segment, a 1024-dimensional feature embedding is extracted. These embeddings serve as input to a linear classifier that is trained to predict the class label for each segment. The classifier consists of a single fully connected neural network layer with 4 outputs corresponding to the classes: baby, dog, meerkat and background noise. A softmax activation function is applied to the output layer, providing probability estimates for each class. The linear classifier is trained using categorical cross-entropy loss and the ADAM optimizer, with hyperparameters tuned for optimal validation performance.

To refine the segment-level predictions and generate coherent sound events, a median filter of kernel size 3 is applied along the temporal dimension of the classifier output. This filtering step smooths the predictions, reduces the effects of transient noise, and enforces temporal consistency.

After filtering, adjacent segments with the same class prediction are merged into continuous sound events. Consequently, the final output of the SED model consists of a set of detected sound events, each characterized by a class label, start time, and end time.

B. Uncertainty Aggregation Strategies for File Querying

In our active learning framework, the querying process operates at file level, meaning that entire audio files are selected for annotation in each iteration. To bridge the gap between segment-level uncertainties and file-level queries, uncertainty aggregation strategies are employed. These strategies compute a file-level uncertainty score by aggregating segment-level entropies within each unlabeled audio file. The entropy of a segment is quantified by its Shannon entropy, defined as $E = -\sum_c p_c \log_2(p_c)$, where p_c represents the predicted probability for class c. We investigate and compare the following aggregation methods.

Top K Entropy is based on the premise that the most uncertain segments provide the most valuable information for improving model performance. It selects the top K highest segment entropies within a file and calculates their average. By focusing on the K most uncertain segments, this strategy aims to query files that contain a few highly ambiguous segments, potentially indicative of rare or challenging sound events. Through empirical evaluation, we found that $K = 10$ strikes a balance between selecting sufficiently uncertain segments while avoiding overrepresentation of minor fluctuations in entropy.

Along with this, the following *baseline strategies* were explored in the experiments.

- *Mean Entropy* is an aggregation strategy that calculates the average entropy across all segments within a file and uses this as representation of the file-level uncertainty. It provides a holistic measure of uncertainty, where all segments are considered equal.
- *Median Entropy* calculates the median entropy across all segments in a file.
- *Mean Event Entropy* uses the mean entropy of all segments which are predicted as events by the model.
- *Random Querying* chooses files at random, without considering entropy. This simulates a scenario where active learning is not applied, serving as a good reference point for all uncertainty aggregation strategies.

In each active learning iteration, all unlabeled audio files in the training set are ranked based on their computed uncertainty scores from the chosen aggregation strategy. Files with the highest uncertainty scores are considered more informative, and are queried for annotation in a batch of a predefined size. The annotator is simulated in these experiments by simply revealing the ground truth labels for the queried audio recordings.

III. Experimental Evaluation and Results

A. Dataset Generation

Experiments were conducted using bioacoustic soundscapes to evaluate the performance of different AL strategies. The datasets provides a controlled and reproducible experimental environment, while closely resembling real-world bioacoustic

data. The datasets were generated by mixing foreground sound events with background noise sourced from recordings of park environments [11]. The foreground events consisted of vocalizations of three classes: baby cries, [12], dog barks [12], and meerkat calls [13]. These classes were chosen to represent diverse sound characteristics and event durations.

datasets were generated with varying characteristics, to assess the robustness of our method under different data conditions. Each dataset was generated using 2 parameters; event ratio r, describing the ratio of files containing events, and SNR. Files with events always contain one to three events (chosen uniformly), with equal probability of each class.

The main dataset used was created with an SNR of 0 and $r = 0.2$, meaning that 20% of the audio files contained sound events, while the remaining 80% consisted solely of background noise. In order to determine if results depend on SNR, we also generated datasets with SNR values of 10dB and -10dB. A dataset with $r = 1.0$, where all files contained at least one event, was also created and evaluated. Each of the mentioned datasets comprised of 2500 audio files, with a fixed duration of 10 seconds per file (equivalent of 175 segments). For each dataset, 80% of the data was generated for training and 20% was generated for testing, where the event ratio was kept within each set division.

B. Experimental Setup

Active learning was simulated using a pool-based setup. We initialized the AL loop with a randomly selected seed set comprising 0.5% of the training data and iteratively queried additional unlabeled files for annotation. The batch sizes were dynamically adjusted throughout the learning process, starting with smaller batches in the early iterations to capture the rapid performance gains of AL, and gradually increasing batch sizes in later iterations. We compared the performance of *Top K Entropy* with different baseline strategies (*Mean Entropy, Median Entropy, Mean Event Entropy* and *Random Querying*). For each strategy, experiments were repeated for 5 or more random seeds to account for variability in initialization and data sampling. Model performance is evaluated using the Intersection over Union (IoU) metric, computed over the entire test set by concatenating data from all testing files. We refer to this metric as total IoU.

C. Comparative Analysis of Aggregation Strategies

Figure 2 presents a comparison of the total IoU performance for the explored aggregation strategies. Our results indicate that *Top K Entropy* achieves a 92% reduction in annotation effort, outperforming *Random Querying* across all annotation budgets, thus demonstrating the effectiveness of uncertainty-based selection. *Top 10 Entropy* outperform the baselines across all annotation budgets, achieving a total IoU comparable to the fully supervised model (trained on 100% labeled data) with only 8% of the training data annotated. This translates to a 92% reduction in annotation effort. *Mean Entropy* also managed to achieve results comparable to a fully supervised model in the early iterations.

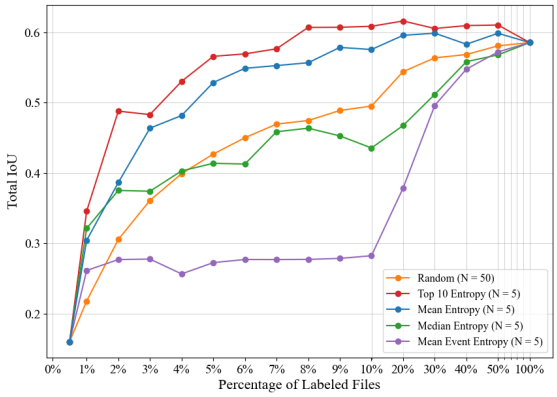


Fig. 2. Total IoU performance of different aggregation strategies and the Random Querying baseline, averaged over 5 seeds. *Top K Entropy* achieves comparable IoU to a fully supervised model while reducing annotation effort by 92%. Results based on data generated with event ratio $r = 0.2$ and SNR $= 0$.

Median Entropy does not provide a substantial improvement in annotation efficiency compared to the random baseline. This is not surprising, considering that the majority of segments in each file purely consist of background noise, where the model is both certain and accurate. *Mean Event Entropy* is significantly worse than the random baseline.

The superior performance of *Top 10 Entropy* and *Mean Entropy* suggests that uncertainty-based querying effectively identifies informative audio files for annotation, leading to higher annotation efficiency.

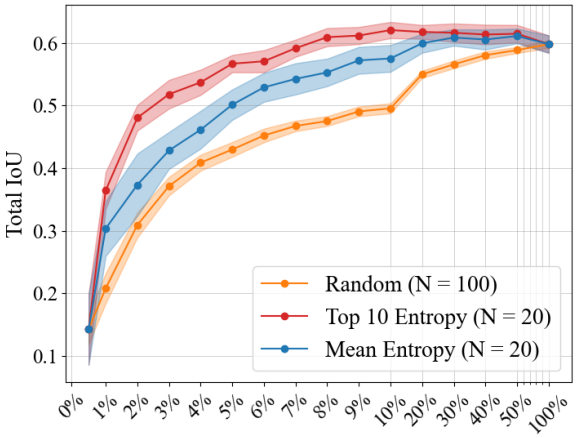


Fig. 3. Total IoU results averaged over 20 seeds (N=20) for *Mean Entropy* and *Top 10 Entropy*. The results for Random Querying baseline are averaged over 20 seeds, with five different initializations for the dataset. Each line is paired with 95% confidence intervals. Results based on data generated with event ratio $r = 0.2$ and SNR $= 0$.

In figure 3, *Top 10 Entropy* and *Mean Entropy* are further compared to *Random Querying*, with 95% confidence intervals. This experiment was run for 20 different seeds to get better approximations of average performance and variance. These results clearly show the superiority of *Top 10 Entropy*. The effectiveness of *Top K Entropy* highlights

the importance of focusing on the most ambiguous segments within audio files, as these segments likely contain novel or challenging sound events that contribute most significantly to model learning. Given that *Top K Entropy* outperforms *Mean Entropy*, it suggests that querying files with the highest peaks in uncertainty is more effective than querying those with the highest average uncertainty.

Diversification strategies were evaluated (Farthest Traversal [5], and Random Selection), and they improved the results for the baseline strategies slightly. The *Top K Entropy* strategy, on the other hand, appears to inherently select diverse batches, as its performance remained unaffected by additional diversification techniques. The *Top K Entropy* strategy consistently had the best scores across all experiments.

D. Exploring Top K Entropy

K was set to 10 in the initial testing of the *Top K Entropy* strategy. In figure 4 a comparison between different values of K is presented.
All tested values of K outperform *Random Querying*, but differences between them are hard to distinguish. Smaller values of K seem to perform slightly better for low annotation budgets. These results show that the *Top K Entropy* is robust with respect to the selection of K. Previously, in figure 3, we observed that the maximum value of $K = 175$ (which corresponds to *Mean Entropy*) performs worse compared to $K = 10$. This suggests that there is an upper limit for a suitable value of K.

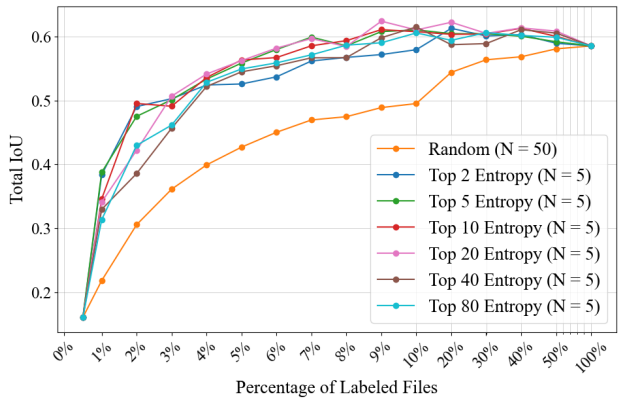


Fig. 4. Total IoU performance averaged over 5 seeds (N=5), for 6 different fixed values of K for the querying strategy *Top K Entropy*) along with the baseline strategy. Results based on data generated with event ratio $r = 0.2$ and SNR $= 0$.

E. Generalization - SNR

The sensitivity of models to varying SNR was evaluated using SNR $= 10$ and SNR $= -10$, showing an improvement and a degradation in performance, respectively. In both cases, the active learning framework is more annotation efficient compared to the baseline. For SNR $= -10$ the relative improvement from the baseline is much higher, compared to when the data is mixed with a higher SNR. This suggests that the relative gain of our approach is bigger given a more challenging SED task. These results also showcase robustness to noise power variability, beneficial for real-world applications where the SNR will fluctuate.

F. Generalization - Event Ratio

To evaluate the impact that event ratio has on active learning performance, a dataset consisting solely of files containing events was used. The *Top 10 Entropy* strategy consistently outperformed the baseline, indicating robustness to changes in event frequency. However, in this domain the *Mean Entropy* strategy did not outperform the random baseline, suggesting it may be better suited for datasets with fewer events. Overall, the active learning strategies demonstrated greater effectiveness when fewer events were present in the data, suggesting that active learning is most beneficial when events are scarce.

IV. Discussion

Our results demonstrate the effectiveness of the *Top K Entropy* aggregation strategy for active learning for bioacoustic SED. The strategy consistently demonstrates superior performance, suggesting that focusing on the most uncertain segments within audio files is crucial for efficient learning. The significant reduction in labeled data required to achieve high performance underscores the practical value of the proposed strategy in this domain.

The analysis of queried files reveals that successful AL strategies tend to address class imbalance and prioritize files with sound events. While diversification techniques showed some potential for further improvement, particularly Farthest Traversal with AudioMAE [14] embeddings, the gains were modest in our experiments. This could be due to the effectiveness of *Top K Entropy* in already selecting a relatively diverse set of informative files. Further investigation into more sophisticated diversification strategies and their interaction with different uncertainty aggregation methods is warranted.

The generalization results across SNR levels further underscore the robustness and practical applicability of the proposed approach. The baseline approaches suffer more from a lower SNR than *Top K Entropy*, suggesting that the approach is robust and stable in different environments. This is important for the deployment of bioacoustic SED systems in diverse ecological settings and under varying environmental conditions.

The generalization results for the dataset containing only files with events further highlights the potential of *Top K Entropy*. As the event ratio increases, the reward for identifying files containing events diminishes, while prioritizing files with a higher number of events or greater event variability becomes more important. The results indicate that *Top K Entropy* adapted better to this change in data distribution than the other baselines.

The success of *Top K Entropy* suggests that prioritizing the most uncertain segments within a file provides a better representation of file-level uncertainty than traditional averaging methods. This aligns with the broader active learning literature, which highlights that focusing on high-entropy regions can

accelerate model convergence. Future studies should explore whether this principle extends to other time-series domains, such as speech recognition and medical diagnostics.

Limitations of this study include the use of soundscapes mixed with known sound event classes, which, while designed to mimic real-world data, may not fully capture the complexities of natural bioacoustic recordings. Future research should validate our findings using real-world bioacoustic datasets and explore how self-supervised learning methods, such as AudioMAE or wav2vec, can enhance uncertainty estimation in active learning frameworks. Furthermore, our study does not fully examine the impact of class imbalance on *Top K Entropy's* effectiveness. If rare events consistently exhibit high entropy, they may be overrepresented in queries, potentially biasing the training process. Future research should investigate adaptive sampling strategies that balance uncertainty with class distribution awareness.

V. Conclusions

This research provides an evaluation of uncertainty aggregation strategies in active learning for efficient annotation in bioacoustic sound event detection. Our findings demonstrate that uncertainty-based AL, and particularly the *Top K Entropy* aggregation strategy, offers a powerful approach to drastically reduce annotation effort while maintaining high SED performance. By prioritizing the most uncertain segments, the *Top K Entropy* strategy reduces annotation requirements by up to 92% while maintaining model performance, demonstrating its potential for scalable bioacoustic monitoring. The robustness of the proposed approach across SNR variations and varying event ratios further underscores its practical applicability for real-world bioacoustic monitoring.

Future research directions include exploring dynamic adaptation of the K parameter in *Top K Entropy* to optimize performance across different data distributions and annotation budgets. Investigating more advanced diversification techniques and their interplay with uncertainty aggregation methods could also lead to further improvements in annotation efficiency. Furthermore, evaluating AL strategies on diverse real-world bioacoustic datasets and exploring the integration of AL with citizen science initiatives could pave the way for scalable and cost-effective biodiversity monitoring solutions. By significantly reducing the annotation bottleneck, active learning has the potential to democratize bioacoustic data analysis, empowering researchers and conservation practitioners to leverage the vast amounts of acoustic information for effective biodiversity monitoring and conservation action.

VI. Acknowledgements

This work was supported by Swedish Foundation for Strategic Research, FID20-0028.

References

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, no. 2017, pp. 1–32, 2022. [Online]. Available: http://arxiv.org/abs/2112.06725

[2] E. Browning, R. Gibb, P. Glover-Kapfer, and K. Jones, "Passive acoustic monitoring in ecology and conservation. technical report," Tech. Rep., 2017.

[3] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[4] B. Settles, "Active learning literature survey," Tech. Rep., 2009.

[5] Z. Shuyang, T. Heittola, and T. Virtanen, "Active learning for sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2895–2905, 2020.

[6] J. M. van Osta, B. Dreis, E. Meyer, L. F. Grogan, and J. G. Castley, "An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the southern black-throated finch (poephila cincta cincta)," *Ecological Informatics*, vol. 77, p. 102233, 2023.

[7] A. Kirsch, J. van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," in *Proc. NeurIPS*, 2019, pp. 7024–7035.

[8] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *Proc. ICLR*, 2020.

[9] P. Ren, Y. Xiao, K. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, K. Chen, and K. Wang, "A survey of deep active learning," *ACM Computing Surveys*, vol. 54, no. 9, p. Article 180, 2021.

[10] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017.

[11] A. Diment, A. Mesaros, T. Heittola, and T. Virtanen, "TUT Rare sound events, Development dataset," Mar. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.401395

[12] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "NIGENS general sound events database," Feb. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2535878

[13] I. Nolasco, "DCASE 2022 Task 5: Few-shot Bioacoustic Event Detection Evaluation Set," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6517414

[14] P.-Y. B. Huang, H. Xu, J. B. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *ArXiv*, vol. abs/2207.06405, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250491011

From Weak to Strong Sound Event Labels using Adaptive Change-Point Detection and Active Learning

1st John Martinsson
Computer Science
Research Institutes of Sweden
Gothenburg, Sweden
john.martinsson@ri.se

2nd Olof Mogren
Computer Science
Research Institutes of Sweden
Gothenburg, Sweden
olof.mogren@ri.se

3rd Maria Sandsten
Centre for Math. Sciences
Lund University
Lund, Sweden
maria.sandsten@matstat.lu.se

4th Tuomas Virtanen
Signal Proc. Research Centre
Tampere University
Tampere, Finland
tuomas.virtanen@tuni.fi

***Abstract*—We propose an adaptive change point detection method (A-CPD) for machine guided weak label annotation of audio recording segments. The goal is to maximize the amount of information gained about the temporal activations of the target sounds. For each unlabeled audio recording, we use a prediction model to derive a probability curve used to guide annotation. The prediction model is initially pre-trained on available annotated sound event data with classes that are disjoint from the classes in the unlabeled dataset. The prediction model then gradually adapts to the annotations provided by the annotator in an active learning loop. We derive query segments to guide the weak label annotator towards strong labels, using change point detection on these probabilities. We show that it is possible to derive strong labels of high quality with a limited annotation budget, and show favorable results for A-CPD when compared to two baseline query segment strategies.**

***Index Terms*—Active learning, annotation, sound event detection, deep learning**

I. Introduction

Most audio datasets today consists of weakly labeled data with imprecise timing information [1], and there is a need for efficient and reliable annotation processes to acquire labels with precise timing information. We refer to such labels as strong labels. The performance of sound event detection (SED) models improve with strong labels [2], and strong labels become especially important when we want to count the number of occurrences of an event class. For example in bioacoustics, where counting the number of vocalizations of an animal species can be used to estimate population density and draw ecological insights [3].

Crowdsourcing the strong labels is challenging and an attractive solution is to crowdsource weak labels to enable reconstruction of the strong labels [4], [5]. Asking the annotator for strong labels requires more work and it can in the worst case lead to the annotator misunderstanding the task [5].

Disagreement-based active learning is the most used form of active learning for sound event detection [6]–[9], focusing on selecting what audio segment to label next. The recordings are either split into equal length audio segments [6], [7], [9] or segments depending on the structure of the sound [8]. Each segment is then given a weak label by the annotator.

This work was supported by The Swedish Foundation for Strategic Research (SSF; FID20-0028) and Sweden's Innovation Agency (2023-01486).

https://github.com/johnmartinsson/adaptive-change-point-detection

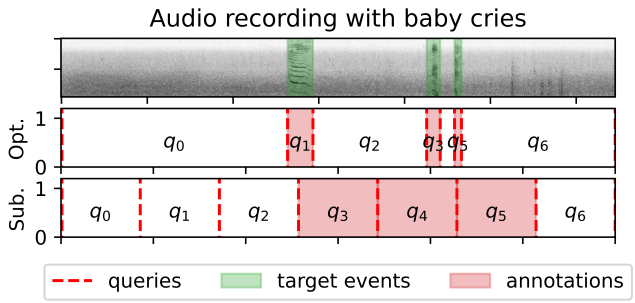


Fig. 1. Illustration of segmentation of an audio spectrogram with three target events shown in shaded green (top panel) into a set of audio query segments $q_0, \ldots, q_6$ using an optimal method w.r.t the derived strong label timings (middle panel) and a sub-optimal method (bottom panel). Resulting annotations, from the weak labels given by the annotator, are shown in shaded red for both methods. Query q_4 for the optimal method is omitted for clarity.

We use a weak label annotator to derive strong labels as in [5], but instead of using fixed length query segments we adapt the query segments to the data, in the setting of active learning. We propose an adaptive change point detection (A-CPD) method which splits a given audio recording into a set of audio segments, or queries. The queries are then labeled by the annotator and the strong labels are derived and evaluated. See Fig. 1 for an illustration where a set of seven queries are used either optimally or sub-optimally for a given audio recording with three sound events. We assume three sound events to be detected in each audio recording as a simplification during method development. We aim to adapt the set of queries in such a way that the information about the temporal activations of the target sounds is maximized. Note that we aim to actively guide the annotator during the annotation of the audio recordings, rather than actively choose which audio recordings to annotate which is typically done in active learning.

II. Sound event annotation using active learning

We consider SED tasks where the goal is to predict the presence of a given target event class. The results can also be generalized into the multi-class setting. Given a restricted annotation budget and no initial labels we aim to derive strong labels using active learning to train a SED system. To this end, we propose the following machine guided annotation process.

Let $\mathcal{D}_L^{(k)}$ denote the set of labeled audio recordings and $\mathcal{D}_U^{(k)}$ the set of unlabeled audio recordings at active learning iteration k. Further, let $\mathcal{A}^{(k)} = \{(s_i^{(j)}, e_i^{(j)}, c_i^{(j)})\}_{i=1\,j=1}^{B\quad k}$ denote the annotations of segments, where s denotes the onset, e the offset, and $c \in \{0, 1\}$, the weak label for each segment i of the B annotated segments in audio recording j.

We start without any labels, $\mathcal{A}^{(0)} = \mathcal{D}_L^{(0)} = \emptyset$, and all audio recordings are unlabeled, $\mathcal{D}_U^{(0)} = \{\mathbf{x}_j\}_{j=1}^N$, where $\mathbf{x}_j \in \mathbb{R}^T$ denotes an audio recording of length T, and N denotes the total number of audio recordings. We then loop for each $k \in \{1, \ldots, N\}$ and:

1) choose a random unlabeled audio recording $\mathbf{x}$ from $\mathcal{D}_U^{(k-1)}$,
2) derive a set of B audio query segments $Q = \{q_i\}_{i=0}^{B-1}$ using a query strategy where $q_i = (s_i, e_i)$ consists of the start s_i and end e_i timings for query i,
3) send the queries to the annotator (returning a weak label for each query) and add the annotations to the set of segment labels $\mathcal{A}^{(k)} = \mathcal{A}^{(k-1)} \cup \{(s_i, e_i, c_i)\}_{i=1}^B$,
4) *In case of A-CPD*: use the annotations $\{(s_i, e_i, c_i)\}_{i=1}^B$ to update the query strategy, and
5) update the labeled recording set $\mathcal{D}_L^{(k)}$ by adding $\mathbf{x}$ and the unlabeled recording set $\mathcal{D}_U^{(k)}$ by removing $\mathbf{x}$.

For brevity we have omitted the dependence on k for $\mathbf{x}_{r_k}$ and $(s_i^{(r_k)}, e_i^{(r_k)}, c_i^{(r_k)})$ in the description of the annotation loop, where $r_k \in \{1, \ldots, N\}$ would denote the randomly sampled audio recording for iteration k. After the annotation loop all N audio recordings have been annotated exactly once with the query method used in step (2), resulting in a set of annotations $\mathcal{A}^{(N)} = \{(s_i^{(j)}, e_i^{(j)}, c_i^{(j)})\}_{i=1\,j=1}^{B\quad N}$.

Note that B is not the number of sound events in the recording, but the number of query segments allowed when annotating the recording. The smallest number of query segments to derive the ground truth strong labels does, however, depend on the number of sound events M in the recording as $2M + 1$ (see Section III-D). A-CPD is developed to provide strong labels using as few as $B = 2M + 1$ queries.

The total annotation budget used will scale with both N and B. Typically we would aim to reduce N by actively sampling the data points to annotate, but we instead aim to reduce B. Think of B as a part of the annotation cost of an audio recording, which can be reduced with maintained label strength by guiding the annotator during the annotation process.

III. Query strategies

In this section we describe the studied query strategies.

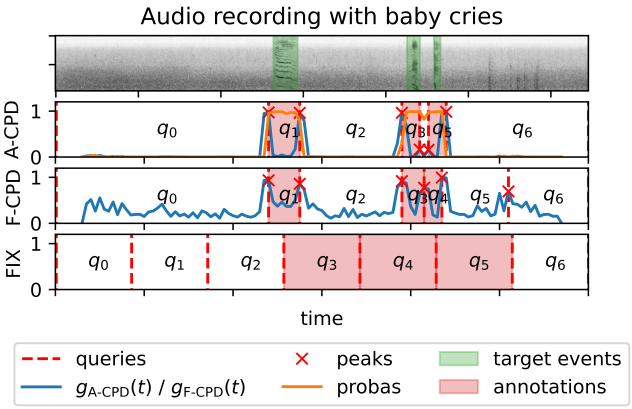


Fig. 2. Qualitative example of how the different query strategies A-CPD, F-CPD and FIX segment a spectrogram of an audio recording with three target events shown in shaded green (top panel) into $B = 7$ queries. A-CPD (second panel) uses change point detection (blue line) on the probability curve from a prediction model (orange line) to detect the $B - 1$ most prominent peaks (red crosses) which are used to construct a set of queries $\{q_0, \ldots, q_{B-1}\}$ (dashed red lines). Each query $q_i = (s_i, e_i)$ is given a weak label $c_i \in \{0, 1\}$ ($c = 1$ shown as shaded red), resulting in the i:th annotation (s_i, e_i, c_i). F-CPD (third panel) uses change point detection directly on the cosine distances in embedding space (blue line) and thereafter constructs queries in the same way as A-CPD. FIX (fourth panel) uses fixed length queries.

A. The adaptive change point detection strategy (A-CPD)

To produce a set of queries for a given audio recording $\mathbf{x}$ at annotation round k we perform three key steps:

1) update a prediction model using the annotations from round $k - 1$ (initialized with pre-training if $k = 0$),
2) predict probabilities indicating the presence of the target class in the recording using the model, and
3) apply change point detection to the probabilities to derive the queries.

The pre-training of the prediction model can be done in a supervised or unsupervised way. The important property is that the model reacts to changes in the audio recording related to the presence or absence of the target class. However, it is not strictly necessary that the model reacts *only* to those changes.

Let $h_k : \mathbb{R}^L \to [0, 1]$ denote a model that predicts the probability of an audio segment of length L belonging to the target event class. In principle, any prediction model can be used. For a given audio recording $\mathbf{x}$ the prediction model $h_k(\cdot)$ is applied to consecutive audio segments to derive a probability curve, shown as the orange curve for A-CPD in Fig. 2. The consecutive audio segments are derived using a moving window of L seconds with hop size $L/4$.

We define the Euclidean distance between two points $t - \alpha$ and $t + \alpha$ on the probability curve as:

$$g_{\text{A-CPD}}^{(k)}(t) = ||h_k(t - \alpha) - h_k(t + \alpha)||, \tag{1}$$

shown as the blue curve for A-CPD in Fig. 2. The previous probability is compared with the next probability in Eq. 1, and $\alpha = L/4$ (hop size) is therefore chosen to ensure a 50% overlap between the audio segments for these probabilities.

Let t be a local optimum of $g^{(k)}_{\text{A-CPD}}(t)$, and all such local optima are called peaks. We rank peaks based on *prominence*. For any given peak t, let t_l and t_r denote the closest local minima of $g_k(\cdot)$ to the left and right of t. The prominence of the peak at t is defined as $|g_k(t) - \max(g_k(t_l), g_k(t_r))|$. Let $\mathcal{T}_{\text{A-CPD}} = \{t_1, t_2, \ldots, t_{B-1}\}$ be the $B-1$ most prominent peaks of a given audio recording such that $t_1 \leq t_2 \leq \cdots \leq t_{B-1}$, shown as red crosses in Fig. 2. The A-CPD query method is then defined as:

$$Q^{(k)}_{\text{A-CPD}} = \{(0, t_1), (t_1, t_2), \ldots, (t_{B-1}, T)\}, \quad (2)$$

which are shown as dashed red lines in Fig. 2, where T is the length of the audio recording and B is the number of queries used. Note that $g^{(k)}_{\text{A-CPD}}(t)$ will gradually become more sensitive towards changes between presence and absence of the target class in the recording with additional annotations, and become less sensitive to other unrelated changes.

B. The fixed change point detection strategy (F-CPD)

The fixed change point detection (F-CPD) method used as a reference derives the queries by computing the cosine distance between the previous embedding at time $t - \alpha$ and the next embedding at time $t + \alpha$:

$$g_{\text{F-CPD}}(t) = 1 - \frac{\mathbf{e}_{t-\alpha} \cdot \mathbf{e}_{t+\alpha}}{||\mathbf{e}_{t-\alpha}||||\mathbf{e}_{t+\alpha}||}, \quad (3)$$

where $\mathbf{e}_t = f_\theta(\mathbf{x}_t)$ denotes the embedding of consecutive audio segments $\mathbf{x}_t$ centered at second t using the embedding function $f_\theta : \mathbb{R}^L \rightarrow \mathbb{R}^K$. The cosine distance curve for an audio recording is shown as the blue line for F-CPD in Fig. 2. This method is similar to [8] except that embeddings are derived for 1.0 seconds of audio instead of 0.02. We therefore directly compare the previous and next embeddings instead of a moving average as in [8].

The most prominent peaks in the cosine distance curve is then selected, $\mathcal{T}_{\text{FIX}} = \{t_1, t_2, \ldots, t_{B-1}\}$, and the set of queries are defined as in Eq. 2, shown as dashed red lines for F-CPD in Fig 2.

C. The fixed length strategy (FIX)

In the fixed length query strategy (FIX) audio is split into equal length segments and then labeled. Let $d = T/B$, then the queries are defined as

$$Q_{\text{FIX}} = \{(0d, 1d), (1d, 2d), \ldots, ((B-1)d, Bd)\}, \quad (4)$$

shown as dashed red lines for FIX in Fig 2. This is the setting most previous active learning work for SED consider.

D. The oracle strategy (ORC)

The oracle query strategy constructs the queries based on the ground truth presence and absence annotations

$$Q_{\text{ORC}} = \{(s_0, e_0), (s_1, e_1), \ldots, (s_{B_{\text{suff}}-1}, e_{B_{\text{suff}}-1})\}, \quad (5)$$

where (s_i, e_i) is the onset and offset for segment i where the target event is either present or not. B_{suff} is the sufficient number of queries to get the true strong labels, which relate to the number of target events M in the given audio recording by $B_{\text{suff}} = 2M + 1$. ORC is undefined for $B < B_{\text{suff}}$.

E. The role of query strategies in the annotation process

The query strategies described in this section are then used in step (2) of the annotation loop described in Section II. Note that when the queries are not adapted to the audio recording multiple events can end up being counted as one. In Fig. 2 we can see this for F-CPD where q_3 and q_4 are directly adjacent, meaning that they are not resolved as two separate events, and for FIX where q_3, q_4 and q_5 are all directly adjacent. A-CPD often resolves all three events. Fig 2 is a qualitative example of all three methods, and quantitative results to further support this claim are provided later in table I.

The FIX length query segments depend on the query timings and target event timings aligning by chance since the query construction is independent of the target events. The A-CPD method aim to create query segments that are aligned with the target events by construction. In addition, the number of queries needed to derive the strong labels scale with the number of target events in the recording for A-CPD, which can be beneficial.

IV. Evaluation

A. Datasets

We create three SED datasets for evaluation, each with a different target event class: Meerkat, Dog or Baby cry. The Meerkat sounds are from the DCASE 2023 few-shot bioacoustic SED dataset [10] and the Dog and Baby cry sounds from the NIGENS dataset [11]. The sounds used for absence of an event are from the 15 background types in the TUT Rare sound events dataset [12].

The audio recordings in each dataset are created by randomly selecting $M = 3$ sound events from that event class and mixing them together with a randomly selected background recording of length $T = 30$ seconds. In this way we know that exactly $B_{\text{suff}} = 2M + 1 = 7$ queries are *sufficient* and *necessary* to derive the ground truth strong labels using a weak label annotator. The mixing is done using Scaper [13] at an SNR of 0 dB. In total we generate $N = 300$ audio recordings using this procedure for each event class as training data and equally many as test data.

The source files used in the mixing uses the supplied splits in [11] and [12], except for the Meerkat sounds where non exist and the split is done on a recording level.

B. Evaluation metrics

We evaluate the methods by annotating the mixed training datasets using the query strategies described in Section II and the annotation loop described in Section III. The quality of the annotations are then measured in two ways: (i) how strong the annotations are compared to the ground truth, and (ii) the test time performance of two evaluation models trained using the different annotations.

The evaluation metrics used in case (i) and (ii) are event-based F_1-score (F_{1e}) and segment-based F_1-score (F_{1s}) [14]. The segment size for F_{1s} is set to 0.05 seconds, and the collar for F_{1e} is set to 0.5 seconds. In case (i) the F_{1s} measures how much of the audio that has been correctly labeled and in

TABLE I
AVERAGE F_{1s}-SCORE AND F_{1e}-SCORE FOR THE TRAINING ANNOTATIONS FOR EACH ANNOTATION PROCESS AND TARGET EVENT CLASS WITH $\beta = 0$

Strategy	Meerkat		Dog		Baby	
	F_{1s}	F_{1e}	F_{1s}	F_{1e}	F_{1s}	F_{1e}
ORC	1.00	1.00	1.00	1.00	1.00	1.00
A-CPD	**0.31**	**0.57**	**0.29**	**0.45**	**0.62**	**0.60**
F-CPD	0.16	0.44	0.21	0.30	0.48	0.45
FIX	0.11	0.00	0.19	0.00	0.41	0.01

case (ii) F_{1s} measures how much of the audio that has been correctly predicted by the evaluation model. The F_{1e} score is only used to measure how close the annotations are to the ground truth labels in the training data.

a) Annotator model: Let $\mathcal{A}_{gt}^{(j)} = \{(s_i, e_i, c = 1)\}_{i=1}^{3}$ denote the set of ground truth target event labels for audio recording j, where s_i is the onset, e_i the offset and $c = 1$ indicate the presence of the target event.

We use $\mathcal{A}_{gt}^{(j)}$ to simulate an annotator for recording j. For a given query segment we check the overlap ratio with the ground truth target event labels. Formally, if there exists an annotation $(s_i, e_i, c_i = 1)$ s.t.

$$\frac{(s_i, e_i) \cap (s_q, e_q)}{|s_i - e_i|} \geq \gamma, \tag{6}$$

holds for the given query segment $q = (s_q, e_q)$, then the annotator returns $c_i = 1$ for query q, and $c_i = 0$ otherwise. Annotation noise is added by flipping the returned label with probability β. In this work $\gamma = 0.5$, and $\beta \in \{0.0, 0.2\}$.

C. Implementation details and experiment setup

a) Prediction model: The prediction model $h_k(\cdot)$ is modeled using a prototypical neural network (ProtoNet) [15]. The prototypes are easily updated at each annotation round k using a running average between each previous prototype and the newly labeled audio embeddings. We model the embedding function $f_\theta(\cdot)$ using BirdNET [16], a convolutional neural network pre-trained on large amounts of bird sounds.

b) Evaluation models: We use two models to evaluate the test time performance of models trained on the annotations obtained using each query strategy: a two layer multilayer perceptron (MLP) and a ProtoNet. The MLP is trained using the Adam optimizer and cross-entropy loss. Each query strategy is run 10 times and the evaluation models are trained on the embeddings using the resulting labeled datasets. ProtoNet is used in two ways: as a prediction model in the proposed A-CPD method, and as an evaluation model.

D. Results

In Table I we show the average F_{1s}-score and F_{1e}-score for the training data annotations over 10 runs for each dataset and with the sufficient nu $B = B_{\text{suff}} = 7$. The A-CPD method outperforms the other methods for all studied target event classes. The standard deviation is in all cases less than 0.03 (omitted from table for brevity), and the baseline query strategies are deterministic when $\beta = 0$.

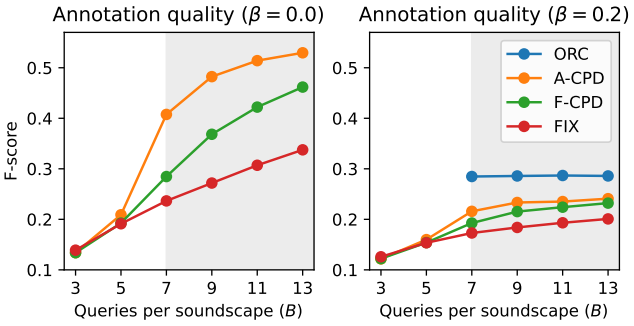


Fig. 3. The average F_{1s}-score over the three classes for each of the studied annotation processes plotted against the number of queries per audio recording, B. The results are shown for an annotator without noise (left) and with $\beta = 0.2$ (right). Note that ORC is 1.0 when $\beta = 0$ and is therefore not shown in the left figure. Shaded region where $B \geq B_{\text{suff}}$.

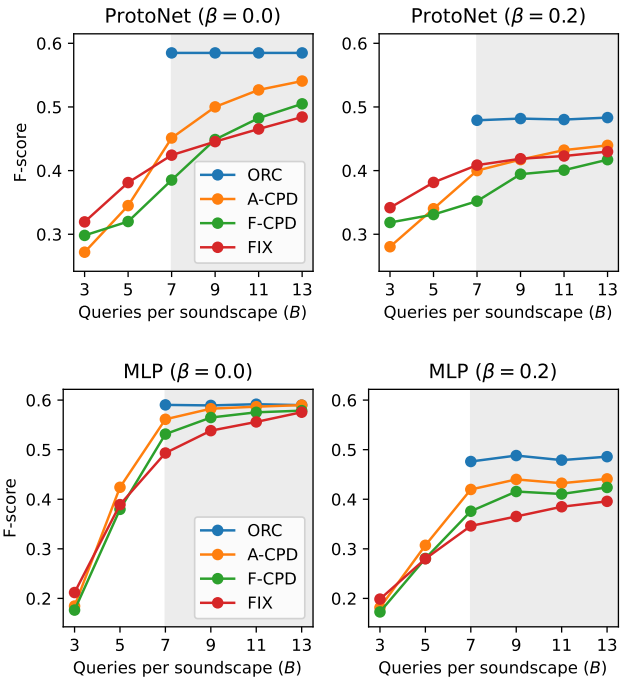


Fig. 4. The average test time F_{1s}-score over the studied sound classes for a ProtoNet (top) and the MLP (bottom) trained with the annotations from each respective annotation process and setting. Shaded region where $B \geq B_{\text{suff}}$.

In Fig. 3 we show the average F_{1s}-score over all runs and event classes for the annotations derived from each query strategy. The proposed A-CPD method has a strictly higher F_{1s}-score than the FIX and F-CPD baselines for all budgets and noise settings. We also see that there is still a significant gap to the ORC strategy. The noisy annotator ($\beta = 0.2$) drastically reduce the label quality for all studied strategies, especially ORC dropping from an F_{1s}-score of 1.0 (omitted from figure) to ≈ 0.28 (large drop due to class-imbalance).

In Fig. 4 we show the average test time F_{1s}-score of

TABLE II
AVERAGE TEST TIME F_{1s}-SCORE FOR PROTONET WITH $\beta = 0$.

Strategy	Meerkat	Dog	Baby
ORC	0.46	0.48	0.81
A-CPD	$\mathbf{0.44} \pm 0.00$	0.20 ± 0.01	$\mathbf{0.71} \pm 0.02$
F-CPD	0.31	0.19	0.66
FIX	0.34	**0.25**	0.68

TABLE III
AVERAGE TEST TIME F_{1s}-SCORE FOR MLP WITH $\beta = 0$.

Strategy	Meerkat	Dog	Baby
ORC	0.43 ± 0.00	0.51 ± 0.01	0.83 ± 0.00
A-CPD	$\mathbf{0.44} \pm 0.00$	$\mathbf{0.43} \pm 0.02$	$\mathbf{0.81} \pm 0.01$
F-CPD	0.38 ± 0.01	0.42 ± 0.02	0.79 ± 0.01
FIX	0.33 ± 0.02	0.40 ± 0.02	0.75 ± 0.02

a ProtoNet (top) and a MLP (bottom) trained using the annotations from each of the studied annotation strategies and settings. The A-CPD method outperforms the other methods when $B \geq 7$. For the ProtoNet the FIX method outperform A-CPD when $B < 7$ and for the MLP the results are similar.

Table II and III show the average F_{1s}-score and standard deviation for the three different event classes for all studied query strategies. The average is over 10 runs, and the number of queries is set to $B = 7$. Table II shows the F_{1s}-score for the ProtoNet evaluation model. A-CPD achieves a higher F_{1s}-score for the meerkat and baby datasets. On average A-CPD outperforms the other methods as seen in Fig. 4. Table III shows the F_{1s}-score for the MLP evaluation model. A-CPD achieves a higher F_{1s}-score for all studied datasets.

E. Discussion

The results in all tables are for the sufficient budget $B = B_{\text{suff}} = 2M + 1$. In practice we do not know B_{suff}. However, the A-CPD method is applicable also for an arbitrary number of sound events in the recording when B is chosen sufficiently large. This choice need to be made for all the studied methods. We show the benefit of A-CPD for differently chosen B in Fig. 3. Estimating B_{suff} based on the audio recording could further reduce the number of queries used and is left as future work.

We chose $\gamma = 0.5$ in the annotator model since the annotator should be able to detect a target event if more than 50% of the event occurs within the query segment. This choice is however non-trivial, and depends on the expertise of the annotator and target class among others. We observe similar results on average as those presented in the paper for $\gamma \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$ (not shown).

We use BirdNET [16] to model the embedding function since we study bioacoustic target classes. However, an embedding function such as PANNs [17] may also be used if the target classes are more general.

V. CONCLUSIONS

We have presented a query strategy based on adaptive change point detection (A-CPD) which derive strong labels of high quality from a weak label annotator in an active learning setting. We show that A-CPD gives strictly stronger labels than all other studied baseline query strategies for all studied budget constraints and annotator noise settings. We also show that models trained using annotations from A-CPD tend to outperform models trained with the weaker labels from the baselines at test time. We note that the gap to the oracle method is still large, leaving room for improvements in future work.

REFERENCES

[1] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[2] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 366–370, 2021.

[3] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack, "Estimating animal population density using passive acoustics," *Biological Reviews*, vol. 88, no. 2, pp. 287–309, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12001

[4] I. Martin-Morato, M. Harju, and A. Mesaros, "Crowdsourcing Strong Labels for Sound Event Detection," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 246–250, 2021.

[5] I. Martin-Morato and A. Mesaros, "Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 902–914, 2023.

[6] Z. Shuyang, T. Heittola, and T. Virtanen, "Active learning for sound event classification by clustering unlabeled data," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 751–755, 2017.

[7] ——, "An active learning method using clustering and committee-based sample selection for sound event classification," *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018 - Proceedings*, pp. 116–120, 2018.

[8] ——, "Active Learning for Sound Event Detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 2895–2905, 2020.

[9] Y. Wang, M. Cartwright, and J. P. Bello, "Active Few-Shot Learning for Sound Event Detection," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1551–1555, 2022.

[10] I. Nolasco, S. Singh, E. Vidana-Villa, E. Grout, J. Morford, M. Emmerson, F. Jensens, H. Whitehead, I. Kiskin, A. Strandburg-Peshkin, L. Gill, H. Pamula, V. Lostanlen, V. Morfi, and D. Stowell, "Few-shot bioacoustic event detection at the DCASE 2022 challenge," no. November, pp. 1–5, 2022.

[11] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS General Sound Events Database," Technische Universität Berlin, Tech. Rep., 2020, arXiv:1902.08314 [cs.SD].

[12] A. Diment, A. Mesaros, T. Heittola, and T. Virtanen, "TUT Rare sound events, Development dataset," Jan. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.401395

[13] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 344–348, 2017.

[14] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences (Switzerland)*, vol. 6, no. 6, 2016.

[15] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, pp. 4078–4088, 2017.

[16] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, no. January, p. 101236, 2021.

[17] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

IMPACTS OF COLOR AND TEXTURE DISTORTIONS ON EARTH OBSERVATION DATA IN DEEP LEARNING

Martin Willbo[1], **Aleksis Pirinen**[1], **John Martinsson**[1,2], **Edvin Listo Zec**[1,3], **Olof Mogren**[1,4], **Mikael Nilsson**[2]
[1]RISE Research Institutes of Sweden
[2]Centre for Mathematical Sciences, Lund University, Sweden
[3]KTH Royal Institute of Technology
[4]Swedish Centre for Impacts of Climate Extremes (climes)
{martin.willbo@ri.se, aleksis.pirinen@ri.se, john.martinsson@ri.se
edvin.listo.zec@ri.se, olof.mogren@ri.se, mikael.nilsson@math.lth.se}

ABSTRACT

Land cover classification and change detection are two important applications of remote sensing and Earth observation (EO) that have benefited greatly from the advances in deep learning. Convolutional and transformer-based U-net models are the state-of-the-art architectures for these tasks, and their performances have been boosted by an increased availability of large-scale annotated EO datasets. However, the influence of different visual characteristics of the input EO data on a model's predictions is not well understood. In this work we systematically examine model sensitivities with respect to several color- and texture-based distortions on the input EO data during inference, given models that have been trained without such distortions. We conduct experiments with multiple state-of-the-art segmentation networks for land cover classification and show that they are in general more sensitive to texture than to color distortions. Beyond revealing intriguing characteristics of widely used land cover classification models, our results can also be used to guide the development of more robust models within the EO domain.

1 INTRODUCTION

Land cover classification is a key application for remote sensing and Earth observation (EO) data, as it provides essential information for various domains, such as urban planning, environmental monitoring, disaster management, and agriculture. Deep neural networks, such as CNNs and transformers, have demonstrated impressive capabilities and results for processing satellite imagery (Florian & Adam, 2017; Wang et al., 2022; Zhao et al., 2023). However, these models, and the methods used to regularize them (e.g. common image augmentation techniques), are mostly developed and tested on standard imagery (e.g. from ImageNet), which may differ from EO images in several aspects. For instance, previous studies have shown that CNNs trained on ImageNet rely more on texture than on color or shape (Hermann et al., 2020). Such dependencies are less explored in the EO domain, where texture and color may change due to factors such as seasonality, weather, and sensor noise. It is thus essential to understand how different types of visual features and data distortions affect the performance and robustness of deep learning models for EO tasks, and to develop new models and methods that are more suitable for EO data (Rolf et al., 2024).

In this work we aim to provide a better understanding of how different types of test time distortions affect the performance of popular models trained on EO data for land cover classification, and to motivate the development of new data augmentation techniques that are more appropriate for EO models. We study the inductive biases and invariances of popular deep learning models for land cover classification, by applying various test time image distortions that the models have not seen during training. We propose a set of image distortion functions that are *independently applied per image and semantic class* in land cover data: **(i)** converting the pixels of a class into gray-scale[1] (color distortion), and **(ii)** randomly swapping pixel values within a class in an image (texture distortion); see Fig. 1. We then evaluate the performance of the models on OpenEarthMap Xia et al.

[1]Other transformations are also explored in the appendix.

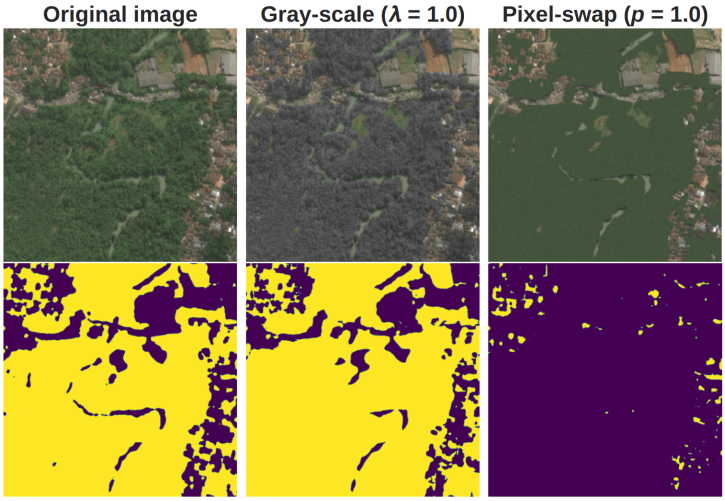


Figure 1: Example image from the training dataset (OpenEarthMap). The class considered here is *tree*. Yellow and dark blue respectively show pixels predicted as *tree* and not *tree*. Top row: Original image, image with **gray-scale** transformation (color distortion) applied, and image with **pixel-swap** transformation (texture distortion) applied, respectively. Note that in the middle, the trees are gray even if they appear to be in color at a glance. Bottom row: Model predictions for the corresponding images in the first row. The transformations are defined in §3; more transformations are explored in the appendix. Predictions made using U-Net-Efficientnet-B4.

(2023), a large-scale and fully labeled benchmark dataset of high-resolution aerial images, under different distortion settings. Our results reveal the strengths and limitations of deep learning models for land cover classification, and offer guidance for future research and improvement.

2 Related Work

Image distortions, such as blur, noise, contrast variation, and JPEG compression, can substantially degrade the accuracy of deep neural networks (DNNs) when applied to the input data. This phenomenon has been demonstrated by Dodge & Karam (2016), who evaluated the impact of different quality distortions on CNNs. Zhou et al. (2017) further investigated the performance of CNNs under blur and noise distortions, and proposed to improve the model robustness by re-training with noisy data. Our work is related to these works, as we also investigate the robustness of DNNs on distorted data. However, our focus is to investigate model sensitivities specifically on EO data, and we leave the development of more robust solutions to future work.

It was shown by Hermann et al. (2020) that common augmentation techniques, such as Random Resized Crop, can introduce bias towards texture rather than shape in the domain of standard image classification. They postulated that aggressive crops may remove distinguishing shape information and push the network to learn to discriminate by texture. We show that DNNs trained on EO data are inherently biased to discriminate by texture even without applying such augmentation techniques. Contemporary work by (Gong et al., 2024) explores segmentation model sensitivity with respect to color variances that are introduced at test time in the standard image domain. They also propose training strategies oriented towards rendering models invariant to color perturbations. However, as our results suggest, segmentation models trained on EO data are already largely color invariant without applying such training strategies. This in turn points to important differences between biases in models trained on data from these two domains.

3 Experimental Setup and Description of Image Distortions

We train and evaluate three models – both CNN- and transformer-based ones – on OpenEarthMap, a large-scale benchmark for land cover classification. It contains 5,000 aerial and satellite images

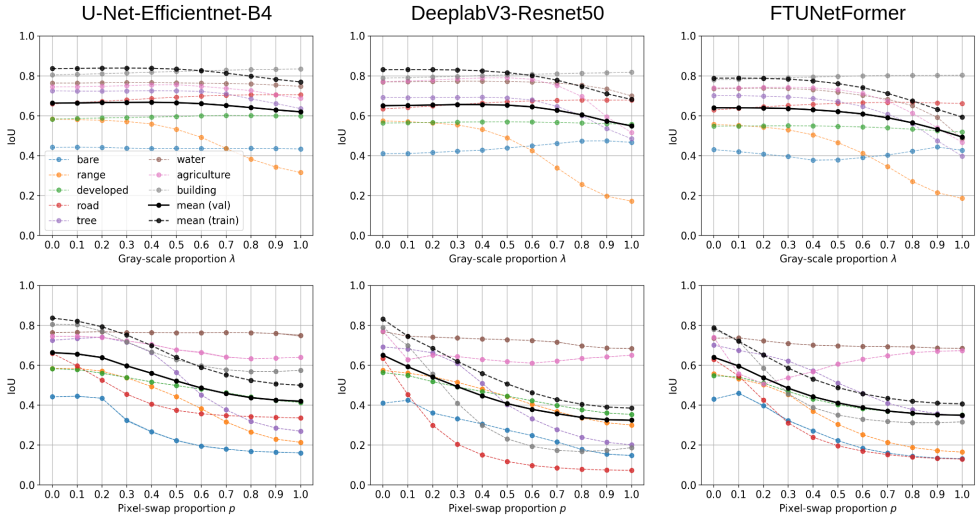


Figure 2: Impact of the **gray-scale** (top) and **pixel-swap** (bottom) transformations at test time on the validation set for the three segmentation models outlined in §3. From left to right: U-Net-Efficientnet-B4, DeeplabV3-Resnet50, and FTUNetFormer. The solid black curve is the mean of the colored curves (validation data), and the dashed black curve is the corresponding mean on training data (included for comparison). Models are generally more sensitive to texture than color distortions. The pixel-swap plot curves are the mean over three realisations of the pixel-swap transform.

(RGB imagery) with 8 class labels at a resolution of 0.25-0.5m. The images span 97 regions from 44 countries across 6 continents, and are split into training, validation, and test sets (we use the official splits). We compare three segmentation models: U-Net-EfficientNet-B4 (Iakubovskii, 2019), DeeplabV3-ResNet50 (Florian & Adam, 2017), and FTUNetFormer Wang et al. (2022). During training, we apply only horizontal and vertical flips (an independent 50% probability for each) as data augmentations. **Thus note that we do not apply any of the distortions described below**[2] **during training** (relevant follow-up experiments would however include investigating the effect of applying them also during model training). We use random crops of size 512×512 with a batch size of 10 during training, and full-size images during evaluation (more details are in the appendix). We ignore the background class; it constitutes $\sim 0.6\%$ of all pixels and is also ignored in the official OpenEarthMap benchmark.

Gray-scale transformation (color distortion). Let $\mathcal{I}_c = \{(i_k, j_k)\}_{k=1}^{K}$ denote the set of pixel positions corresponding to class c in a given RGB aerial or satellite image $\boldsymbol{I}$. Let $\boldsymbol{I}(i,j) \in [0,255]^3$ denote the pixel (three channels) at coordinate $(i,j) \in \mathcal{I}_c$, and let $\lambda \in [0,1]$. Then, the gray-scale transformation for image $\boldsymbol{I}$ and class c is defined as $\boldsymbol{I}(i,j) = (1-\lambda)\boldsymbol{I}(i,j) + \lambda \boldsymbol{G}(i,j)$, where $\boldsymbol{G}$ is a corresponding gray-scale image derived by the *rgb2gray* transformation from Scikit-image. Other pixels are left unchanged. The gray-scale image is duplicated over the three color channels, i.e. $\boldsymbol{G}(i,j) \in [0,255]^3$ has the same value in each element. See Fig. 1 for an example with $\lambda = 1$.

Pixel-swap transformation (texture distortion). For a pixel-swap proportion p we randomly sample pK pixel positions without replacement from $\mathcal{I}_c$ and randomly permute the pixel values at these positions. Other pixels are left unchanged here as well. See Fig. 1 for an example with $p = 1$.

4 Experimental Results

Fig. 2 demonstrates the impact of the gray-scale and pixel-swap distortions (at varying intensities, see x-axes) on test time predictions on unseen validation data. Recall that the models were trained without any such distortions. We see that for all models, but particularly for U-Net-Efficientnet-B4 (left), there is quite a strong invariance to color distortions (top row), i.e. even though the images are severely altered and moved far in image space, performance does not suffer much in general.

[2]Please also see to the appendix for additional test time image distortions and associated results.

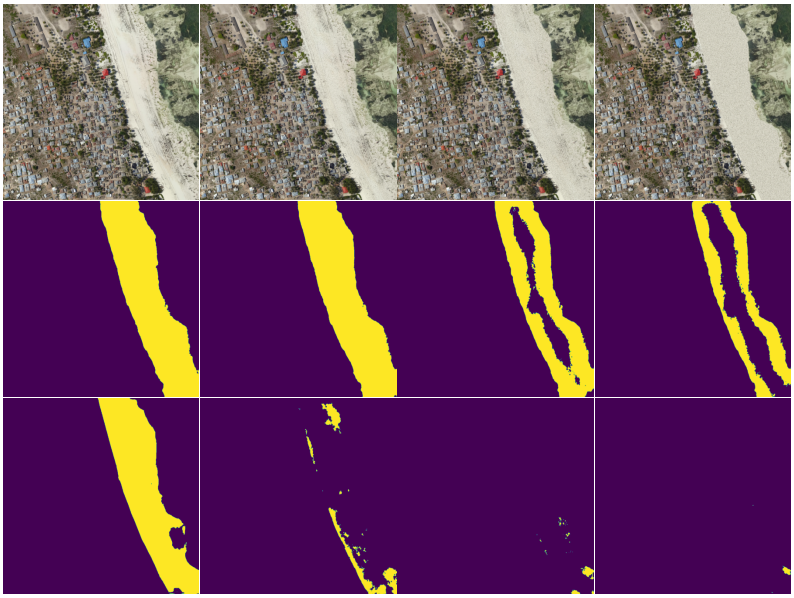


Figure 3: **Top:** Zanzibar region, **pixel-swap** transformation on the *bare* class with proportion p swapped, where $p \in \{0, 0.33, 0.66, 1\}$ (left to right). **Middle:** Corresponding model predictions, where yellow and dark blue respectively show pixels predicted as *bare* and *not bare*. The border region remains correctly classified regardless of transformation intensity, so the surrounding context is critical. **Bottom:** Same as middle, but predictions obtained from the same images and distortions as in the first row, but where all pixels except *bare* ones have been masked out in the images by replacing them with the per-channel mean of the training set (more such results are in the appendix). The importance of context is clear. Predictions made using the U-Net-Efficientnet-B4 model.

For some classes however (e.g. *range*), performance deteriorates as the images become fully gray-scale transformed. We also see that although the models have similar mIoUs (black curves) under no gray-scale transformation, and similar degradation trends, the transformer-based model (right in Fig. 2) seems to be the most sensitive to color distortions.

As for texture distortion (pixel-swaps, see bottom row) we see more rapid performance drops in general. For smaller distortions, e.g. pixel-swap proportions of 0.1-0.3, we see performance degradations for several classes and on average for all three models. At these lower distortion proportions it is difficult for the human eye to notice any changes, as can be seen in the examples in the appendix. However, note that even though the models are shown to be sensitive to texture distortions, none of the IoUs approach zero as the intensity of the transformation grows. In Fig. 3 we show how the model prediction for the *range* class for one of the images changes as the intensity of the pixel-swap transformation increases. Note how the interior of the region associated with the class becomes increasingly misclassified, while the border remains correctly classified. This coupled with the third row of Fig. 3 clearly indicates that the network leverages surrounding context for its predictions. *We refer the reader to the appendix for more quantitative and qualitative results.*

Finally, we note that *range* and *tree* are among the classes that are most affected by degradations – be them color- or texture-based – even though they are among the most common classes in the training set (measured in total number of pixels). Thus, somewhat counter-intuitively, the robustness to distortions for a given class is not related to the relative amount of training data for said class.

More on the importance of surrounding context. We here further examine how model predictions for a given class are affected by context from neighboring pixels of other classes (beyond the qualitative example in Fig. 3, more of which are found in the appendix). As a basis for this investigation we use the pixel-swap transformation, but here we also replace all pixel values *not* related to the class under investigation with the per-channel mean of the training data set, in order to remove all surrounding context. See Fig. 8 and Fig. 9 in the appendix for two qualitative examples. We see that removing context affects model predictions significantly, and in particular that the accuracy

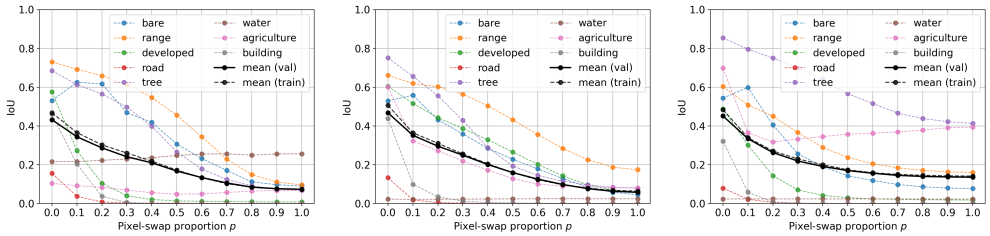


Figure 4: Impact of the **pixel-swap** transformation where all pixels except for the class under investigation are replaced by the per-channel mean of the training set. Results are shown for the three segmentation models outlined in §3. From left to right: U-Net-Efficientnet-B4, DeeplabV3-Resnet50, and FTUNetFormer. The solid black curve is the mean of the colored curves (validation data), and the dashed black curve is the corresponding mean on training data (included for comparison). Models perform significantly worse in general, compared to the case where context is kept intact (see also Fig. 5).

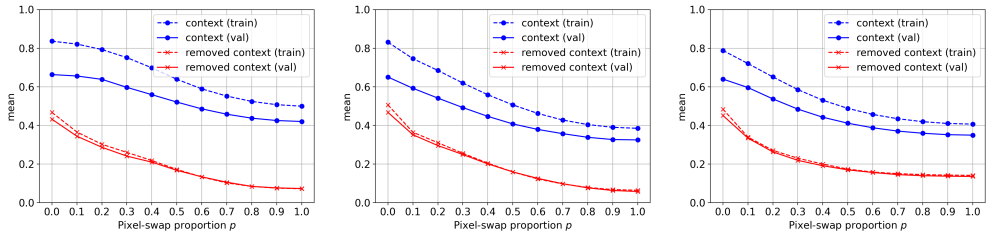


Figure 5: Comparison of keeping context intact (blue) and removing context (red), with various proportions of the **pixel-swap** transformation. The blue curves are identical to the means (in black) of Fig. 2. The red curves are identical to the means (in black) of Fig. 4. From left to right: U-Net-Efficientnet-B4, DeeplabV3-Resnet50, and FTUNetFormer. There is a significant performance drop at all proportions p, even when no pixel-swap is applied ($p = 0$), and the difference between the training and validation set is vastly smaller when surrounding context is removed.

deteriorates much faster as the pixel-swap proportion p increases, compared to the setting where the context is kept intact. In Fig. 8 we see that the border predictions also fail when context is absent, while in Fig. 9 we see a significant drop in prediction accuracy even at $p = 0$.

In Fig. 4, the effect of context removal is quantitatively examined on the whole validation set (including also average results on the training set, for comparison). Comparing with Fig. 2, we see that removing context yields a significant performance drop in general. There is also a smaller gap in performance between the training and validation sets when context is removed, even for $p = 0$ (i.e. with no pixel-swap applied). This suggests that removing surrounding information inhibits the model from making correct predictions for the class of interest, whether or not this is data on which the model has been trained. To make these comparisons easier, see also Fig. 5.

5 Conclusions

In this paper we have investigated the impacts of several test time color and texture distortions on EO imagery. Our experiments – conducted using popular CNN- and transformer-based models – suggest that deep networks, which have not been exposed to these distortions in training, are relatively robust to EO image color distortions but are sensitive to texture distortions. Further, our experiments indicate that models use the surrounding context when making predictions and are sensitive to changes in this context. These empirical findings, while intriguing on their own, also point to many future areas of potential research and improvements regarding land cover classification and EO tasks more broadly. For example, since our results suggest that there is a large variation in how different classes are affected by the various distortions, it may be possible to leverage these insights to develop effective class-dependent data augmentation techniques in the EO domain.

REFERENCES

Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2016.

L-CCGP Florian and Schroff Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *Conference on computer vision and pattern recognition (CVPR). IEEE/CVF*, volume 6, 2017.

Yunpeng Gong, Jiaquan Li, Lifei Chen, and Min Jiang. Exploring color invariance through image-level ensemble learning. *arXiv preprint arXiv:2401.10512*, 2024.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19000–19015. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/db5f9f42a7157abe65bb145000b5871a-Paper.pdf.

Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*, 2024.

Scikit-image. rgb2gray. https://scikit-image.org/docs/stable/auto_examples/color_exposure/plot_rgb_to_gray.html.

Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023.

Shengyu Zhao, Kaiwen Tu, Shutong Ye, Hao Tang, Yaocong Hu, and Chao Xie. Land use and land cover classification meets deep learning: A review. *Sensors*, 23(21):8966, 2023.

Yiren Zhou, Sibo Song, and Ngai-Man Cheung. On classification of distorted images with deep convolutional neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1213–1217. IEEE, 2017.

Appendix: Impacts of Color and Texture Distortions on Earth Observation Data in Deep Learning

In this appendix we provide several additional results (both quantitative and qualitative) for the image distortions introduced in the main paper, as well as for additional distortions defined in this appendix. We also provide more details regarding our experimental setup.

Additional experimental setup details. For convenience, before expanding with additional details, we first attach here some of the implementation details that were already explained in the main paper. During training, we apply only horizontal and vertical flips (an independent 50% probability for each) as data augmentations. *Thus note that we do not apply any of the color- or texture-based distortions during training.* All three models (see §3 in the main paper) are trained and evaluated on images resized to 512×512 or 1024×1024, however during training we sample random crops of size 512×512 from these images, for 2,000 epochs with a batch size of 10, and the best ones are selected based on the validation mIoU. We use a standard cross-entropy loss and the Adam optimizer (Kingma & Ba, 2014) with learning rate 0.0002, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. We ignore the background class; it constitutes $\sim 0.6\%$ of all pixels and is also ignored in the official OpenEarthMap benchmark.

Additional color distortion experiments. Here we include experimental results for additional forms of color distortion, which we commonly refer to as *color-duplication*, as described next.

Color-duplication: As in the main paper, let $\mathcal{I}_c = \{(i_k, j_k)\}_{k=1}^{K}$ denote the set of pixel positions corresponding to class c in a given RGB aerial or satellite image $\boldsymbol{I}$. The color-duplication transformation is somewhat similar to the gray-scale transformation defined in the main paper. However, instead of a convex combination between the class c-related pixels of the original image $\boldsymbol{I}$ and a gray-scale counterpart $\boldsymbol{G}$, we first copy one color channel $\boldsymbol{I}_h$ (with $h \in \{R, G, B\}$ and where R, G and B respectively denote the red, blue and green color channels), and then construct the color-duplicated image $\boldsymbol{I}^D = [\boldsymbol{I}_h; \boldsymbol{I}_h; \boldsymbol{I}_h]$. Given $\boldsymbol{I}^D$, the class c-related color duplication at mixing proportion λ is given by $\boldsymbol{I}(i,j) = (1-\lambda)\boldsymbol{I}(i,j) + \lambda \boldsymbol{I}^D(i,j) = (1-\lambda)[\boldsymbol{I}_R; \boldsymbol{I}_G; \boldsymbol{I}_B](i,j) + \lambda[\boldsymbol{I}_h; \boldsymbol{I}_h; \boldsymbol{I}_h](i,j)$, where $(i,j) \in \mathcal{I}_c$. For example, in the red-duplication case with $h = R$, the transformation is given by $(1-\lambda)[\boldsymbol{I}_R; \boldsymbol{I}_G; \boldsymbol{I}_B](i,j) + \lambda[\boldsymbol{I}_R; \boldsymbol{I}_R; \boldsymbol{I}_R](i,j) = [\boldsymbol{I}_R; (1-\lambda)\boldsymbol{I}_G + \lambda \boldsymbol{I}_R; (1-\lambda)\boldsymbol{I}_B + \lambda \boldsymbol{I}_R](i,j)$. As usual, $\boldsymbol{I}$ is left unchanged for pixels of other classes. Finally, note that we still normalize the model input with the per-channel means of the training set, even if the channels are changed.

The results for the color-duplication transformations (red-, green- and blue-duplications) are shown in Fig. 6. As can be seen, despite the extreme nature of this type of color distortion, most classes are only marginally affected by it. Hence, taken into account also the gray-scale experiments from the main paper, these results suggest that deep learning models are not very sensitive to color transformations of EO imagery.

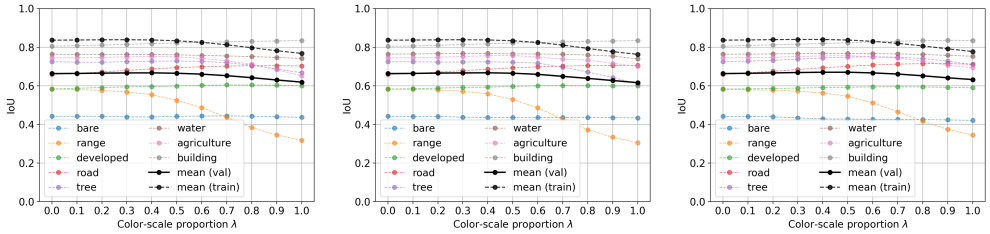


Figure 6: Impact of the **color-duplication** transformations for the U-Net-Efficientnet-B4 model. From left to right: Red-duplication, green-duplication, and blue-duplication. The solid black curve is the mean of the colored curves (validation data), and the dashed black curve is the corresponding mean on training data (included for comparison). The performance degradation trends are similar for all three transformations, and are also comparable to the gray-scale color distortion (see main paper). All taken together, these results suggest that deep networks are quite robust to color distortions on EO imagery.

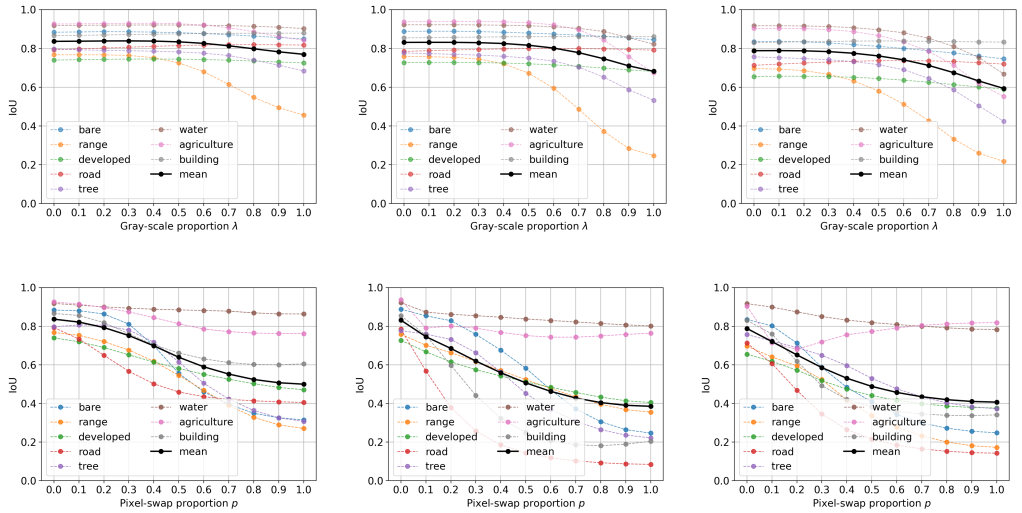


Figure 7: Impact of the **gray-scale** (top row) and **pixel-swap** (bottom row) transformations at test time on the training set for the three segmentation models outlined in §3 (main paper). From left to right: U-Net-Efficientnet-B4, DeeplabV3-Resnet50, and FTUNetFormer. The solid black curve is the mean of the colored curves. As is the case on validation data (cf. Fig. 2 in the main paper), models are generally more sensitive to texture than color distortions on training data as well.

More results for the image distortion experiments in the main paper. Here we show more visual results for the gray-scale and pixel-swap distortion experiments (i.e. the distortions investigated in the main paper); see Fig. 10 - 17. In particular, we show qualitative results for each of the eight classes in the dataset, as we change the intensity of the respective distortions. Note that the illustrated distortion intensity ranges vary between the examples; these range choices are to a large extent guided by the class-specific sensitivities seen in Fig. 2 (cf. main paper), and are set in such a way that interesting accuracy degradations are highlighted where possible. All model predictions in these qualitative examples are from the U-Net-Efficientnet-B4 model. Finally, for completeness we also show a plot similar to Fig. 2 (main paper), but for the training set – see Fig. 7.

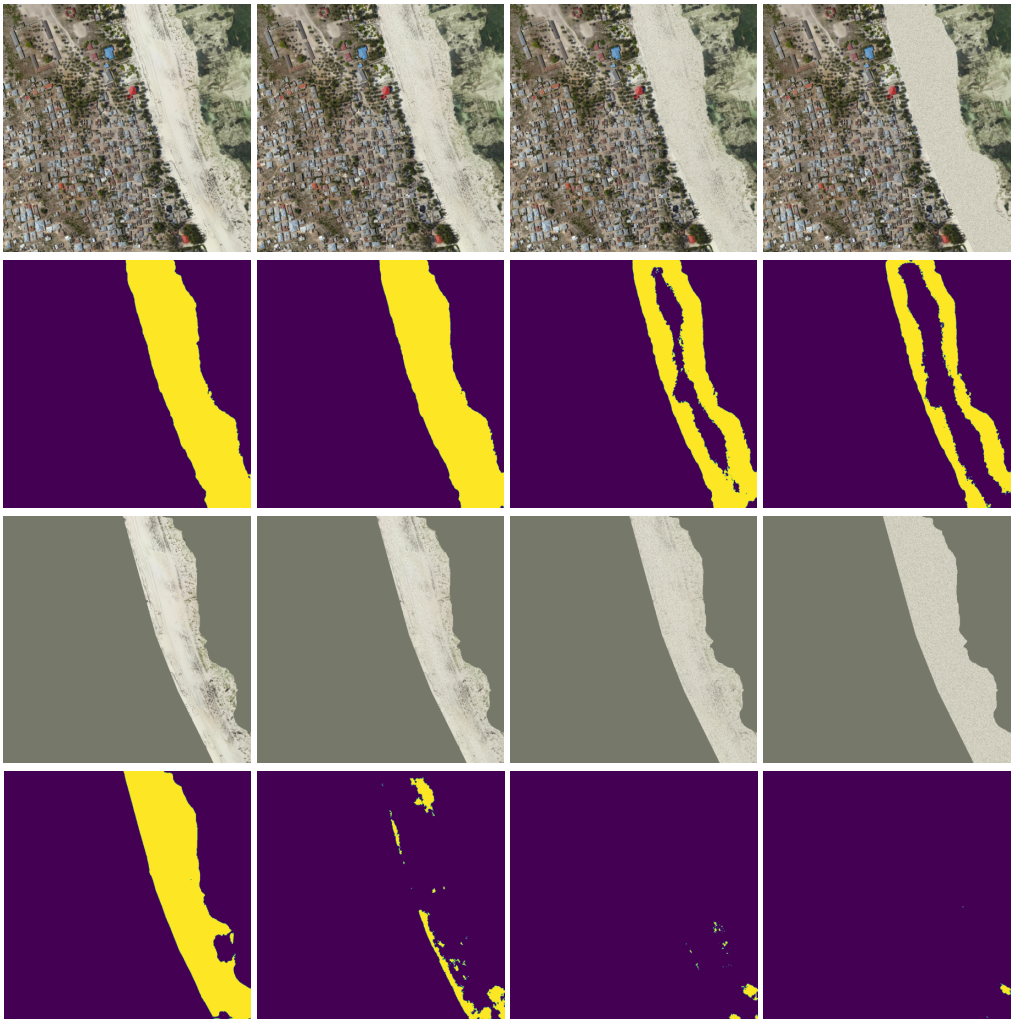


Figure 8: Zanzibar region, **pixel-swap** transformation on the *bare* class (expanded version of Fig. 3 (main paper)). Top two rows: Input images and model predictions. Bottom two rows: Input images, where pixels *not* of the *bare* class have been replaced by the per-channel mean of the training set, and model predictions below. Pixel-swap is applied to the input images with proportion p swapped, where $p \in \{0, 0.33, 0.66, 1\}$ (from left to right). Note in the bottom row how the prediction accuracy deteriorates much faster as p increases, compared to the case when the surrounding context is kept intact (second row). Also note how the border is no longer correctly classified for the input images when the surrounding context is removed. Predictions made using the U-Net-Efficientnet-B4 model.

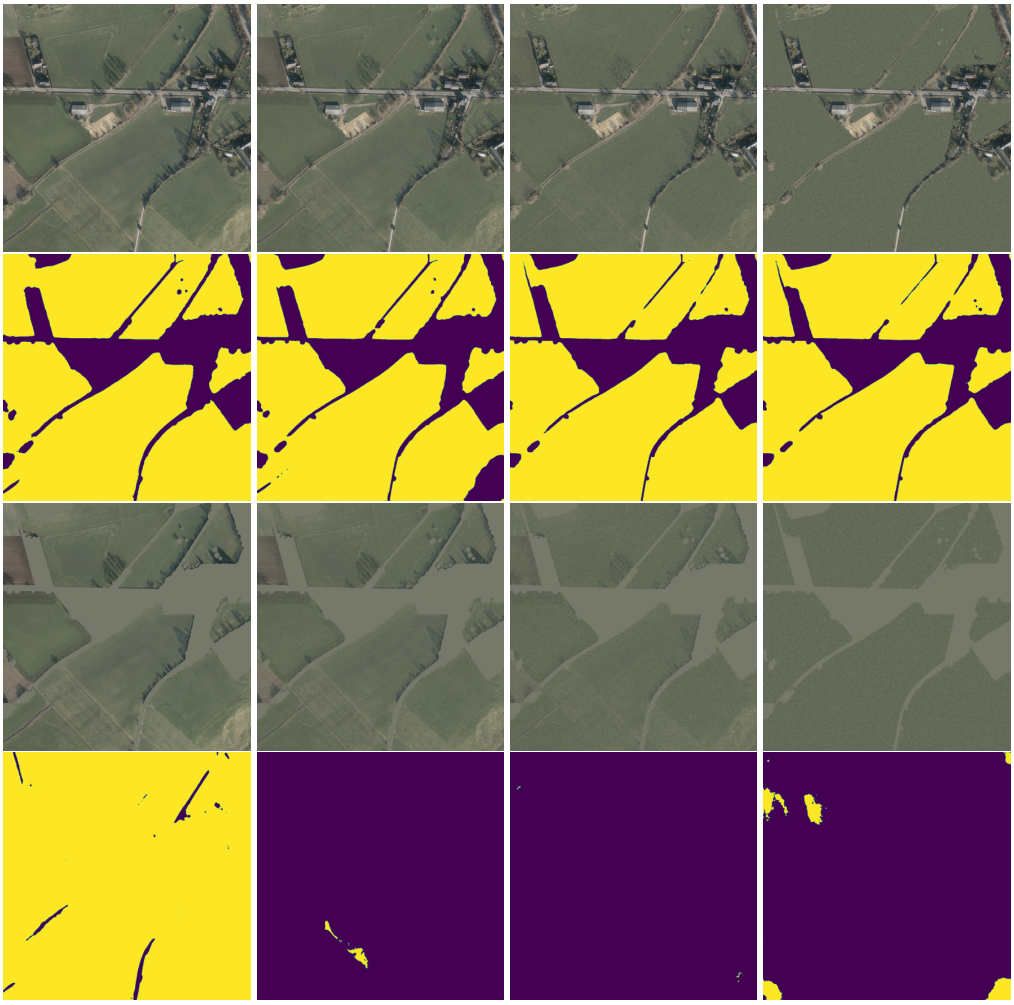


Figure 9: Aachen region, **pixel-swap** transformation on the *agriculture* class. Top two rows: Input images and model predictions. Bottom two rows: Input images, where pixels *not* of the *agriculture* class have been replaced by the per-channel mean of the training set, and model predictions below. Pixel-swap is applied to the input images with proportion p swapped, where $p \in \{0, 0.33, 0.66, 1\}$ (from left to right). Removing the surrounding context yields significantly worse predictions for each proportion p (bottom row), whereas predictions remain accurate independently of p when context is kept (second row). Predictions made using the U-Net-Efficientnet-B4 model.

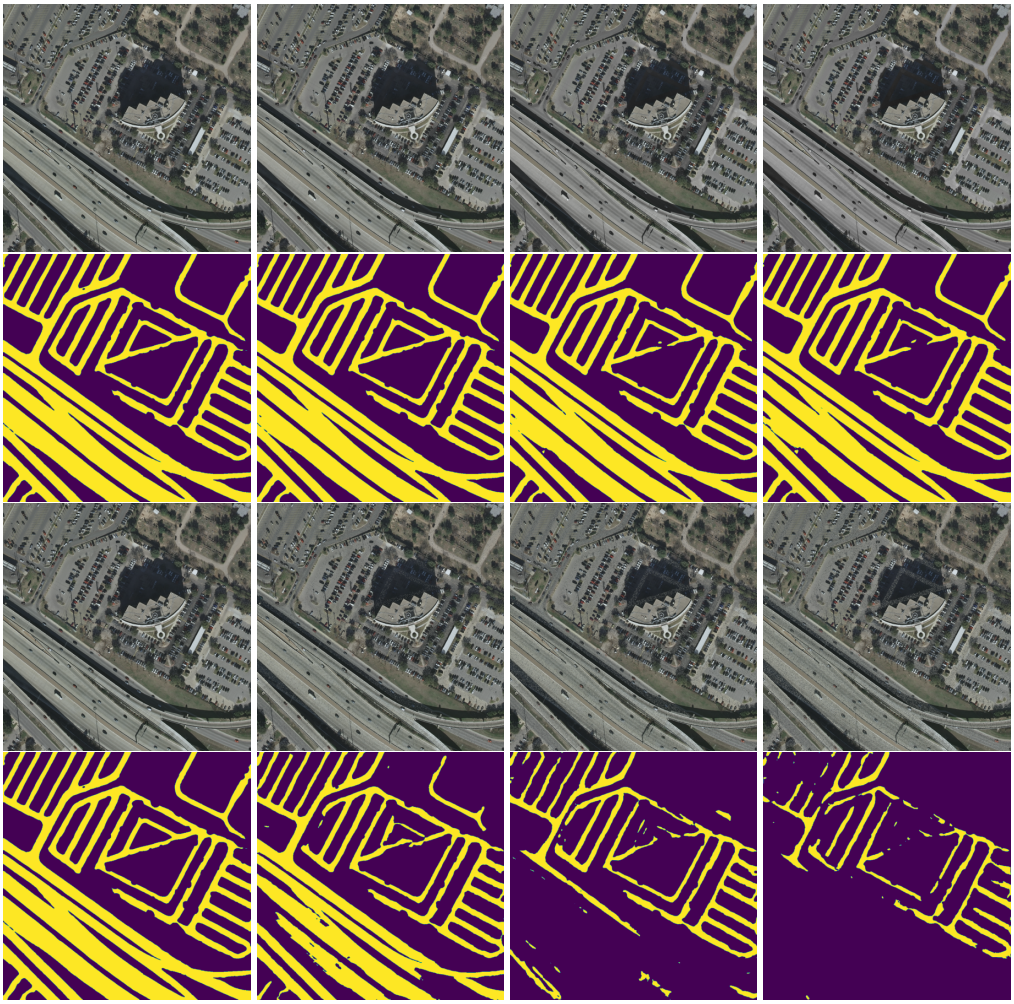


Figure 10: Austin region, transformations on the *road* class. The top two rows show the **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions below. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, where $p \in \{0, 0.1, 0.2, 0.3\}$ (from left to right) and corresponding model predictions below. We see that the predictions are very robust with respect to color distortion (top), and very sensitive to texture distortion (bottom).

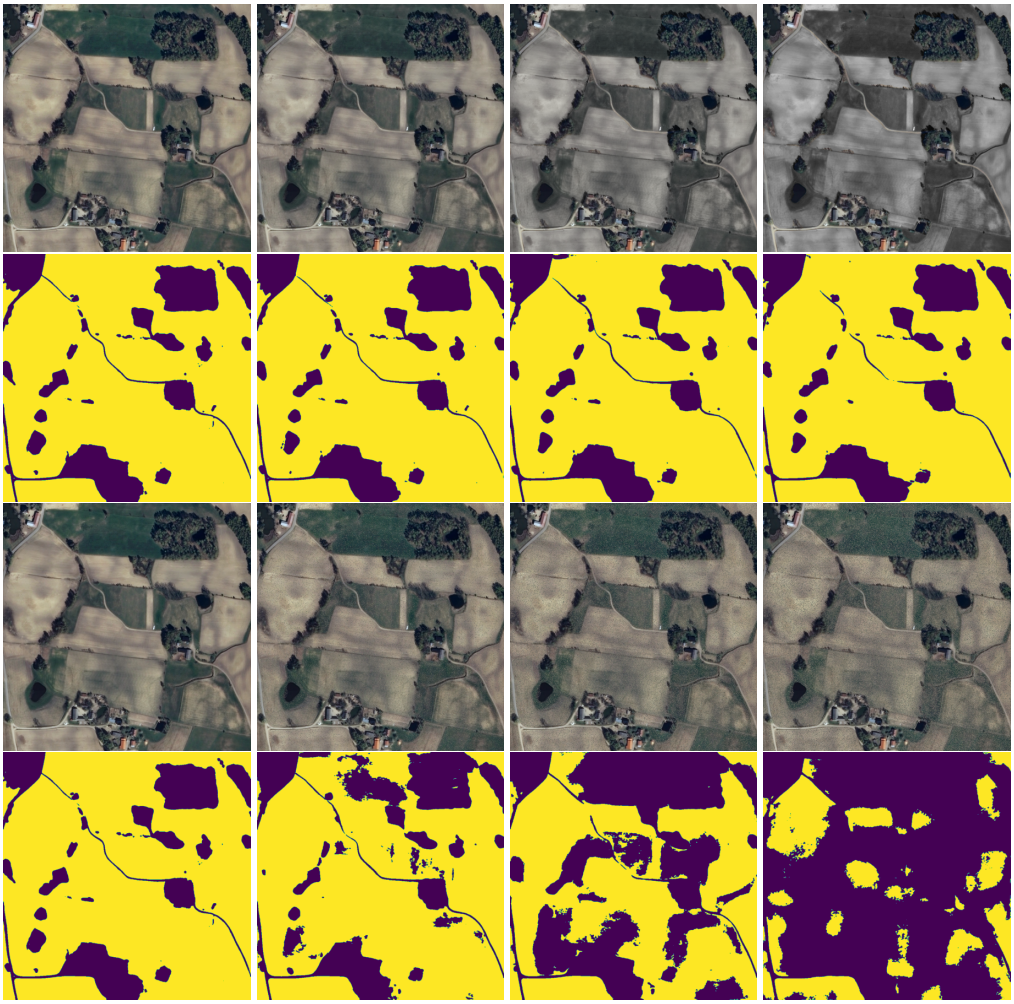


Figure 11: Pomorskie region, transformations on the *agriculture* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.1, 0.2, 0.3\}$ (from left to right) and corresponding model predictions below. We see that the predictions are very robust with respect to color distortion (top), and very sensitive to texture distortion (bottom)

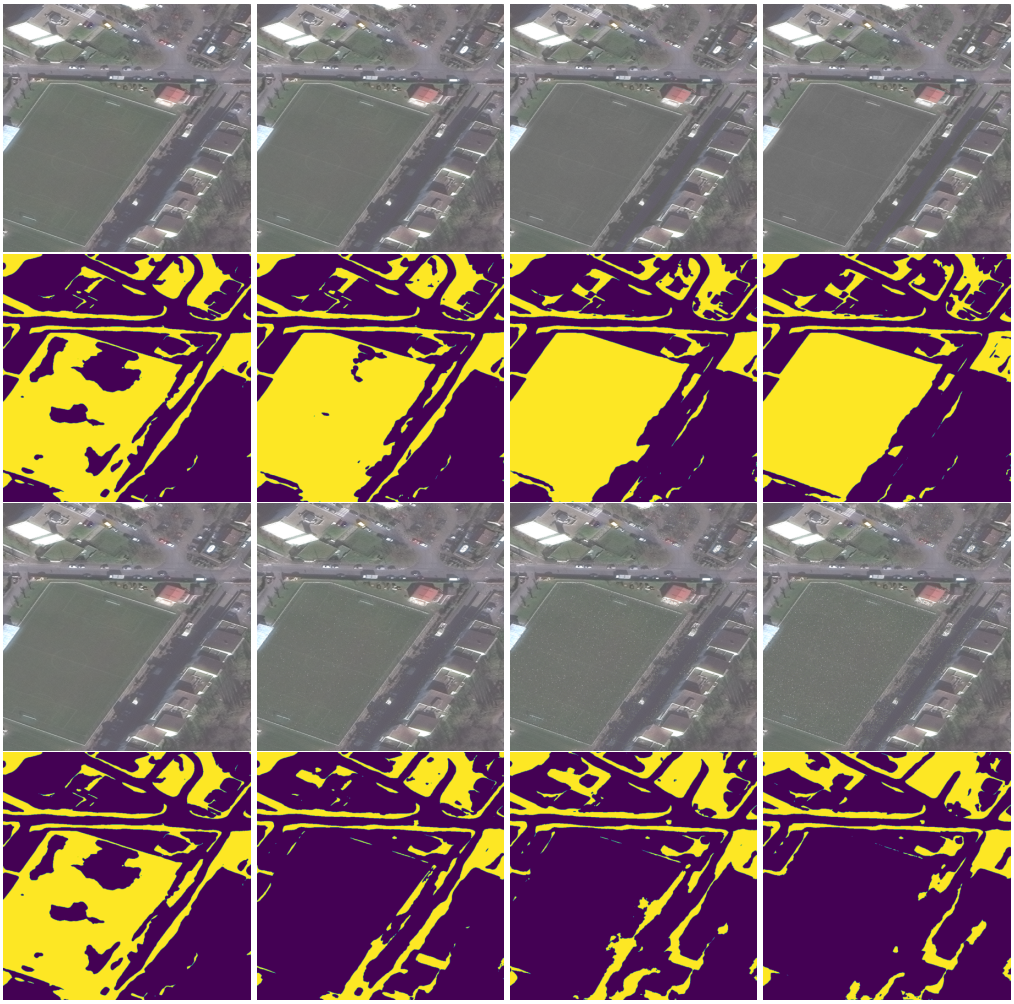


Figure 12: Paris region, transformations on the *developed* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.1, 0.2, 0.3\}$ (from left to right) and corresponding model predictions below. We see that the predictions are very robust with respect to color distortion (top), and very sensitive to texture distortion (bottom).

Figure 13: Mahe region, transformations on the *water* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions below. We see that the predictions, in contrast to most other classes, are very robust with respect to color distortion (top) and texture distortion (bottom).

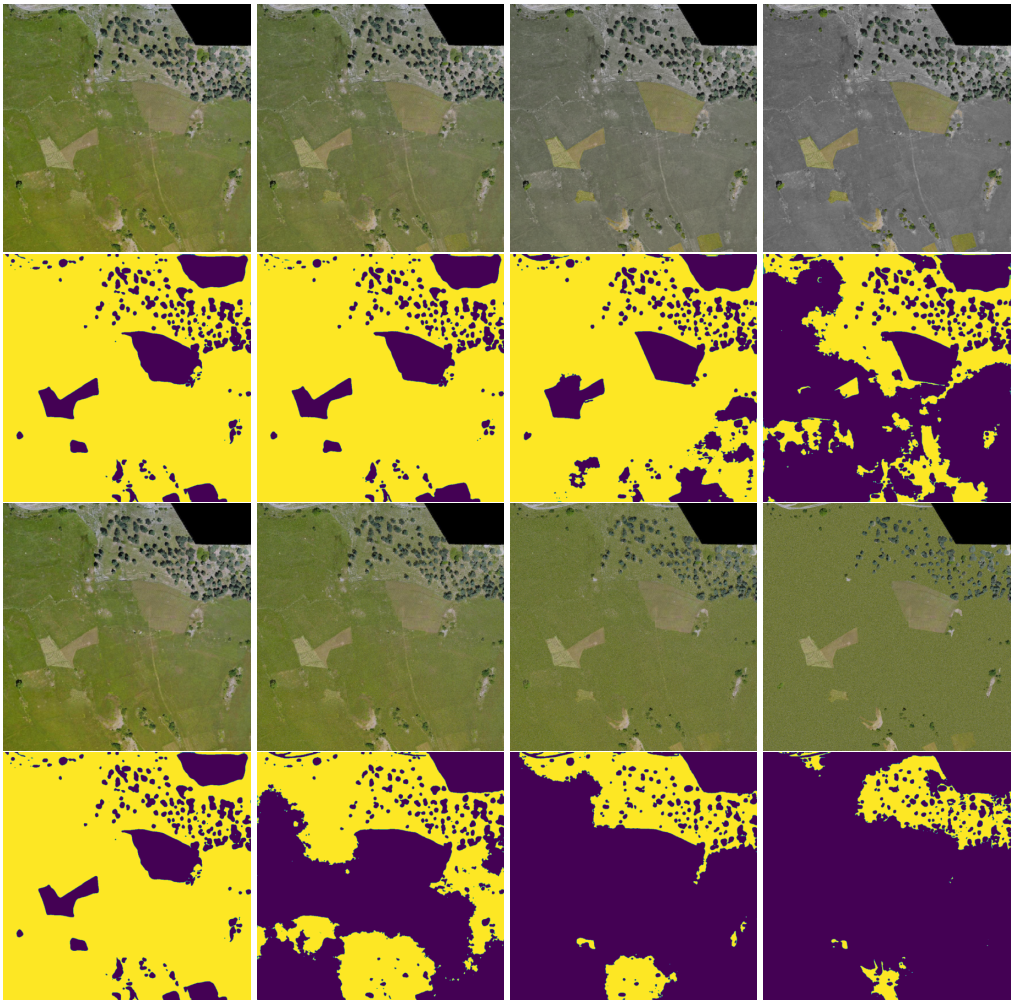


Figure 14: Svaneti region, transformations on the *range* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions below. We see sensitivities with respect to color distortion (top), especially with gray-scale proportion $\lambda = 1$. The model predictions are however more sensitive to texture distortion (bottom).

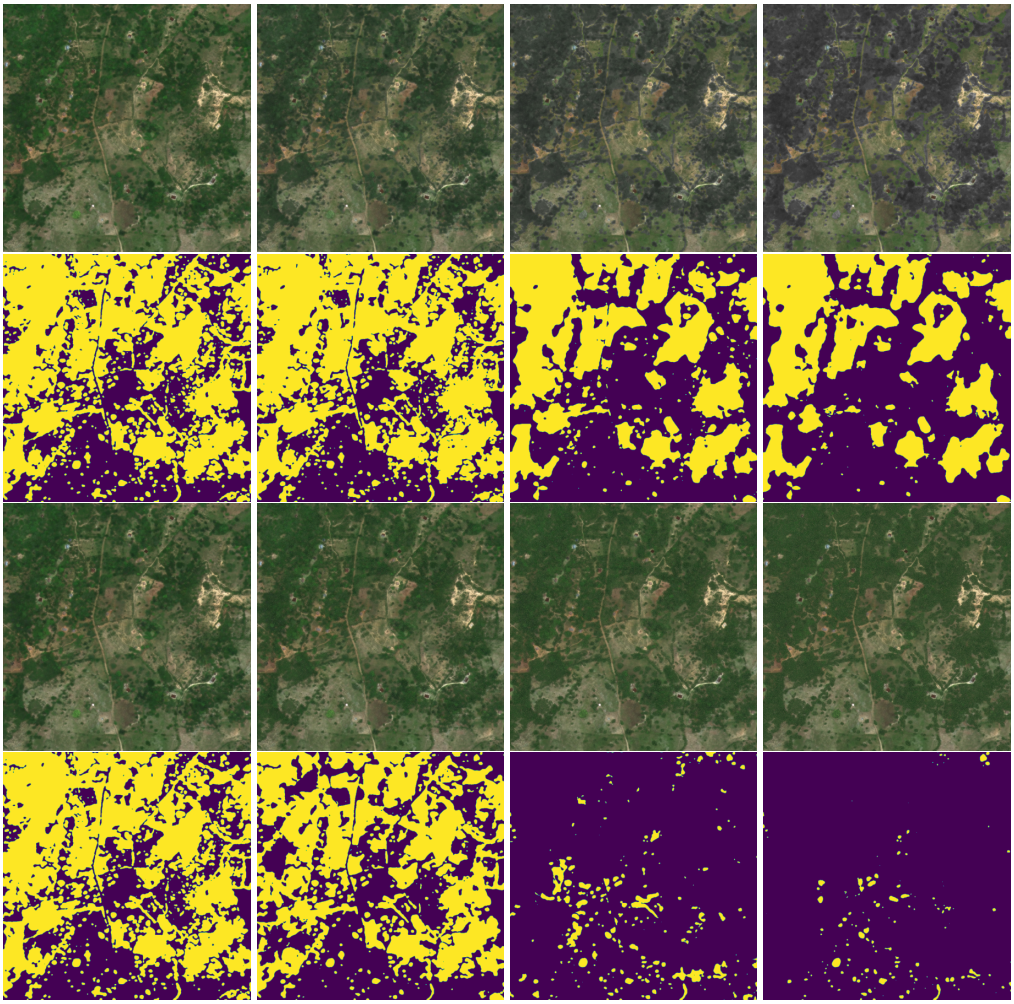


Figure 15: Jeremie region, transformations on the *tree* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.2, 0.4, 0.6\}$ (from left to right) and corresponding model predictions below. We see that the predictions are robust with respect to color distortion with exception for gray-scale proportion $\lambda = 1$ (top), and sensitive to texture distortion (bottom).

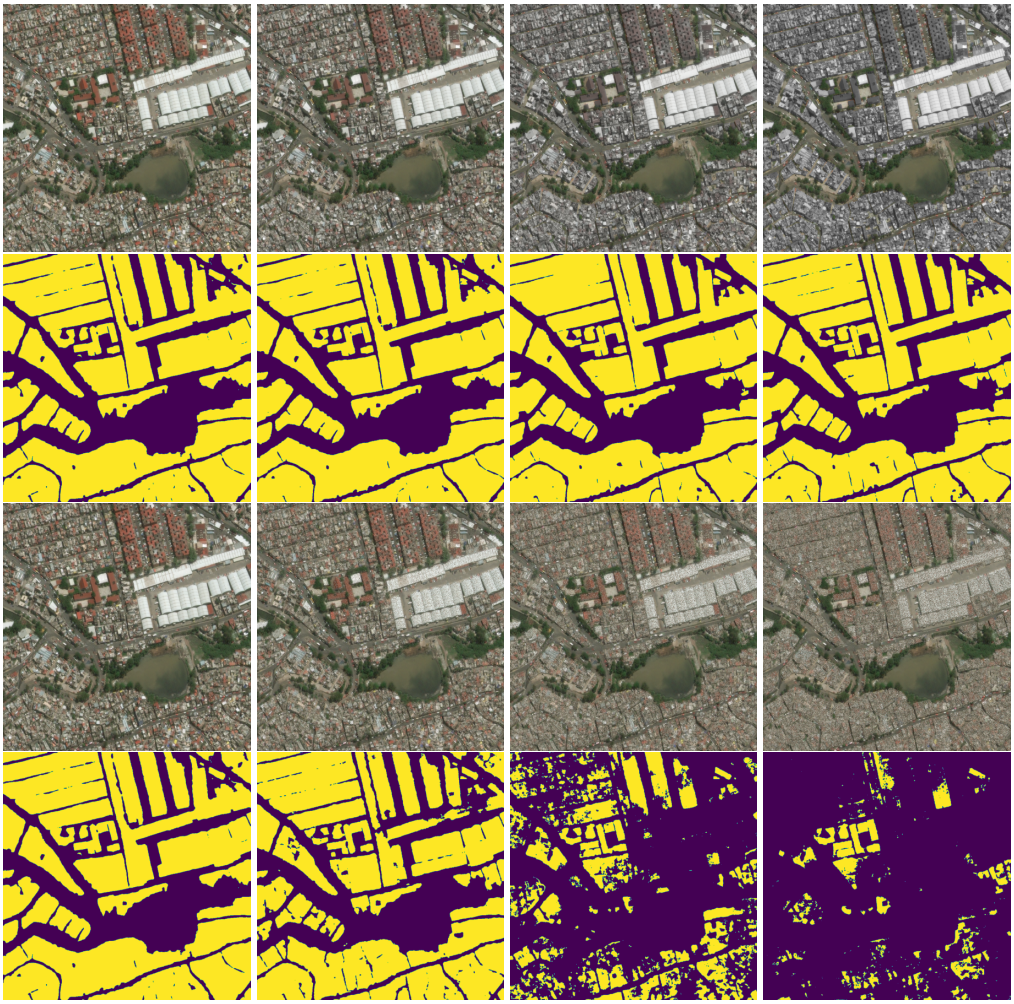


Figure 16: Mexico City region, transformations on the *building* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.2, 0.4, 0.6\}$ (from left to right) and corresponding model predictions below. We see that the predictions are very robust with respect to color distortion (top), and sensitive to texture distortion (bottom).

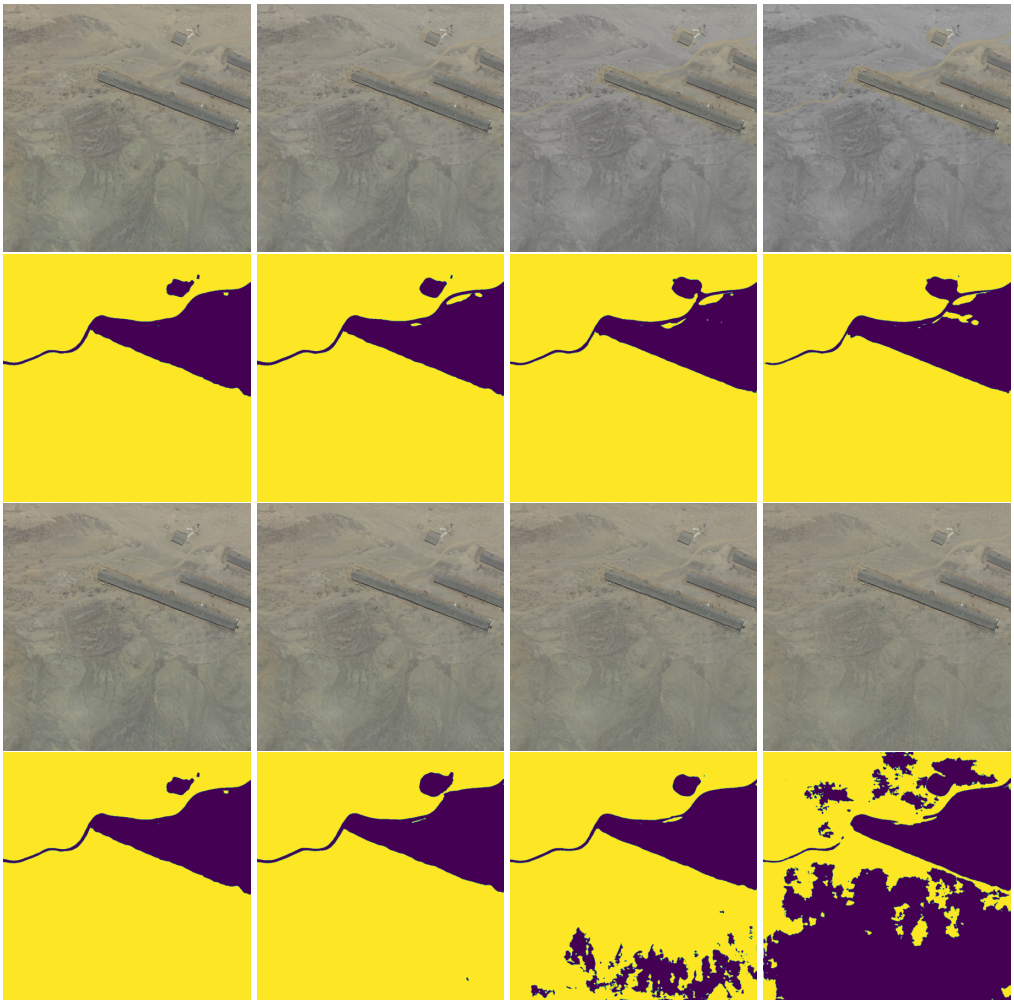


Figure 17: Lima region, transformations on the *bare* class. The top two rows show **gray-scale** transformed images with gray-scale proportion $\lambda \in \{0, 0.33, 0.66, 1\}$ (from left to right) and corresponding model predictions. The bottom two rows show **pixel-swap** transformed images with proportion p swapped, $p \in \{0, 0.1, 0.2, 0.3\}$ (from left to right) and corresponding model predictions below. We see that the predictions are very robust with respect to color distortion (top), and very sensitive to texture distortion (bottom).

Efficient Node Selection in Private Personalized Decentralized Learning

Edvin Listo Zec*[1,2], Johan Östman[3], Olof Mogren[1], and Daniel Gillblad[3]

[1]RISE Research Institutes of Sweden
[2]KTH Royal Institute of Technology
[3]AI Sweden
edvin.listo.zec@ri.se

Abstract

Personalized decentralized learning is a promising paradigm for distributed learning, enabling each node to train a local model on its own data and collaborate with other nodes to improve without sharing any data. However, this approach poses significant privacy risks, as nodes may inadvertently disclose sensitive information about their data or preferences through their collaboration choices. In this paper, we propose Private Personalized Decentralized Learning (**PPDL**), a novel approach that combines secure aggregation and correlated adversarial multi-armed bandit optimization to protect node privacy while facilitating efficient node selection. By leveraging dependencies between different arms, represented by potential collaborators, we demonstrate that PPDL can effectively identify suitable collaborators solely based on aggregated models. Additionally, we show that PPDL surpasses previous non-private methods in model performance on standard benchmarks under label and covariate shift scenarios.

1 Introduction

Collaborative machine learning is a recent paradigm where multiple actors train a joint model without revealing their local datasets [1]. Instead, only the locally trained model parameters are shared among the actors. In applications pertaining to sensitive data, e.g., healthcare and banking, where it may be challenging to collect the data in a single location, collaborative learning has the potential to unlock a plethora of novel collaborations. Collaborative learning is typically distinguished with regard to the underlying network topology. To this end, federated learning (FL) refers to a star topology where an orchestrating parameter server receives model updates from the actors, aggregates the updates, and broadcasts the aggregate. Decentralized learning (DL) constitutes arbitrary network topologies without an orchestrator, i.e., actors in the network learn by exchanging model updates within their neighborhood [2]. Actors within DL are typically referred to as nodes.

There are inherent risks and limitations with FL, such as that it may be challenging to find a trustworthy third party due to regulations or the desire for autonomy (e.g. for hospitals, banks, or other big corporations). Further, FL scales poorly in the number of nodes due to the communication bottleneck and the server constitutes a single-point-of-failure [2]. This has motivated research on fully decentralized systems, which eliminate the need for a central server. Instead, model parameters are directly communicated between peers in the learning setup using a communication protocol, such as gossip learning [3]. However, this approach is not well-suited for non-iid settings, where multiple distinct learning objectives may be present. In such cases, node selection during training is crucial for achieving efficient and effective learning.

The idea of each node identifying useful peers in the network to train a personalized model was proposed in [4]. Therein, nodes jointly learn a collaboration graph, via an alternating optimization method, that dictates whom to communicate to. A score-based method, decentralized adaptive clustering (DAC), was presented in [5] where each node scores its neighboring peers based on the the empirical loss, obtained by evaluating the received model parameters on the local dataset. While DAC manages to find beneficial nodes and identifies heterogeneous clusters in the network, model parameters from the nodes' training updates are still communicated over the network in plain text and the peers receiving the updates must hence be trusted. As such, DAC is vulnerable to inference attacks. This raises the question of how to ensure the privacy of the model parameters in decentralized machine learning systems. In many privacy-critical applications, differential privacy [6] is used in conjunction with FL to protect the data of nodes. Although this adds a layer of privacy, it comes at the expense of a deterioration in model performance.

In this work, we overcome this problem and introduce a communication-efficient and privacy-preserving algorithm named **Private Personalized Decentralized Learning (PPDL)**. We use multi-armed bandits to find beneficial collaborators and secure aggregation [7, 8] to hide individual updates. Our method works in a serverless decentralized setting, but can also apply to standard FL. We protect against inference attacks by only observing aggregated models.In our proposed method, a peer only observes an aggregate of model parameters, which substantially lessens the risk of inference attacks as compared to previous works.

Since a node only receives an aggregate of the parameter updates of M nodes at a given point in time,

*Corresponding Author.

Proceedings of the 6th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2025.

it cannot infer a score on the similarity of any one of the peers in the aggregate (as in DAC); such a score can only be computed for the aggregate. Instead, our solution exploits dependencies between different group selections and makes use of adversarial multi-armed bandit optimization to efficiently find the subsets of peers that are beneficial for collaboration. Our experimental evaluations demonstrate that our approach offers a competitive solution for personalized decentralized learning that preserves data privacy under covariate shift and label shift and efficiently finds the beneficial collaborators within the network. Our solution has a communication efficiency and performance similar to that of previous methods, but adds a higher level of privacy.

2 Decentralized learning by finding useful collaborations

Problem formulation. We consider several DL tasks over a network of K nodes, each with a *private* data distribution $\mathcal{D}_i$ over the inputs $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$. Each node $i \in [K]$ has a model f_i with parameters $w_i \in \mathbb{R}^d$ and a loss function $\ell(f_i(w_i; x), y) : \mathbb{R}^d \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Each note wants to minimize its expected loss over its data,

$$w_i^\star = \arg\min_{w_i \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \left[\ell(f_i(w_i; x), y)\right]. \quad (1)$$

A challenge is to find similar nodes to collaborate with, without sharing data. If the distributions are substantially dissimilar, collaboration may result in decreased performance compared to local training without collaboration. In situations where some of the other nodes in the network have similar local data distributions, it may be beneficial to collaborate towards the goal in (1) by means of exchanging and aggregating model parameters.

However, revealing details of node data may be difficult or impossible due to privacy reasons. To address this issue, we propose a method for identifying nodes with similar local datasets in a private manner. We assume the nodes communicate over a network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N} = \{1, \ldots, K\}$ are the nodes and $\mathcal{E} = \{(i, j) : i, j \in \mathcal{N}, i \neq j\}$ are the edges between the nodes. The neighborhood of node $i \in \mathcal{N}$ is denoted by $\mathcal{N}_i = \{j : (i, j) \in \mathcal{E}, j \in \mathcal{N}\}$. Like [5, 9], node i want to find a set of nodes $\mathcal{M}_i \subseteq \mathcal{N}_i$ to exchange models with. In each round, the learning proceeds as follows. First, each node $i \in \mathcal{N}$ selects a set $\mathcal{M}_i \subseteq \mathcal{N}_i$ to receive model updates from. Second, the nodes in $\mathcal{M}_i$ submit their local models securely to node i by using secure aggregation, e.g., [8]. Third, node i computes the aggregated model from the nodes in $\mathcal{M}_i$ and aggregates it with its local model after which local training is initiated using the updated model.

Privacy. Although FedAvg is commonly advertised as being private, recent results have demonstrated attacks able to recover training data from the models [10]. To protect the nodes from such attacks, we utilize secure aggregation to ensure that a node who queried multiple model parameters from a subset of its neighbors only get to observe an aggregate of those models. The design of secure aggregation schemes is outside of the scope of this work but may be achieved for arbitrary networks by using Shamir's secret sharing scheme [11] as demonstrated in [8]. For our purposes, we assume that a node i queries a set $\mathcal{M}_i^{(t)} \subseteq \mathcal{N}_i$ of size M in round $t \in [T]$ and observes only the aggregate $\bar{w}_i^{(t)} = \sum_{j\in\mathcal{M}_i^{(t)}} \beta_j w_j$ where $\beta_j \geq 0$ satisfy $\sum_{j\in\mathcal{M}_i^{(t)}} \beta_j = 1$. Consequently, node i is presented with $C_i = \binom{|\mathcal{N}_i|}{M}$ different groups of nodes to choose among where we assume $|\mathcal{N}_i| \geq M$ for all $i \in [N]$. For example, in a fully connected network consisting of $K = 100$ nodes and secure aggregation schemes where $M = 2$ and $M = 3$, we have 4,851 and 156,849 different groups, respectively.

Multi-armed bandits. We have a challenging group-selection problem with many groups and few rounds. A node can only evaluate a group by its local accuracy, which is stochastic and non-stationary due to other nodes' actions. We use an online learning approach and model the problem for each node as an adversarial multi-armed bandit with C_i arms and T rounds [12].

The performance of a bandit algorithm is measured by pseudo-regret, which compares the expected rewards of the best arm and the algorithm. For adversarial bandits, the pseudo-regret per round decreases as $\mathcal{O}(\sqrt{C_i/T})$ [13]. This means a large C_i, as in our case, an algorithm cannot be expected to perform well in a few rounds. However, this assumes independent rewards; if rewards are dependent, pulling an arm can give information about other arms and reduce exploration [14].

In our problem, some groups share nodes. The number of groups that share u nodes with a given group is $\binom{M}{u}\binom{N-M-1}{M-u}$. For example, in a fully connected network with $N = 100$ and $M = 3$, there are 13,680 and 288 groups that share one and two nodes, respectively, with a given group. So, selecting one group out of the 156,849 could inform about 13,968 groups. To leverage this idea, we use of pseudo-rewards, as presented in [14].

Let the different groups available to node i be indexed from $1, \ldots, C_i$ and, w.l.o.g., let the reward from choosing group $j \in [C_i]$ at time t satisfy $r_j^{(t)} \in [0, 1]$. We define the pseudo-rewards $s_{l,j}^{(t)}(\alpha_j^{(t)}) \in [0, 1]$ as an upper bound on the expected reward on $r_l^{(t)}$ given that we observe $r_j^{(t)}$ for $j \in [C_i]$ and $l \in [C_i] \setminus \{j\}$. This is mathematically represented as:

$$\mathbb{E}\left[r_l^{(t)} | r_j^{(t)} = \alpha_j^{(t)}\right] \leq s_{l,j}^{(t)}(\alpha_j^{(t)}). \quad (2)$$

For $j = l$, we let $s_{j,j}^{(t)} = r_j^{(t)}$. Note that setting $s_{l,j}^{(t)}(\alpha_j^{(t)}) = 1$ for all $j, l \in [C_i]$, $l \neq j$ and $t \in [T]$, results in recovering the uncorrelated multi-armed bandit setting. Note that the inequality in (2) must be satisfied in order to achieve zero-regret asymptotically in the

number of rounds [14]. However, as our objective is to simply identify nodes with similar local data distributions within a fixed number of training rounds, the choice of pseudo-reward in (2) will mainly serve to trade-off between exploitation and exploration.

To use the correlated bandit framework in our setting, we notice that groups with large overlap have many parameters in common in the aggregation step, hence, it seems plausible that also their expected rewards should be closer than groups with less overlap. Therefore, we design the pseudo-rewards between two groups to be decreasing in the number of overlapping nodes. Furthermore, it is expected that the discrepancy in accuracy between groups with large overlap decreases over time, hence the time dependency in (2). Let $u_{l,j} \in \{0, \dots, M-1\}$ denote the number of overlapping nodes between group l and group j. We consider pseudo-rewards of the form

$$s_{l,j}^{(t)}(\alpha_j^{(t)}) = \min\left\{\alpha_j^{(t)} + \frac{q(t)}{u_{l,j}}, 1\right\} \tag{3}$$

where $q : [T] \to \mathbb{R}_+$ is a non-increasing function in time, i.e., $q(t_2) \leq q(t_1)$ for $t_2 > t_1$. We make this choice as the variance between node models is anticipated to decrease as models converge.

2.1 Private multi-armed bandits for node selection

In this section, we present our bandit algorithm for a node. For ease of notation, we omit the node index. Let $k^{(t)} \in [C_i]$ be the group chosen at time t and let $n_{k^{(t)}}(t)$ be the number of times it has been chosen. The reward from choosing group $j \in [C_i]$ is defined as $\mu_j(t) = \frac{\sum_{\tau=1}^{t} \mathbf{1}\{k^{(\tau)}=j\} r_j^{(\tau)}}{n_j(t)}$ and the pseudo-reward for group $l \in [C_i] \setminus \{j\}$ when group $j \in [C_i]$ is selected, is given by $\phi_{l,j}(t) = \frac{\sum_{\tau=1}^{t} \mathbf{1}\{k^{(\tau)}=j\} s_{l,j}^{(\tau)}(r_j^{(\tau)})}{n_j(t)}$. We reduce the problem size by selecting only competitive arms, i.e., arms whose minimum pseudo-rewards are higher than the maximum reward. To this end, we define the set of significant arms as $\mathcal{S}_i^{(t)} = \{j \in [C_i] : n_j(t) > t/N\}$ and let $\bar{k}^{(t)} = \arg\max_{l \in \mathcal{S}_i^{(t)}} \mu_l(t)$. The set of empirically competitive arms is defined as

$$\mathcal{A}_i^{(t)} = \left\{ j \in [C_i] : \min_{l \in \mathcal{S}_i^{(t)}} \phi_{j,l}(t) \geq \mu_{\bar{k}^{(t)}}(t) \right\} \cup \{\bar{k}^{(t)}\}. \tag{4}$$

Note that $\mathcal{A}_i^{(t)}$ is not monotonically decreasing in t as arms may be non-competitive in one round and competitive in the next. Once $\mathcal{A}_i^{(t)}$ has been obtained, an arbitrary multi-armed bandit algorithm may be applied over the set of arms. As we consider adversarial rewards, we employ the *Tsallis-Inf* algorithm that is known to achieve a pseudo-regret with the optimal scaling [15], where large $q(t)$ encourages exploration whereas a small $q(t)$ encourages exploitation.

3 Experiments

Our code is made available upon publication to encourage reproducibility [1]. All experiments were carried out on an Nvidia 3090 Ti GPU. We conduct experiments on various cluster configurations and employ the CIFAR-10 and Fashion-MNIST datasets, which are commonly used in the literature for decentralized machine learning evaluations on covariate and label shift, see Section 3.1 [16]. We follow previous work [5, 9] and assume a fully connected graph among the nodes. Our algorithm aims to find a sub-graph for each node that maximizes its local task performance. In other words, we want to find the best collaborators for each node based on its local, private data.

Baselines. In all experiments we use decentralized adaptive clustering (**DAC**) [5] as a baseline for comparison, as it is most similar to our work. In addition, we also make comparisons to a random gossip communication protocol (denoted **Random**) and an oracle (denoted **Oracle**) that has perfect information of cluster assignments and only communicates (randomly) within these. Moreover, we also compare with local training on the nodes where no communication is allowed (denoted **Local**).

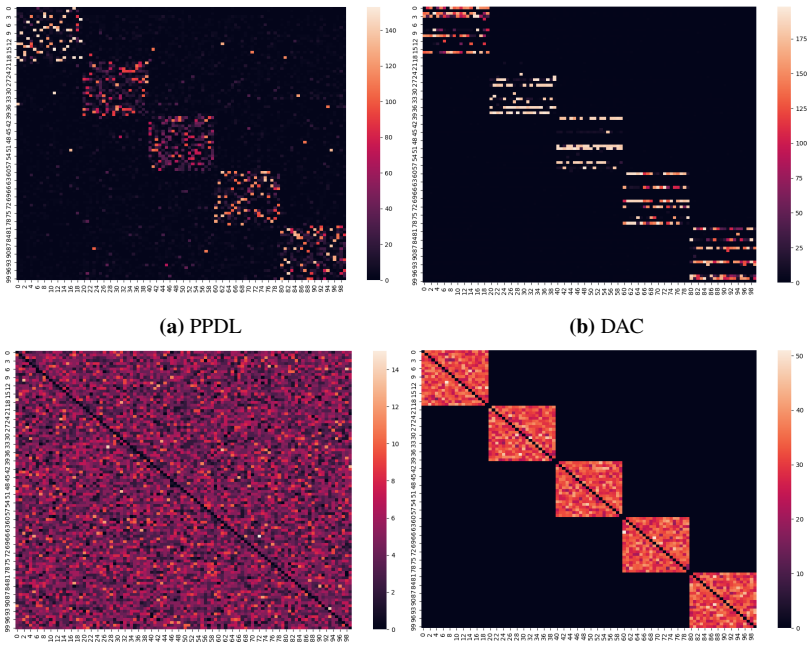
Covariate shift. To evaluate the performance of our method under non-iid data distributions, we replicate some of the experiments outlined in [5] for covariate shift with 100 nodes by dividing the data uniformly into four partitions, each with images rotated $0°, 90°, 180°$ and $270°$, respectively. We also experiment with heterogeneous cluster sizes by dividing the data into clusters of $0°, 180°, 350°$ and $10°$ rotation, with $70, 20, 5$ and 5 nodes in each cluster, respectively.

Label shift. Moreover, we also conduct experiments on label shift. As in [5], for the CIFAR-10 dataset we divide the data into two clusters based on labels, one for animal images and one for vehicle images. Additionally, we extend the experiments on label shift where we partition the data such that each node only has two labels, and these labels are grouped into clusters of five, where each cluster contains 20 nodes with the same two labels.

In our experiments, we evaluate all models on a test set with the same distributional shift as the training set for each node in the network. This is because the goal is to solve the local learning task for each node as effectively as possible. We use early stopping locally on each node.

Model and data. We use the same CNN architecture as [5], with three convolutional and two fully connected layers. We simulate 100 nodes for CIFAR-10 and Fashion-MNIST, and average results over three runs. Each node has equal data samples, uses the Adam optimizer and batch size of 8, and samples $M = 3$ other nodes per round. We train for three local epochs and 200 rounds. We use two $q(t)$ in (3): constant (PPDL) and exponentially decaying (PPDL-var), tuned by validation. We also tune learning rates using a validation set.

[1] https://github.com/edvinli/ppdl



(a) PPDL **(b)** DAC

(c) Random **(d)** Oracle

Figure 1. Heatmaps visualising how often node x communicates with node y for the four different methods on the CIFAR-10 dataset with 5 clusters.

Table 1. CIFAR-10 label shift test accuracy with 5 clusters.

Method	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Mean
PPDL	74.70	68.51	73.78	77.74	74.60	73.87
PPDL-var	71.82	80.40	78.72	82.06	76.05	77.81
DAC	**79.41**	**76.83**	**78.52**	**80.58**	**76.94**	**78.46**
Random	68.90	63.31	66.70	69.98	69.92	67.76
Local	77.11	88.79	82.60	61.83	68.32	75.73
Oracle	88.65	91.25	84.69	81.54	79.62	85.15

Table 2. Test accuracies for covariate shift on CIFAR-10 and Fashion-MNIST, with varying node numbers per cluster (70,20,5,5). Mean values over clusters are also provided.

	CIFAR-10				
Method	0°	180°	350°	10°	Mean
PPDL	52.37	45.21	50.60	51.03	49.80
PPDL-var	**54.63**	**47.27**	**51.84**	**53.30**	**51.76**
DAC	53.70	47.73	52.84	51.35	51.41
Random	54.85	44.70	52.64	52.43	51.16
Local	34.06	31.64	29.92	32.68	31.91
Oracle	55.04	46.80	38.35	38.00	44.55
	Fashion-MNIST				
PPDL	**84.62**	**81.81**	**81.01**	**82.11**	**82.39**
PPDL-var	80.68	81.12	80.42	80.66	80.72
DAC	82.48	80.44	79.85	80.43	80.80
Random	84.26	79.61	78.42	78.99	80.32
Local	78.72	76.83	77.40	77.26	77.55
Oracle	83.00	81.93	79.01	79.76	80.93

Table 3. Test accuracies for covariate shift on CIFAR-10 and Fashion-MNIST, with the same number of nodes per cluster (25). Mean values over clusters are also provided.

	CIFAR-10				
Method	0°	90°	180°	270°	Mean
PPDL	43.48	43.31	43.73	43.10	43.40
PPDL-var	45.06	44.05	44.60	43.14	44.22
DAC	**45.21**	**45.08**	**45.18**	**45.78**	**45.31**
Random	41.35	41.19	42.39	41.46	41.60
Local	32.01	32.34	31.47	33.07	32.22
Oracle	49.47	49.66	49.57	48.43	49.28
	Fashion-MNIST				
PPDL	80.69	81.12	80.43	80.66	80.73
PPDL-var	**80.81**	**81.71**	**82.36**	**80.19**	**81.26**
DAC	78.83	79.51	78.69	79.02	79.01
Random	80.20	80.72	79.3	79.99	80.05
Local	78.84	79.36	79.98	77.04	78.81
Oracle	82.86	83.18	84.25	83.79	83.52

Table 4. CIFAR-10 label shift test accuracy with 'animal' and 'vehicle' clusters.

Method	Vehicles	Animals	Mean
PPDL	51.86	36.31	43.81
PPDL-var	**52.86**	**36.33**	**44.60**
DAC	52.78	33.87	43.32
Random	44.79	30.00	37.40
Local	51.10	35.11	43.11
Oracle	57.17	39.74	48.45

Covariate shift. Tables 2 and 3 show the results of our covariate shift experiments with two cluster setups. Our method, PPDL, performs similarly to DAC, but with secure aggregation for privacy. Random favors large clusters and penalizes small ones, as seen in Table 2. DAC and PPDL avoid collaborating with "poisonous" nodes by their sampling schemes, improving test accuracy in the 180° cluster. Oracle has low test accuracies for small clusters, likely due to limited data (only 5 nodes per cluster). For the smallest clusters, 350° and 10°, PPDL and DAC find similar nodes in the large 0° cluster, improving performance. Thus, DAC and PPDL increase performance and fairness for smaller clusters that differ from large ones. The Fashion-MNIST results are less different between methods, likely due to the easier problem than CIFAR-10. Also, rotating images may not be challenging for small CNNs, as they can learn rotation-invariant representations with enough data. We analyze harder label shift problems ne

Label shift. The results of our label shift experiment with two clusters (animals and vehicles) are presented in Table 4. We observe that both PPDL and DAC perform well, with PPDL achieving superior results. The highest accuracy is achieved with PPDL-var, in which $q(t)$ is exponentially decayed. We note that Random performs worse than local training without collaboration, likely due to model poisoning caused by nodes communicating with incorrect clusters. For Random, the node models learn different representations for the different clusters, and when merging models from two distinct clusters, the resulting model is inferior due to the significant dissimilarity between the models, a phenomenon known as *client drift*. Both DAC and PPDL are able to mitigate this problem by identifying useful collaborators.

The results of our five-cluster experiment on CIFAR-10 are presented in Table 1, where each cluster consists of two unique labels. We observe that Random performs worse than the Local baseline on average also in this setting. Our experiments also reveal a high degree of variance within a cluster for the Local baseline, which can be attributed to the small size of node data. In contrast, the PPDL and DAC methods perform comparably and are able to correctly identify beneficial collaborators, as depicted in Figure 1.

4 Related work

Decentralized learning. Previous studies have demonstrated the effectiveness of gossip algorithms, as highlighted in references such as [3, 17, 18]. Furthermore, collaborative gossip algorithms, where nodes possess distinct local tasks, have been investigated in the context of multi-task learning (MTL) as seen in [4, 19]. While gossip learning has been demonstrated to be effective in convex optimization, its application in non-convex optimization, which is required for training deep neural networks, has not been as extensively studied. One of the first works that explored the use of gossip-based optimization for non-convex deep learning was conducted on convolutional neural networks (CNNs) in [20]. The authors demonstrated that high accuracies could be achieved at low communication costs using a decentralized and asynchronous framework. However, it is important to note that gossip learning is not well-suited for non-iid settings, where several distinct learning objectives may be present. Indeed, a protocol based on random communication between nodes does not take into consideration the benefits of node selection during training.

In centralized FL, methods based on hard clustering [21–23] can efficiently identify node clusters, but they limit the collaboration of nodes to their own clusters. This prevents nodes from utilizing useful information from similar clusters in forming a global model. Recent works have advanced decentralized learning of deep neural networks on non-iid data. [9] used expectation-maximization, while [24] improved node selection and communication cost with gradient-based cosine similarity and model pruning. [25] identified similar nodes by empirical loss, but only allowed hard clustering. [5] proposed a decentralized adaptive clustering algorithm that used empirical loss similarity to discover beneficial peers, but without privacy guarantees for model weights. This probability vector is then used for sampling similar nodes in the next communication round for each node, allowing for soft cluster assignments and communication within the entire graph. Empirical results demonstrate the effectiveness of this method in identifying clusters of nodes and improving the performance of the models. Although this method identifies useful node collaborations, there are a lot of privacy risks as model weights are being shared without any privacy guarantees.

Secure aggregation. Secure aggregation is a method to enhance node privacy in FL by protecting against server inference attacks [7, 26]. The idea relies on random masking of the node models, before uploaded to the server, such that the masks cancel out when models are aggregated. Extensions based on secret sharing schemes [11] have been proposed, e.g., [27]. For decentralized learning, where the communication topology may be arbitrary, only few works have considered privacy. One protocol for secure aggregation over arbitrary networks is presented in [8]. Specifically, for node i, the scheme consists of two phases: i) node $j \in \mathcal{N}_i$ broad-

casts a public key that is used to privately collect shares of a random mask generated by node $l \in \mathcal{N}_i$, ii) node i receives the masked models and the aggregated shares of the random masks at from each node and reconstructs the aggregated masks to recover the aggregated model. Note that all of the above schemes require the models to be mapped to a finite field, an operation that may impact the training. A step towards avoiding this step for secure aggregation over connected graphs was recently proposed in [28].

Multi-armed bandits. Random node sampling in FL and DL can be improved by biasing towards nodes' local losses [29]. Multi-armed bandits for node selection were introduced in [30] with rewards based on node latency and objective to minimize training time. Extensions for model averaging [31], asynchronous FL [32], and dropout and fairness handling [33] were proposed. However, multi-armed bandits may perform poorly when the number of arms is large or dependent. Dependency-based clustering [34] and pseudo-reward shrinking [14] are two methods to exploit dependencies and reduce the number of arms.

5 Conclusions and future work

We introduce **Private Personalized Decentralized Learning** (**PPDL**), a novel privacy-preserving node selection approach for personalized decentralized deep learning based on adversarial multi-armed bandits. Our approach uses secure aggregation to hide individual node metrics and exploits node dependencies to sample groups of collaborators efficiently. To the best of our knowledge, this is the first privacy-preserving node selection scheme for decentralized learning. We show that **PPDL** achieves comparable performance to existing (non-private) techniques on multiple experiments, while also providing privacy protection with secure aggregation.

For future research, it would be interesting to explore aggregation methods for models trained on different datasets in order to enhance the robustness of nodes to merging with other clusters. Another direction is to understand how privacy is affected by the number of nodes participating in the secure aggregation. Intuitively, as shown in, e.g., Section V.A in [8], privacy improves with larger group sizes.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Vol. 54. 2017, pp. 1273–1282.

[2] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. "Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[3] D. Kempe, A. Dobra, and J. Gehrke. "Gossip-based computation of aggregate information". In: *44th Annual IEEE Symposium on Foundations of Computer Science*. 2003, pp. 482–491.

[4] V. Zantedeschi, A. Bellet, and M. Tommasi. "Fully Decentralized Joint Learning of Personalized Models and Collaboration Graphs". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. 2020, pp. 864–874.

[5] E. Listo Zec, E. Ekblom, M. Willbo, O. Mogren, and S. Girdzijauskas. "Decentralized adaptive clustering of deep nets is beneficial for client collaboration". In: *FL-IJCAI'22: International Workshop on Trustworthy Federated Learning* (2022).

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[7] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. "Practical secure aggregation for privacy-preserving machine learning". In: *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*. 2017.

[8] K. Tjell and R. Wisniewski. "Private Aggregation With Application to Distributed Optimization". In: *IEEE Control Systems Letters* 5.5 (2021), pp. 1591–1596.

[9] Y. Sui, J. Wen, Y. Lau, B. L. Ross, and J. C. Cresswell. *Find Your Friends: Personalized Federated Learning with the Right Collaborators*. 2022.

[10] D. I. Dimitrov, M. Balunovic, N. Konstantinov, and M. Vechev. "Data Leakage in Federated Averaging". In: *Transactions on Machine Learning Research* (2022).

[11] A. Shamir. "How to Share a Secret". In: *Commun. ACM* 22.11 (Nov. 1979), pp. 612–613.

[12] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. "The Nonstochastic Multiarmed Bandit Problem". In: *SIAM Journal on Computing* 32.1 (2002), pp. 48–77.

[13] J.-Y. Audibert, S. Bubeck, et al. "Minimax Policies for Adversarial and Stochastic Bandits." In: *COLT*. Vol. 7. 2009, pp. 1–122.

[14] S. Gupta, S. Chaudhari, G. Joshi, and O. Yagan. "Multi-Armed Bandits With Correlated Arms". In: *IEEE Transactions on Information Theory* 67.10 (2021), pp. 6711–6732.

[15] J. Zimmert and Y. Seldin. "Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits". In: *J. Mach. Learn. Res.* 22.1 (July 2022).

[16] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. *Advances and Open Problems in Federated Learning*. 2019.

[17] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. "Randomized gossip algorithms". In: *IEEE transactions on information theory* 52.6 (2006), pp. 2508–2530.

[18] R. Ormándi, I. Hegedűs, and M. Jelasity. "Gossip learning with linear models on fully distributed data". In: *Concurrency and Computation: Practice and Experience* 25.4 (2013), pp. 556–571.

[19] P. Vanhaesebrouck, A. Bellet, and M. Tommasi. "Decentralized Collaborative Learning of Personalized Models over Networks". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 2017.

[20] M. Blot, D. Picard, M. Cord, and N. Thome. "Gossip training for deep learning". In: *arXiv preprint arXiv:1611.09726* (2016).

[21] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. "An efficient framework for clustered federated learning". In: *Advances in Neural Information Processing Systems* 33 (2020).

[22] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. *Three approaches for personalization with applications to federated learning*. 2020.

[23] F. Sattler, K.-R. Müller, and W. Samek. "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints". In: *IEEE transactions on neural networks and learning systems* 32.8 (2020), pp. 3710–3722.

[24] Z. Ma, Y. Xu, H. Xu, J. Liu, and Y. Xue. "Like Attracts Like: Personalized Federated Learning in Decentralized Edge Computing". In: *IEEE Transactions on Mobile Computing* (2022).

[25] N. Onoszko, G. Karlsson, O. Mogren, and E. L. Zec. "Decentralized federated learning of deep neural networks on non-iid data". In: *2021 ICML Workshop on Federated Learning for User Privacy and Data Confidentiality* (2021).

[26] J. H. Bell, K. A. Bonawitz, A. Gascón, T. Lepoint, and M. Raykova. "Secure Single-Server Aggregation with (Poly)Logarithmic Overhead". In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1253–1269.

[27] J. So, C. He, C.-S. Yang, S. Li, Q. Yu, R. E. Ali, B. Guler, and S. Avestimehr. *LightSecAgg: a Lightweight and Versatile Design for Secure Aggregation in Federated Learning*. 2021.

[28] K. Tjell and R. Wisniewski. *Privacy in Distributed Computations based on Real Number Secret Sharing*. 2021.

[29] Y. Jee Cho, J. Wang, and G. Joshi. "Towards Understanding Biased Client Selection in Federated Learning". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. 2022.

[30] W. Xia, T. Q. S. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu. "Multi-Armed Bandit-Based Client Scheduling for Federated Learning". In: *IEEE Transactions on Wireless Communications* 19.11 (2020), pp. 7108–7123.

[31] T. Kim, S. Bae, J.-w. Lee, and S. Yun. *Accurate and Fast Federated Learning via Combinatorial Multi-Armed Bandits*. 2020.

[32] H. Zhu, J. Kuang, M. Yang, and H. Qian. "Client Selection with Staleness Compensation in Asynchronous Federated Learning". In: *IEEE Transactions on Vehicular Technology* (2022).

[33] T. Huang, W. Lin, L. Shen, K. Li, and A. Y. Zomaya. "Stochastic Client Selection for Federated Learning With Volatile Clients". In: *IEEE Internet of Things Journal* 9.20 (2022), pp. 20055–20070.

[34] R. Singh, F. Liu, Y. Sun, and N. Shroff. *Multi-Armed Bandits with Dependent Arms*. 2020.

Fully Convolutional Networks for Dense Water Flow Intensity Prediction in Swedish Catchment Areas

Aleksis Pirinen[1], Olof Mogren[1] and Mårten Västerdal[2]
[1]RISE Research Institutes of Sweden
[2]Department of City Planning and Sustainability, Kungsbacka municipality
{aleksis.pirinen@ri.se, olof.mogren@ri.se, marten.vasterdal@kungsbacka.se}

***Abstract*— Intensifying climate change will lead to more extreme weather events, including heavy rainfall and drought. Accurate stream flow prediction models which are adaptable and robust to new circumstances in a changing climate will be an important source of information for decisions on climate adaptation efforts, especially regarding mitigation of the risks of and damages associated with flooding. In this work we propose a machine learning-based approach for predicting water flow intensities in inland watercourses based on the physical characteristics of the catchment areas, obtained from geospatial data (including elevation and soil maps, as well as satellite imagery), in addition to temporal information about past rainfall quantities and temperature variations. We target the one-day-ahead regime, where a fully convolutional neural network model receives spatio-temporal inputs and predicts the water flow intensity in every coordinate of the spatial input for the subsequent day. To the best of our knowledge, we are the first to tackle the task of *dense* water flow intensity prediction; earlier works have considered predicting flow intensities at a sparse set of locations at a time. An extensive set of model evaluations and ablations are performed, which empirically justify our various design choices. Code and preprocessed data have been made publicly available at `https://github.com/aleksispi/fcn-water-flow`.**

I. INTRODUCTION

As climate change intensifies, hydrological conditions will change. This will manifest itself both in the form of water shortages and as flooding in cases of intense precipitation. According to the Swedish Environmental Protection Agency, the climate in Sweden is becoming warmer and wetter [1], and municipalities are encouraged to increase their climate adaptation efforts, especially regarding mitigating the risks of, and damages associated with, flooding [12], [15]. The effects of climate change on rainfall-runoff will be more severe further north [14]. At the same time, the hydrological conditions in Sweden have been severely disturbed during the last two hundred years, with wetlands being drained and natural streams being straightened, which will further increase the effects of extreme weather events.

Hydrological modeling can shed light on the dynamics of water flow and how it is affected by various aspects of the environment. This can in turn allow for making informed decisions about the efficacy of nature-based climate change adaptation techniques such as wetland restoration, urban greening, and soil protection. Traditional hydrological models are based on expert knowledge and physical properties such as the preservation of volume, which have to be specified a priori. These work well for a certain domain if they are properly calibrated, but have difficulties generalizing to wider environmental categories [17]. Statistical data driven modeling, including machine learning (ML), is an alternative which has the potential to become more robust if it can be trained on a large enough dataset with a suitable learning signal. This way, not only can the flow intensity be estimated for any given water course following a heavy precipitation event, but also the response time of the given area, i.e. an estimation of the time lapse from the precipitation event to peak flow. Such information is vital to better understand flood risks and the effects of flood and drought mitigation, as well as general hydrological implications from changes in land use.

In this work we propose an ML-based approach for water flow intensity prediction that leverages the physical characteristics of a catchment area. We target the one-day-ahead regime, where a fully convolutional neural network [11] receives spatio-temporal inputs and predicts the water flow intensity at every coordinate for the subsequent day (the same modeling should however be able to handle other time horizons with minor modifications). Two important novelties of our proposed approach are:

- In addition to temporal data (past rainfall and temperatures), we include spatial data as inputs to the modeling, provided as satellite imagery and several derived GIS layers. This allows the model to build internal representations about relationships between temporal and spatial aspects of the local environment (including land cover, soil depth and moisture, and elevation).
- Using a fully convolutional model, we tackle the task of *dense* water flow intensity prediction, as opposed to only predicting flow intensities for a sparse set of spatial locations. To the best of our knowledge, we are the first to consider the dense prediction task.

The remainder of this paper is organized as follows. In Section II we provide a brief overview of the related work. Then, in Section III, we describe in detail the data we have used for modeling, training and evaluation. In Section IV we explain our approach for tackling the water flow intensity prediction task, and our proposed approach is empirically evaluated against alternative methods in Section V-B. Finally, the paper is concluded in Section VI.

II. RELATED WORK

Water flow prediction (also known as *stream forecasting*, or *rainfall-runoff modeling*) for rivers in the U.S. have been modeled using Long Short-Term Memory (LSTM [7]) networks [8], [5], [4], [6]. The modeling follows a traditional setup inspired by earlier physics-based hydrological models such as the U.S. National Water Model (NWM), based on WRF-Hydro [2]. Jia et al. [8] modeled river segments using an LSTM network with graph convolutions. One segment in the river network corresponds to a distance that the water flows during approximately one day. Input features include daily average precipitation, daily average air temperature, date of the year, solar radiation, shade fraction, potential evapotranspiration, elevation, length, slope, and width of each segment. Models were trained using a physics-informed setup where a traditional flow model acted as a teacher for the machine learning model. LSTM networks have also been used for post-processing the output from the NWM [5]. Similar to our work, most of these prior works have focused on next-day predictions. However, there are examples of hourly predictions [6].

Others have also employed convolutional neural networks (CNNs) for stream forecasting [3], [18], [13]. However, in contrast to us, these works do not incorporate spatial data from satellites or GIS, but instead model only the much lower-dimensional data (in single coordinates or very small neighborhoods, not entire areas as in our setup) provided as a feature vector for each time step, similar to the models using LSTMs. More broadly, deep learning has been used for many related tasks, such as groundwater level estimation [21], water quality estimation [19], and rainfall-runoff [10].

While some of the above mentioned works on stream flow estimation include information about the near environment (such as elevation and slope), none of them use detailed spatial information inputs as is proposed in our work. The use of fully convolutional neural networks to encode this information, in combination with traditional inputs such as rainfall and temperature, has the potential of representing more complex relationships and can result in a more detailed view of the near environment. It also enables us to perform *dense* water flow intensity prediction, different to prior works.

III. DATASET DESCRIPTION

We use data from 12 locations in Sweden, based on where the Swedish Meteorological and Hydrological Institute (SMHI) has stations for measuring weather and water flow data. These locations are Jönköping (Tabergsån), Knislinge (Almaån), Krycklan, Skivarp (Skivarpsån), Skövde (Ösan), Torup (Kilan), Tumba (Saxbroån), Dalbergsån, Degeå, Hässjaån, Lillån, and Lillån-Blekinge (see Figure 1).

In each location we have access to the following spatial data layers (see also Figure 2):

- satellite RGB image (Sentinel-2) from the Land Survey of Sweden (Lantmäteriet), 10m $\times$ 10m resolution;
- elevation map from the Land Survey of Sweden, 50m $\times$ 50m resolution;

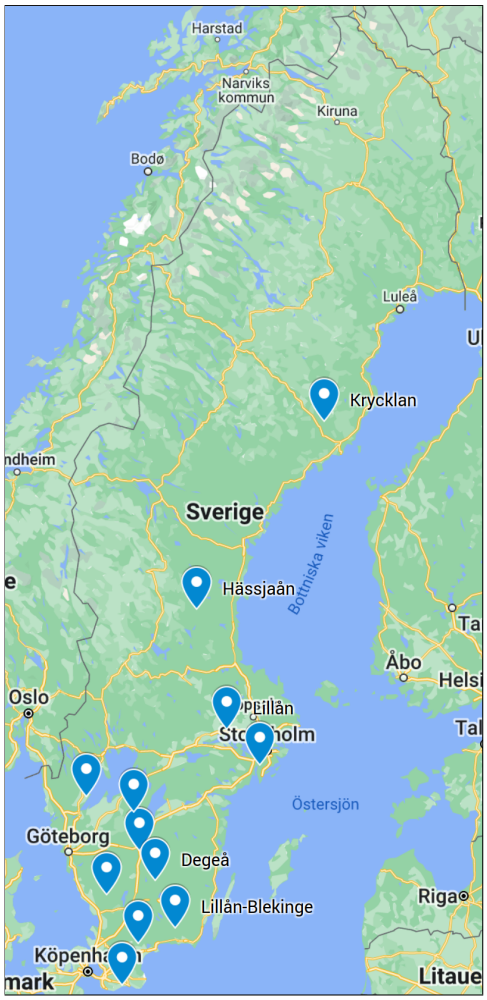


Fig. 1. The data used in this paper comes from 12 locations in Sweden: Jököping (Tabergsån), Knislinge (Almaån), Krycklan, Skivarp (Skivarpsån), Skövde (Ösan), Torup (Kilan), Tumba (Saxbroån), Dalbergsån, Degeå, Hässjaån, Lillån, and Lillån-Blekinge. Note that Lillån and Lillån-Blekinge are far apart (Lillån is north-west of Stockholm).

- terrain slope map from the Land Survey of Sweden, 50m $\times$ 50m resolution;
- soil moisture map the Swedish University of Agricultural Sciences, 2m $\times$ 2m resolution;
- land cover map the Swedish Environmental Protection Agency, 10m $\times$ 10m resolution;
- soil type map from the Geological Survey of Sweden, 10m $\times$ 10m resolution;
- soil depth map from the Geological Survey of Sweden, 10m $\times$ 10m resolution;
- hydraulic conductivity map from the Geological Survey of Sweden, 100m $\times$ 100m resolution;

The elevation map provides each coordinate's elevation above the ocean level, whereas the terrain slope map provides the slope of each coordinate (obtained by computing differences between adjacent coordinates of the elevation map).

All input data were provided as spatially aligned sets of raster maps, each of size 825×1244 pixels. For all locations

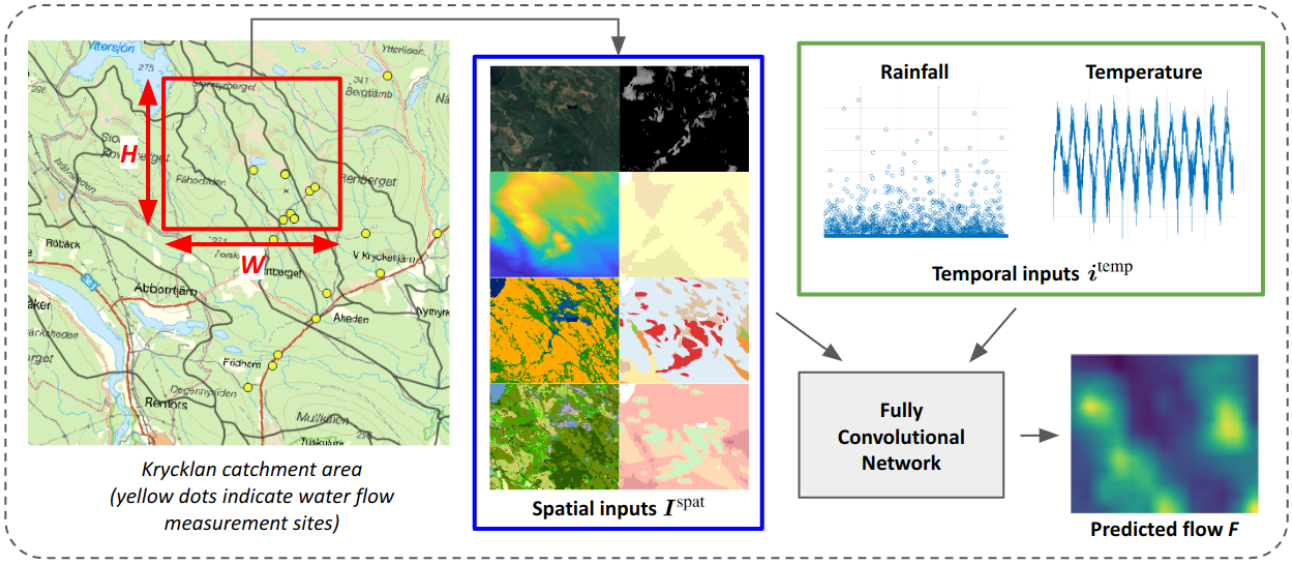


Fig. 2. Overview of our machine learning (ML) approach for dense water flow intensity prediction in catchment areas. The fully convolutional neural network receives both spatial and temporal inputs and produces a map of predicted water flow intensities. The spatial input $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$ represents relevant properties of the region (red rectangle, which can be located anywhere) in which to perform water flow intensity prediction (e.g. elevation map, soil moisture, land cover). The temporal inputs $\boldsymbol{i}^{\text{temp}} = \{r_j, \tau_j\}_{j=t-T}^{j=t-1}$ are the daily average rainfall amounts r_j and temperatures τ_j over the past T days, provided at times $t-T, t-T+1, \ldots, t-1$. The model then produces the flow map $\boldsymbol{F} \in \mathbb{R}^{H \times W}$, where each pixel (k, l) represents the model's estimate of the daily average water flow intensity at location (k, l) in the input map during day t. Note that our model is only trained in a very sparse set of coordinates (since water flow intensity measurement stations are very scarcely located), but the model nonetheless predicts flow intensities in every coordinate, not only in those for which contain flow intensity measurement stations.

except the Krycklan catchment area, there is exactly one water flow intensity measurement station. In Krycklan there are 14 such measurement stations, so in total there are 28 water flow intensity measurement stations in the dataset. In each measurement site, the daily average water flow intensity (m^3/sec) is provided. The specific length of each time series varies, but generally span several decades (in average about three decades).

In each location we also have access to two additional time series – daily cumulative rainfall (mm) and daily average temperature (degrees Celsius). These were obtained from the Swedish Meteorological and Hydrological Insitute (SMHI), and are often measured some distance away from the water flow intensity measurement sites (typically 1-3 kilometers).

It is common with missing measurements in the time series. For those time series which are used as model inputs (see Section IV), this is remedied by linearly interpolating the missing values between end points. Note that this is not done for the regressor (water flow intensity), since we only want the model to learn on and be evaluated on actual measurements.

Data preprocessing. We perform normalization of both the spatial and temporal data. Specifically, the spatial inputs are normalized to the $[0, 1]$-range by dividing with the maximum value (layer-wise) across all locations. A similar $[0, 1]$-normalization is performed for the temporal inputs (rainfall, temperature, and water flow intensity). We also tried another common normalization technique, where the variables are normalized to zero mean and unit variance, but empirically found that the $[0, 1]$-normalization works best in our setup.

IV. METHODOLOGY

In this section we provide an overview of the approach that we have developed for tackling the water flow intensity prediction task. See Figure 2 for an overview of our model and setup.

Our model leverages both spatial inputs $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$ and temporal inputs $\boldsymbol{i}^{\text{temp}} = \{r_j, \tau_j\}_{j=t-T}^{j=t-1}$ (cf. Section III), in order to predict water flow intensities f_t at time[1] t. The task of the model is to predict the water flow intensity $f_t^{h,w}$ at every coordinate (h, w) in a given geographical area of size $H \times W$ given $\boldsymbol{I}^{\text{spat}}$ and $\boldsymbol{i}^{\text{temp}}$. For the temporal inputs, we have chosen to only use readily available rainfall $r_{t-T}, \ldots, r_{t-1}$ and temperature data $\tau_{t-T}, \ldots, \tau_{t-1}$ for the past T days (with $T = 20$ in our setup), and not water flow intensity data $f_{t-T}, \ldots, f_{t-1}$, which is often unavailable in practice. In particular, note that water flow intensity is only measured at a very sparse subset of all coordinates in each location – and there are many locations in Sweden (and beyond) where no such measurement setups exist at all. Hence, for the model to be useful in a much larger set of contexts, it does not rely on past water flow intensities as input. However, in Section V we also compare with model variants that include past water flow intensities when predicting future flow intensities.

The spatial input $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$ contains relevant information regarding land and topological properties that

[1]Similar to most prior works, we target next-day prediction, but the model and approach can be extended to predict further into the future.

affect the water flow intensity in any given coordinate. These spatial input layers were introduced in Section III. In our setup we let $H = W = 100$, which corresponds to a real-world area of size 1km $\times$ 1km. The number of layers C is 10 in our case (three layers for the RGB satellite images, and one layer each for the other types of spatial input).

Since the task is to predict the water flow intensity in every coordinate in a map of size $H \times W$, based (in part) on spatial inputs of size $H \times W \times C = 100 \times 100 \times 10$, we have opted for a fully convolutional neural network[2] (FCN) [11]. This architecture expects a spatial input at one end, and gives a spatial output at the other end. To achieve this, we first concatenate the spatial and temporal data $\boldsymbol{I}^{\text{spat}}$ and $\boldsymbol{i}^{\text{temp}}$ into a unified input $\boldsymbol{I} \in \mathbb{R}^{H \times W \times (C+2T)}$. The first C channels are identical to $\boldsymbol{I}^{\text{spat}}$, while the last $2T$ channels are obtained by tiling rainfall $r_{t-T}, \ldots, r_{t-1}$ and temperature data $\tau_{t-T}, \ldots, \tau_{t-1}$ into $H \times W$-dimensional maps that are concatenated along the channel-dimension (each such map contains HW copies of a single value r_i or τ_i, for $i \in \{t-T, \ldots, t-1\}$). Given an input $\boldsymbol{I}$, the water flow intensity mapping is straightforward: $\boldsymbol{F} = g_{\boldsymbol{\theta}}(\boldsymbol{I})$, where $\boldsymbol{\theta}$ denotes the learnable parameters of the FCN g. We also evaluate and compare with other architecture variants in Section V.

A. Model Training

We randomly set aside 9 of the 12 data locations for training and 3 for validating the models. Specifically, the models are evaluated in Jönköping (Tabergsån), Hässjaån and Lillån, and trained on the other 9 locations; please refer to Section III for details about the dataset. In particular, note that Lillån-Blekinge is in the training set, but it is at a vastly different location than the validation set location Lillån (cf. Figure 1).

Each training input $\boldsymbol{I}$ in a batch is generated by randomly sampling a location and time period from the training set. Once a specific location has been randomly selected, we randomly sample a sub-region of size $H \times W$ which contains a water flow measurement site at the given location – see Figure 3. This results in the spatial input $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$. After having sampled a spatial location, we then concatenate the temporal information $\boldsymbol{i}^{\text{temp}}$ from a randomly sampled time interval (of T consecutive days), to obtain the input $\boldsymbol{I} \in \mathbb{R}^{H \times w \times (C+2T)}$.

There are roughly $25 \cdot 100 \cdot 100 = 250,000$ different spatial training inputs (25 measurement sites, and at each site there are roughly $100 \cdot 100$ possible locations for an enclosing rectangle of size $H \times W = 100 \times 100$ – see Figure 3). Note however that there is a spatial overlap between all different rectangles at a given site, which significantly reduces the data variability, compared to if all $250,000$ different rectangles would have come from different locations. The sites have on average roughly three decades of daily rainfall, temperature and water flow averages, which means there are $30 \cdot 365 \approx 11,000$ different temporal inputs per site. Hence, in total

[2]We use the *FCN8* model from the open-source FCN library [20].

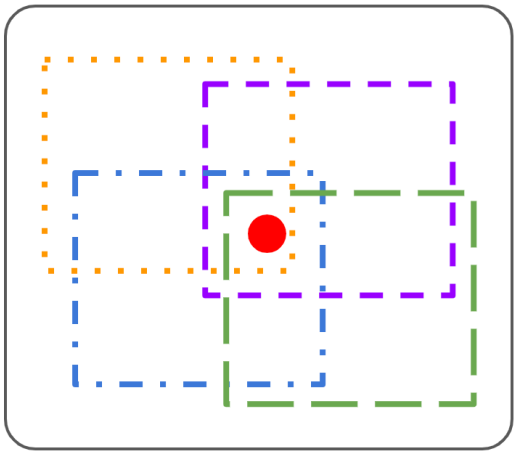


Fig. 3. Examples of possible sampled spatial locations (colored, dashed rectangles) that contain a water flow intensity measurement station (red dot). The random sampling increases the data variability (compared to e.g. always requiring the measurement station to be at the center). Since each possible spatial location has size $H \times W$, with $H = W = 100$ in our setup, the union of all possible such rectangles covers roughly 200×200 pixels, which corresponds to a real-world area of size 2km $\times$ 2km.

there are roughly $250,000 \cdot 11,000 = 2.75 \cdot 10^9$ spatio-temporal training inputs. Again, however, note that there is an overlap between a large majority of these training inputs, so the training set is effectively much smaller.

For model evaluation (see Section V), the end objective is to minimize the root-mean-square error (RMSE) of the predicted water flow intensities (thus note that the error is measured in m^3/sec), assessed based on ground truth flow intensities. During training, in order to balance loss smoothness with robustness to outliers, we use the Huber loss:

$$\mathcal{L}(f, f^{\text{gt}}) = \begin{cases} \frac{1}{2}\left(f - f^{\text{gt}}\right)^2 & \text{if } \|f - f^{\text{gt}}\| \leq \delta \\ \delta\left(\|f - f^{\text{gt}}\| - \frac{1}{2}\delta\right) & \text{else} \end{cases} \quad (1)$$

where f and f^{gt} denote predicted and ground truth flow intensity, respectively. We set $\delta = 1$ by default, as it is shown to result in the best performance (see Section V, where we also compare with other loss functions). The Huber loss can be seen as a combination of the commonly used MSE- and L1-losses, where the MSE-loss is applied when the error is smaller than the threshold δ, and the L1-loss is applied otherwise.

Note that for each predicted flow map $\boldsymbol{F}$, the loss (1) is only given for an extremely sparse set of coordinates (most commonly in a single point). This is because in the ground truth flow map $\boldsymbol{F}^{\text{gt}}$, we only have access to water flow measurements in very few coordinates (since the measurement sites are so sparsely located in the data). Despite this extreme loss sparsity, we show in Section V that the model generalizes well to unseen data. This finding is is in line with earlier works that have shown that it is possible to train semantic segmentation models from extremely few annotated pixels [16].

TABLE I

EXPERIMENTAL RESULTS ON THE VALIDATION SET FOR OUR MAIN MODEL, ITS ABLATED VARIANTS, AND BASELINES. WE REPORT THE ROOT-MEAN-SQUARE ERROR (RMSE; LOWER IS BETTER). COLUMN 1 REPRESENTS OUR MAIN FCN MODEL, CF. SECTION IV. COLUMNS 2-4 REPRESENT MODEL VARIANTS WHICH OMIT SOME OF THE SPATIAL INPUT LAYERS. COLUMNS 4-6 REPRESENT VARIANTS WHICH OMIT SOME TEMPORAL INPUTS. COLUMNS 7-8 REPRESENT BASELINE METHODS AGAINST WHICH TO COMPARE THE RESULTS IN COLUMNS 1-5. NOTE THAT *PREVIOUS FLOW* LEVERAGES PAST WATER FLOW INTENSITY INFORMATION THAT IS UNAVAILABLE TO THE OTHER APPROACHES.

Main model	**No-elev**	**Only-elev**	**No-soil**	**No-temp**	**No-rain**	**Half-time-hist**	**Mean-per-site**	**Previous flow**
1.35	4.74	3.22	4.49	1.91	5.91	1.40	2.05	0.59

For model parameter optimization, we resort to Adam [9] with batch size 64 and learning rate $2 \cdot 10^{-4}$. The model is trained for $250,000$ batches, which takes about 48 hours on the GPU-equipped (Titan V100) work station that is used for experimentation. To improve model generalization towards unseen data, we resort to the customary deep learning training technique of augmenting the data by horizontal and vertical flips of the inputs (an independent probability of 50% per flip).

V. EXPERIMENTS

In this section we present the results of our empirical model evaluations on the validation set. We first describe the various baselines and model variants in Section V-A. Then, in Section V-B, the empirical results are presented.

A. Baselines and Model Variants

We compare our main model described in Section IV against the following baselines:

- ***Mean-per-site*:** For each water flow intensity time series $\boldsymbol{f}^i = \{f_j^i\}_{j=t_1}^{j=t_{N^i}}$, where i indexes the i:th spatial location for a water flow measurement site, we return the mean $\hat{f}^i$ and use that as the predicted water flow intensity at time t (for each day t) at the i:th site. Note that this provides the optimal prediction (in terms of RMSE) in case only spatial information would be used as model input.
- ***Previous flow*:** Provides f_{t-1} as the predicted water flow intensity at time t. Note that this baseline leverages information that our model does not have access to; our model only obtains past rainfall and temperature information, not past water flow intensities.

We also train and evaluate the following variants of our proposed ML model:

- ***No-elev*:** This model omits the elevation and terrain slope maps from the set of spatial input maps.
- ***Only-elev*:** For the spatial part of the input, this model only uses the elevation and terrain slope maps. It omits the other spatial input layers.
- ***No-soil*:** This model omits the soil information spatial layers (soil type, soil moisture, soil depth, land cover) from the set of spatial input maps.
- ***Half-time-history:*** Uses temporal information from the past $T = 10$ days (instead of $T = 20$ as is default).
- ***No-temp*:** This model does not use temperature information as a model input.
- ***No-rain*:** This model does not use rainfall information as a model input.
- ***Flow-(t-k)*:** In addition to all the spatial and temporal inputs of our main model, this model has water flow intensity information $f_{t-T-k+1}, f_{t-T-k+2}, \ldots, f_{t-k}$ as an additional temporal input when predicting the water flow intensity f_t at time t. We train and evaluate models with $k \in \{1, 2, 3\}$, i.e. models that have temporal information up to between three and one day prior to the day for which flow intensities are predicted.

Finally, we also train and evaluate the effect of variations to the main model architecture (cf. Section IV):

- ***Alt-rain-temp*:** Uses a more efficient temporal input representation, which results in the input $\boldsymbol{I}$ having dimension $H \times W \times (C+T)$ instead of $H \times W \times (C+2T)$. This is achieved by having two unique values per temporal layer (instead of only one), where every second element (spatially) is a rainfall measurement, and every second element is a temperature measurement. Note that the convolutional filters (even the first one) will have sufficiently receptive fields to observe all relevant temporal inputs in this case as well.
- ***FC-early*:** Instead of performing a concatenation of the raw temporal data along the channel dimension, this model first processes the temporal inputs through two fully connected (FC) layers, the last of which produces a $20,000$-dimensional vector. It then reshapes this vector into size $H \times W \times C^{\text{temp}} = 100 \times 100 \times 2$ and concatenates with $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$. This data volume of dimension $H \times W \times (C + C^{\text{temp}})$ is then run through all the layers of the FCN, as for the main model.
- ***FC-mid*:** Similar to *FC-early*, this model first processes the temporal inputs through two FC layers, but here the resulting vector has dimension $2888 = 38 \cdot 38 \cdot 2$. It then reshapes this vector into size $H^{\text{mid}} \times W^{\text{mid}} \times C^{\text{temp}} = 38 \times 38 \times 2$. Different to *FC-early*, this model does not perform the concatenation with the raw spatial data $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$; instead it first processes $\boldsymbol{I}^{\text{spat}}$ through the first third of the convolutional layers of the FCN. It then performs the concatenation at this stage of the FCN, followed by joint processing for the remaining two thirds of the network.

B. Empirical Results

As mentioned in Section IV-A, we randomly set aside 9 of the 12 data locations for training and 3 for validating the models. The results of our experiments on the validation set

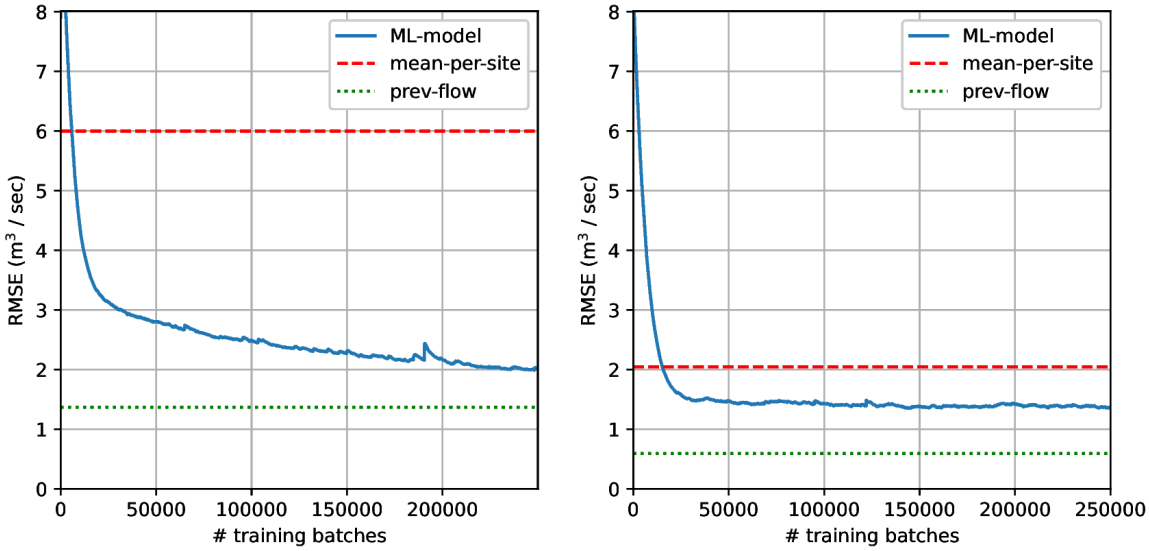


Fig. 4. Training (left) and validation (right) RMSE curves during model training for our main ML model (FCN) described in Section IV. It can be seen that the validation RMSE curve of our model flattens (marginally decreases) throughout training, i.e. the model does not begin to overfit on the training data despite the relatively small size of the training set. *Mean-per-site* and *previous flow* represent baselines, of which *previous flow* can be seen as an oracle that leverages past water flow intensity information that is unavailable to our model.

TABLE II

COMPARISON TO MODELS WHICH OBTAIN PAST WATER FLOW INTENSITIES AS INPUT. COLUMN 1 IS OUR MAIN MODEL THAT DOES NOT OBTAIN PREVIOUS FLOWS AS INPUT. PROVIDING PAST FLOW INFORMATION IMPROVES PREDICTION ACCURACY SIGNIFICANTLY, BUT NOTE THAT IN MANY CASES SUCH INFORMATION IS NOT AVAILABLE.

No flow	**Flow-(t-3)**	**Flow-(t-2)**	**Flow-(t-1)**
1.35	0.94	0.80	0.63

are shown in Figure 4 and Table I - IV. The evaluation metric that we report is the root-mean-square error (RMSE).

Due to the small size of the overall dataset (12 distinct locations), we have not yet considered a proper train-val-test split of the data. This will be done once more data of the appropriate type has been acquired. Currently however, when comparing other model variants and baselines to our main model, we report in the tables the *best* validation RMSE that was obtained during training of the respective model variant. Different to many of the alternative approaches, however, our main model's RMSE on the validation set monotonically improves throughout training (see Figure 4), and hence the results of the main model have not been 'cherry picked' at a certain optimal iteration number based on the validation set. Thus any reported improvements of the main model relative to alternatives may in fact be larger if assessed on a withheld test set.

Main results. It can be seen in Table I that our main model outperforms its ablated variants *no-elev*, *only-elev*, *no-soil*, *no-temp* and *no-rain*. In particular, the elevation and terrain slope maps are crucial, as is past rainfall information. Past temperature information is not as important, but omitting it still results in a higher error. Using rain and temperature information from the past $T = 10$ (instead of $T = 20$; see *half-time-hist*) days leads to similar results for this data.

Furthermore, our main FCN method is significantly better than the *mean-per-site* baseline, which indicates that our model has learnt to properly leverage spatio-temporal information. Our approach does however not outperform *previous flow*, which is a very strong baseline that leverages past water flow intensity information. Such information is not available to our model, and is often hard to come by in practical scenarios. Model variants which obtain previous water flow intensity information are however evaluated in Table II.

In Figure 4 we show training and validation RMSE curves during model training for our main FCN model. Note that the validation RMSE curve flattens (marginally decreases) throughout training, i.e. the model does not begin to overfit on the training data despite the relatively small size of the training set.

Effect of providing previous water flow intensity information as model input. As seen in Table II, models that receive flow information for $T = 20$ past consecutive days until three (*flow-(t-3)*), two (*flow-(t-2)*), or one (*flow-(t-1)*) day before the prediction day are significantly more accurate at predicting water flow intensity. Note however that in many practical scenarios such information is not available.

Effect of loss function. In Table III we compare the effect of using different loss functions during training; cf. (1). It is clear that the Huber loss (with $\delta = 1.0$ or $\delta = 1.1$) yields the best results, whereas the L1-loss results in the worst results.

TABLE III

LOSS ANALYSIS. THE HUBER LOSS YIELDS THE LOWEST RMSE, WITH $\delta = 1.0$ AND $\delta = 1.1$ BEING BEST. THE HUBER LOSS OUTPERFORMS THE MSE LOSS, AND THE L1 LOSS YIELDS POOR RESULTS.

Huber-1.0	MSE	L1	Huber-0.8	Huber-1.1
1.35	1.55	3.17	1.47	1.35

TABLE IV

MODEL ARCHITECTURE COMPARISONS. THE *FC-EARLY* AND *FC-MID* ARCHITECTURES YIELD SIGNIFICANTLY WORSE RESULTS THAN THE MAIN MODEL AND THE *ALT-RAIN-TEMP* ARCHITECTURE.

Main model	Alt-rain-temp	FC-early	FC-mid
1.35	1.47	2.57	2.49

Effect of model architecture. In Table IV we compare the effect of using different model architectures. The more efficient *alt-rain-temp* architecture yields almost as good results as our main architecture, so it would be suitable to consider if compute is a limiting factor. The *FC-early* and *FC-mid* architectures yield significantly worse results.

Qualitative examples. Several qualitative examples for our main model are shown in Figure 5 - 6. As can be expected, higher water flow intensities are typically predicted where the terrain slope is high.

VI. DISCUSSION AND CONCLUSIONS

In this work we have introduced a fully convolutional approach for dense water flow intensity prediction in catchment areas. Our specific results were shown for Swedish basins, but the general methodology is expected to be transferable to other geographical regions.

The proposed model is able to learn and generalize from a limited training dataset. In this work, we have used training data from merely 25 measurement points (28 in total; 3 were used for evaluation). The fact that we obtain such high performance may be attributed to the training setup. In particular, the model generalization is alleviated by the fact that the model sees many slight variations of each measurement site during training, since there are many ways to select a viewpoint around a given measurement site (cf. Figure 3). To the best of our knowledge, this is the first work which models water flow intensity using a fully convolutional neural network, which allows us to provide *dense* flow predictions – in effect, we predict one flow intensity per coordinate, even though we only have annotations for 28 specific coordinates.

Since our main FCN method is significantly better than the *mean-per-site* baseline, we conclude that our model has learnt to properly leverage spatio-temporal information. Our approach does however not outperform the *previous flow* baseline, which is a very strong baseline that leverages past water flow intensity information. Such information is not available to our model, and is often hard to come by in practical scenarios. As can be seen in Table II, the FCN model variants which utilize past flow information also obtain better results. A potential avenue of future work is thus to consider our setup through a privileged learning lens, wherein flow information could be leveraged during training, but where the model must perform inference using only rainfall and temperature information (in addition to spatial information).

We hope that our work will serve as a solid stepping stone and an inspiration for further research within dense water flow modeling, which in turn could deliver useful information when it comes to future climate adaptation planning (e.g. within flood risk management) in Sweden and beyond.

REFERENCES

[1] Claes Bernes. En varmare värld: Växthuseffekten och klimatets förändringar-tredje upplagan, 2017.

[2] B Cosgrove, David Gochis, Edward P Clark, Zhengtao Cui, Aubrey L Dugger, Greg M Fall, Xia Feng, Mark A Fresch, Jonathan J Gourley, Sadiq Khan, et al. Hydrologic modeling at the national water center: Operational implementation of the wrf-hydro model to support national weather service hydrology. In *AGU Fall Meeting Abstracts*, volume 2015, pages H53A–1649, 2015.

[3] Shiheng Duan, Paul Ullrich, and Lele Shu. Using convolutional neural networks for streamflow projection in california. *Frontiers in Water*, 2:28, 2020.

[4] Jonathan M Frame, Frederik Kratzert, Daniel Klotz, Martin Gauch, Guy Shelev, Oren Gilon, Logan M Qualls, Hoshin V Gupta, and Grey S Nearing. Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13):3377–3392, 2022.

[5] Jonathan M Frame, Frederik Kratzert, Austin Raney, Mashrekur Rahman, Fernando R Salas, and Grey S Nearing. Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, 57(6):885–905, 2021.

[6] Martin Gauch, Juliane Mai, and Jimmy Lin. The proper care and feeding of camels: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, 135:104926, 2021.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[8] Xiaowei Jia, Jacob Zwart, Jeffrey Sadler, Alison Appling, Samantha Oliver, Steven Markstrom, Jared Willard, Shaoming Xu, Michael Steinbach, Jordan Read, et al. Physics-guided recurrent graph model for predicting flow and temperature in river networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 612–620. SIAM, 2021.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[12] K Mossberg Sonnek, A Carlsson-Kanyama, and C Denward. Värmens påverkan på samhället–en kunskapsöversikt för kommuner med faktablad och rekommendationer vid värmebölja. *Myndigheten för samhällsskydd och beredskap, Report No.: MSB870*, 2015.

[13] Ana Ramos Oliveira, Tiago Brito Ramos, and Ramiro Neves. Streamflow estimation in a mediterranean watershed using neural network models: A detailed description of the implementation and optimization. *Water*, 15(5):947, 2023.

[14] T Salmonsson. Assessing the impacts of climate change on runoff along a climatic gradient of sweden using persist, slu masters thesis, 2013, 2014.

[15] Lisbeth Schultze, Carina Keskitalo, Irene Bohman, Robert Johannesson, Erik Kjellström, Henrik Larsson, Elisabet Lindgren, Sofie Storbjörk, and Gregor Vulturius. National expert council for climate adaptation assessment report nr 1. 2022.

[16] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1687–1697, 2021.

[17] Wen-Ping Tsai, Dapeng Feng, Ming Pan, Hylke Beck, Kathryn Lawson, Yuan Yang, Jiangtao Liu, and Chaopeng Shen. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature communications*, 12(1):5988, 2021.

[18] Song Pham Van, Hoang Minh Le, Dat Vi Thanh, Thanh Duc Dang, Ho Huu Loc, and Duong Tran Anh. Deep learning convolutional neural network in rainfall–runoff modelling. *Journal of Hydroinformatics*, 22(3):541–561, 2020.

[19] Charuleka Varadharajan, Alison P Appling, Bhavna Arora, Danielle S Christianson, Valerie C Hendrix, Vipin Kumar, Aranildo R Lima, Juliane Müller, Samantha Oliver, Mohammed Ombadi, et al. Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrological Processes*, 36(4):e14565, 2022.

[20] Ketaro Wada. pytorch-fcn: PyTorch Implementation of Fully Convolutional Networks. `https://github.com/wkentaro/pytorch-fcn`, 2017.

[21] Andreas Wunsch, Tanja Liesch, and Stefan Broda. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, 25(3):1671–1687, 2021.

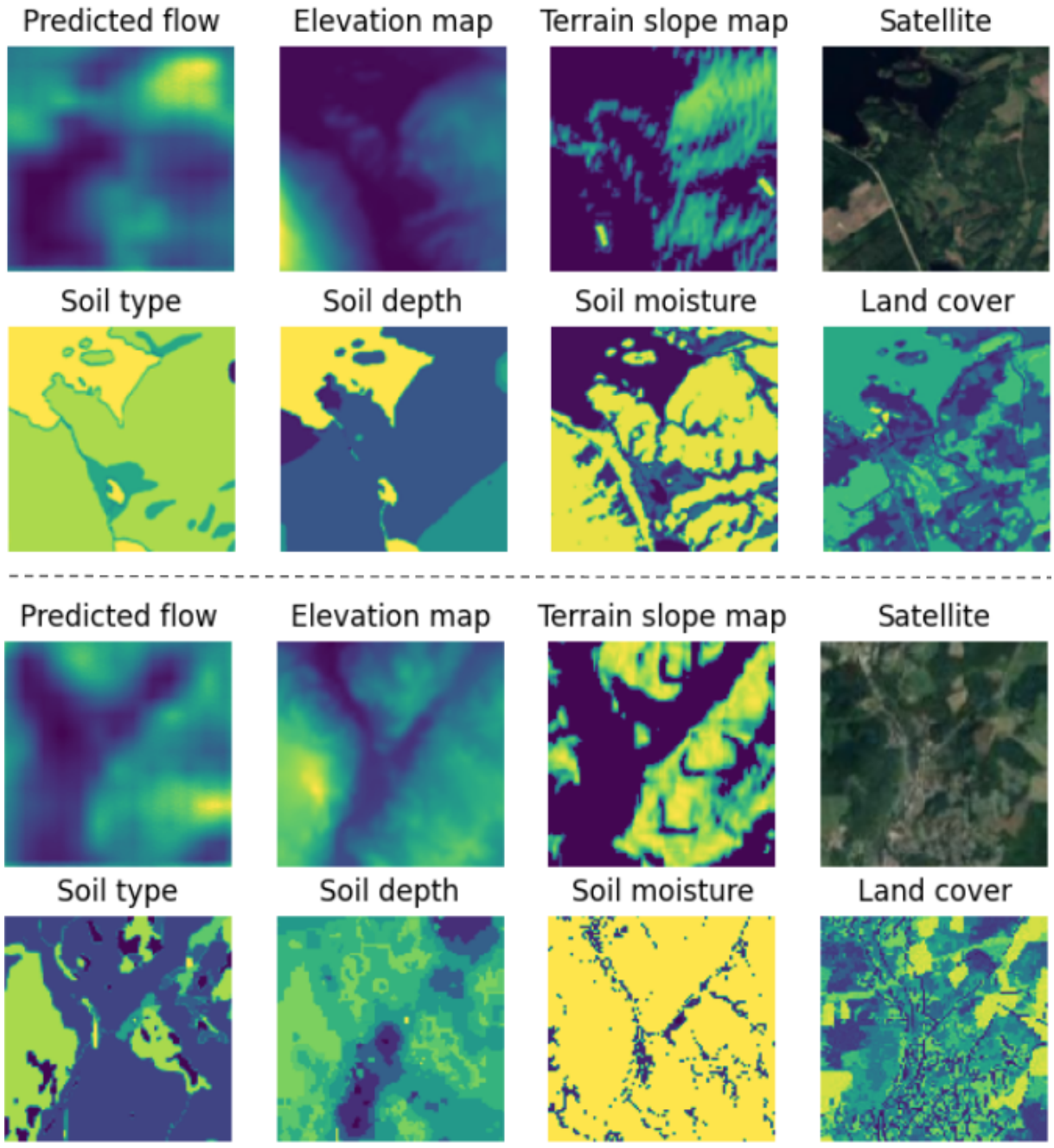


Fig. 5. Two qualitative examples for our main model on the validation set (examples differentiated by the dashed horizontal line). In each example, the top-left image represents the predicted water flow intensities at the given area; darker blue means lower intensity, while brighter yellow means higher intensity. The other seven images represent various spatial input layers to the FCN model. For all images except the satellite image, the maximum color intensity is individually normalized so that variations within images become as visible as possible. As can be expected, in both examples, higher flow intensities are typically predicted where the terrain slope is higher. In the example above the dashed line, the model also predicts relatively high flow intensities on the lake that can be observed at the top-left of the satellite image. Note however that the training set contains no ground truth water flow intensities on lakes, and thus the model has never been able to adapt to what is reasonable in terms of flow intensity on lakes.

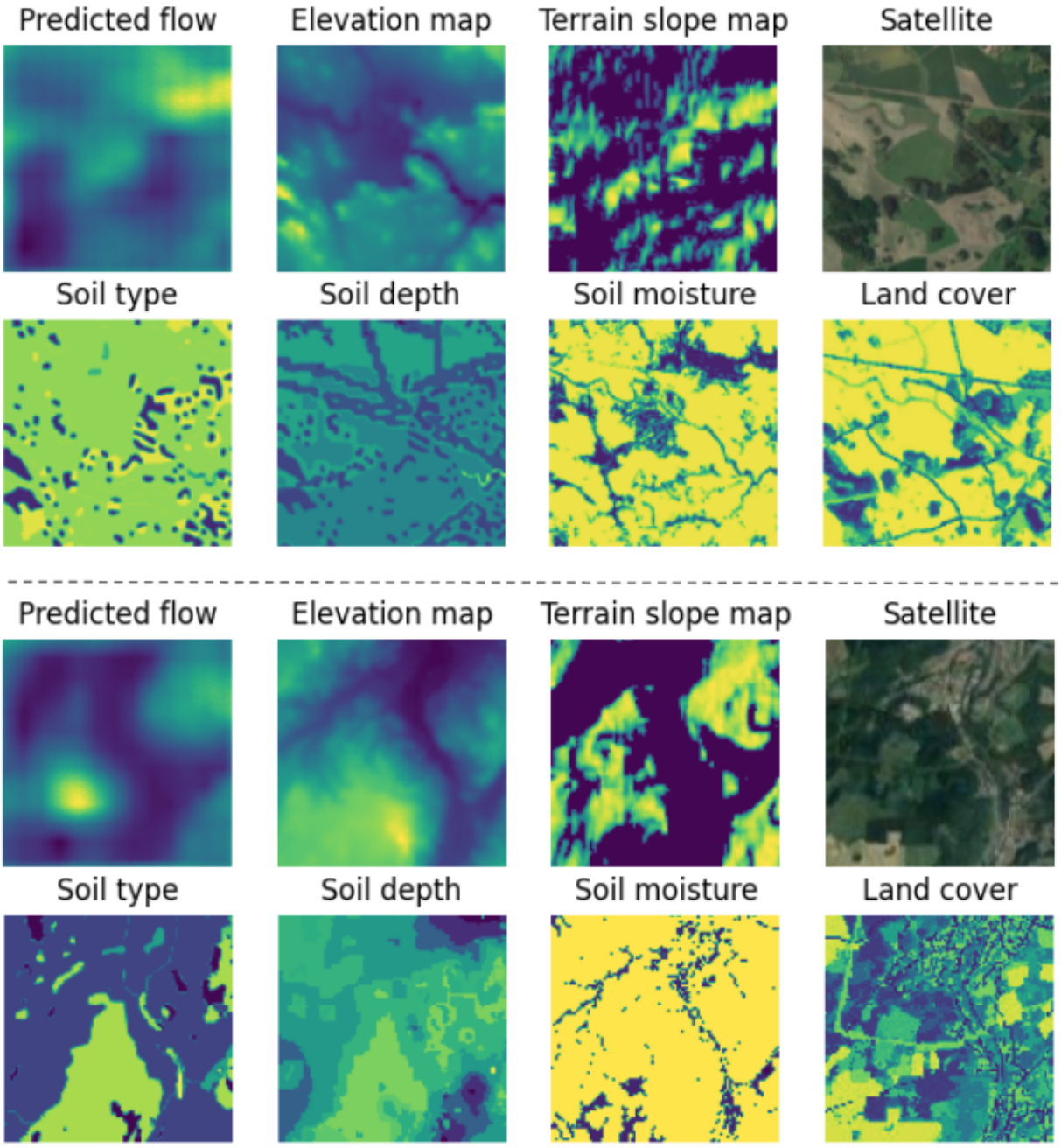


Fig. 6. Two additional qualitative examples for our main model on the validation set (examples differentiated by the dashed horizontal line). In the example above the dashed line, the predicted flow is moderately high within most of the map. A peak in terms of predicted flow can be seen at a corresponding peak within the terrain slope map. In the example below the dashed line, in can again be seen that the predicted flow is typically relatively higher where the terrain slope map is higher

FEW-SHOT BIOACOUSTIC EVENT DETECTION USING A PROTOTYPICAL NETWORK ENSEMBLE WITH ADAPTIVE EMBEDDING FUNCTIONS

Technical Report

John Martinsson[1,2*], *Martin Willbo*[1], *Aleksis Pirinen*[1], *Olof Mogren*[1], *Maria Sandsten*[2]

[1]Computer Science, RISE Research Institutes of Sweden, Sweden
{john.martinsson, martin.willbo,aleksis.pirinen,olof.mogren }@ri.se
[2] Centre for Mathematical Sciences, Lund University, Sweden
{maria.sandsten}@matstat.lu.se

ABSTRACT

In this report we present our method for the DCASE 2022 challenge on few-shot bioacoustic event detection. We use an ensemble of prototypical neural networks with adaptive embedding functions and show that both ensemble and adaptive embedding functions can be used to improve results from an average F-score of 41.3% to an average F-score of 60.0% on the validation dataset.

Index Terms— Machine listening, bioacoustics, few-shot learning, ensemble

1. INTRODUCTION

In few-shot bioacoustic event detection, the task is to predict the start time and the end time of certain bioacoustic events in a set of sound recordings from natural environments. The few-shot test set contains recordings for which only the first five examples of the bioacoustic event class of interest has been annotated, hence few-shot, and the goal is to detect the remaining events of this class in the rest of the recording. In figure 1 we show the setup of the task for one of the task recordings with predictions from our method. We are also given a base training set with annotated bioacoustic events of classes which are disjoint from the classes in the few-shot test set. During development, the few-shot test set is emulated using a supplied few-shot validation set where all events have been annotated as well, but where only the first five are used to infer the remaining events.

The need for solutions to this problem is motivated by the increasing amounts of audio data which are being recorded through acoustic monitoring devices, and where automated analysis is necessary to go through all of the collected data. Annotation efficient methods which can learn from very little annotated data is promising way forward.

The key contributions in this work are as follows:

- we train an embedding function to solve a multi-class sound event detection task since two different sound events can occur (partially) at the same time,
- we adapt the embedding function to the bioacoustic events we want to detect at inference time using the few-shot examples,
- we use ensemble predictions from multiple models trained on different time-frequency transforms to reduce false-positives, and
- we perform a comparison of three different time-frequency transformations.

*Thanks to the Swedish foundation for strategic research for funding.

2. METHOD

In this section we present our method, which is an ensemble of prototypical neural networks [1] with adaptive embedding functions. We describe how each embedding function is trained, how these are adapted using the few-shot examples, and how they are used to produce an ensemble prediction at test time.

2.1. Training embedding function

The base training data consists of sound recordings with annotations for 47 known event classes and one "unknown" event class. We are given the start and end times $\mathcal{A} = \{(s_i^k, e_i^k)\}_{i=1}^N$ of these classes, where (s_i^k, e_i^k) denotes the start and end time of sound event class k for annotation i. We model the 47 known sound event classes and the "unknown" sound event class in the same way, which yields a total of $K = 48$ classes. The goal is to learn an embedding function from this base training data. There is overlap in the annotations, i.e. two different sound events can occur (partially) at the same time, and we therefore treat this as a multi-class problem.

We assume a fixed length audio segment $x \in \mathbb{R}^T$ that consists of T consecutive audio samples is fed to the embedding function $f_\theta^T : \mathbb{R}^T \to \mathbb{R}^M$ (see section 2.4 for further details), where $M \ll T$. We split the audio recordings into audio segments $x_i \in \mathbb{R}^T$ by sliding a window of size T with a hop size of $T/2$ over each recording. For each audio segment x_i, a target vector $y_i \in \{0,1\}^{K \times n}$ is derived. If $n = T$ it means that the target contains one label per audio sample. Choosing $n < T$ means that the temporal resolution for the target is reduced. This results in a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ which defines the sound event detection task used to train the embedding function.

A prediction of the target classes for a given audio segment x_i is derived by $\hat{y}_i = h_\phi(f_\theta^T(x_i))$, where $h_\phi(\cdot)$ is a linear layer followed by an element-wise sigmoid activation function, and $f_\theta^T(\cdot)$ is a convolutional neural network where the first layer is a (non-learnable) time-frequency transform.

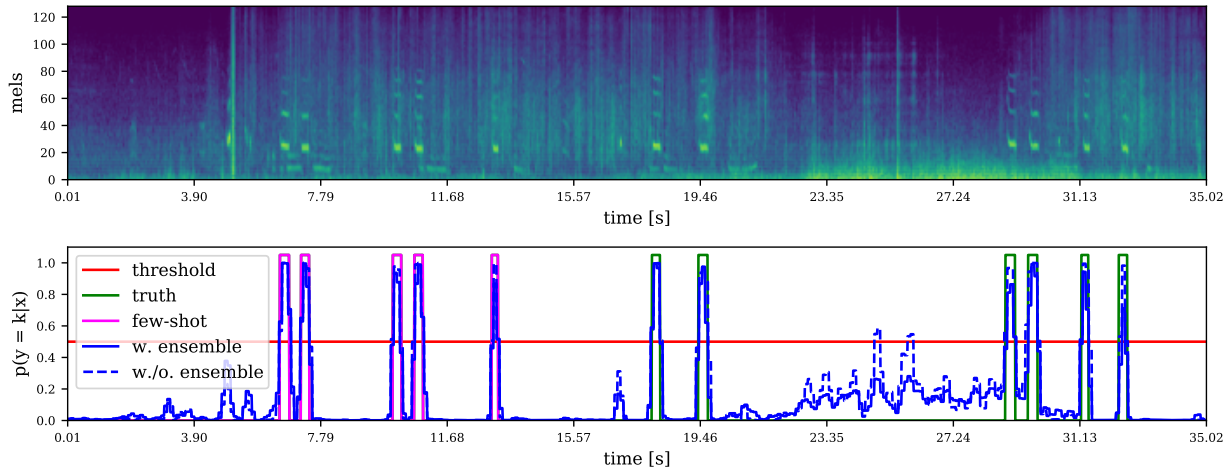


Figure 1: A log Mel spectrogram of part of a sound recording (top) and examples of predictions (bottom) from an ensemble prototypical network (solid blue line) and a prototypical network (dashed blue line) as well as the given few-shot examples (purple line) and remaining ground truth events (green line). The decision threshold τ of 0.5 (red line).

The loss function is the mean element-wise binary cross-entropy between the target y_i and the prediction $\hat{y}_i$, where the mean is taken over the class dimension K and the temporal dimension n.

For a fixed T, we train a set of C different embedding functions, all together parametrized as $\Theta = \{\theta_1, \ldots, \theta_C\}$, by varying the randomly initialized weights of the neural network, the training and validation split of the base training data, and the time-frequency transform in the first layer of the embedding function.

2.2. Prototypical network at test time

At test time we are given a sound recording and the $M = 5$ first examples of the class of interest. We denote these $A_p = \{(s_i, e_i)\}_{i=1}^{M}$ and call them the *positive* sound events. We assume that the gaps between these annotations are background noise and let $A_n = \{(e_i, s_{i+1})\}_{i=1}^{M-1}$ denote the start and end time of the $M-1$ first *negative* sound events. This is not necessarily true since an annotator may miss events, but we assume the likelihood of this to be low.

Let $l_i = e_i - s_i$ be the length of annotation i. If $l_i < T$ we "expand" the annotation with the $(T - l_i)/2$ preceding and subsequent audio samples to get an audio segment of length T and then we split this into segments of length T by sliding a window of size T over the signal with a hop size of $T/16$. Let S_p denote the set of positive audio segments derived from these annotated start and end times, and let S_n denote the set of negative audio segments. We use the embedding function f_θ^T and define the prototypes as

$$c_k = \frac{1}{|S_k|} \sum_{x \in S_k} f_\theta^T(x) \tag{1}$$

and derive a pseudo-probability of audio segment x belonging to sound class k from the prototypical network by

$$p_\theta(y = k|x) = \frac{\exp(-d(f_\theta^T(x), c_k))}{\sum_{k'} \exp(-d(f_\theta^T(x), c_{k'}))}, \tag{2}$$

where $k \in \{n, p\}$ and $d(z_i, z_j)$ denotes the Euclidean distance between z_i and z_j.

The query set S_q is derived by sliding a window of size T over the signal with a hop size of $T/2$. The reason for setting the hop size relative to T is that this means that we do equally many predictions for each audio sample in the validation recordings.

2.3. Our contributions

We now present the two main contributions of this paper: i) adapting the embedding function, and ii) using an ensemble of predictions.

Adapting the embedding function. We use the annotated positive events $A_p = \{(s_i, e_i)\}_{i=1}^{M}$ and compute the set of event lengths $L = \{e_i - s_i\}_{i=1}^{M}$. We choose $T \in \{T_1, 2^1 T_1, 2^2 T_1, 2^3 T_1\}$ such that $\sqrt{(T - l_{\text{min}}/2)^2}$ is minimized, where l_{min} is the shortest event length in L.

We choose $T_1 = 2048$ which is 0.09 seconds at a sampling rate of 22050 Hz so that we can plausibly detect the shortest events in the few-shot validation set. We limit the amount of memory needed during training and inference by only doubling up to three times. We have not extensively evaluated the effect of these choices and adding embedding functions trained on even shorter and longer segments may be beneficial, but of course comes at a computational cost during training.

Ensemble. Let $\Theta = \{\theta_i\}_{i=1}^{C}$ denote the set of parameters of C different adaptive prototypical network models. Then we define

$$p_\Theta(y = k|x) = \frac{1}{C} \sum_{\theta \in \Theta} p_\theta(y = k|x) \tag{3}$$

as in [2], which can be viewed as a uniformly-weighted mixture of experts. We say that x belongs to the positive event class if $p_\Theta(y = p|x) > \tau$ and the negative otherwise. This is done for every $x \in S_q$. Finally, if the query is classified as positive event then the start and end time associated with that query is used as the predicted positive event timings.

2.4. Details of the embedding function

The embedding function consists of a time-frequency transform followed by a convolutional neural network, both of which are briefly described below.

Time-frequency transform. The first layer of the embedding function is a time-frequency transform. Let $E(t, f)$ denote a Mel spectrogram with 128 Mel bins derived from an audio segment x. Then

$$\mathrm{S}(t, f) = 10 \log_{10} \frac{E(t, f)}{E_{\max}}, \tag{4}$$

and

$$\mathrm{PCEN}(t, f) = \left(\frac{E(t, f)}{(\epsilon + (E^t * \phi_T)(t, f))^\alpha} + \delta \right)^r - \delta^r. \tag{5}$$

We either use $\mathrm{S}(t, f)$ as the first layer embedding function, or we use one of two different parameter configurations for per-channel energy normalization (PCEN) [3], one of which is developed for speech audio and one of which is developed for bioacoustics as suggested in [4].

Convolutional neural network. The convolutional neural network used is an adapted version of the 10-layer residual neural network [5] implementation used in the baseline model for the challenge. Specifically, we i) add the classification head $h_\phi(\cdot)$ so that we can model the defined multi-class task, ii) use the same number of filters in every convolutional layer, and iii) reduce the max pooling along the time-dimension when audio segments are too short.

2.5. Post-processing

Since we get one prediction $x \in S_q$ of size T for each query audio segment, this limits how long predictions we can make with the model. To solve this, we simply merge all overlapping predicted positive events into one detected event with a single start and end time.

A predicted positive event will only be considered to be a match with a true positive event during evaluation if they have an intersection-over-union (IoU) of at least 0.3. We therefore remove predictions which are shorter than $0.3 * l_{\text{avg}}$ or longer than $(1/0.3) * l_{\text{avg}}$, where l_{avg} is the average event length of the given five annotations. Since predictions of these lengths can on average not be matched with true events as the evaluation is defined.

3. DATA

To highlight the types of variation that the model needs to handle, we have used the few-shot examples to compute the mean event length, the mean gap length, and the density of these sound events – see table 1. The mean event length is defined as the mean length of the five annotated events; the mean gap length is defined as the mean length of the *unannotated* gaps between the five annotated events; and the density is the sum of the time of the five annotated events divided by the total time spanned by the start of the first annotated event and the end of the last annotated event.

The validation set consists of three different subsets: HB, ME, and PB. We present the statistics for each subset in table 1. The HB subset contains long events with low noise, where the events are longer than the gaps between them, and as a result has a very high event density. The ME subset contains short events with low noise, where the events are shorter than the gaps between them, and as a result has a low event density. The PB subset contains very short events with very high noise, where the events are much shorter than the gaps between them, and has a result a very low event density.

Subset	Mean event length	Mean gap length	Mean density
HB	11.25 ± 3.11	6.12 ± 5.39	0.73 ± 0.12
ME	0.22 ± 0.03	1.40 ± 0.04	0.17 ± 0.02
PB	0.12 ± 0.08	59.89 ± 55.55	0.01 ± 0.02

Table 1: Validation data statistics.

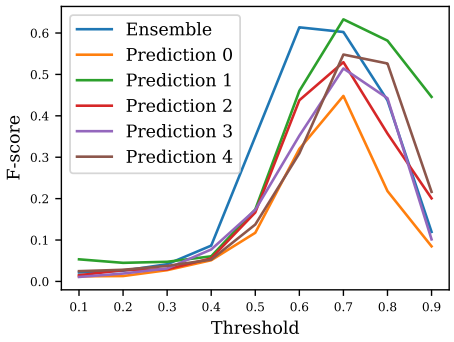


Figure 2: Comparing an ensemble of five predictions using embedding functions trained on PCEN (biodiversity) time-frequency transformations of the audio segments with each of these predictions itself.

4. EXPERIMENTS AND RESULTS

We have trained each embedding function using the Adam [6] optimizer with a learning rate of $1e-3$. The network is trained on a random split with 80% training data and validated on the remaining 20%. The training proceeds until we have observed no reduction in validation loss for the last 10 epochs and the model with the lowest validation loss is chosen as the final model. The temporal span of the targets have been fixed to $n = 16$, meaning that we have 16 targets for any given audio segment.

In figure 2 we compare the F-measure achieved on the few-shot validation set when using an ensemble of five predictions with using each of these predictions by themselves. The achieved F-measure by the ensemble is better than the best of these individual predictions for $0.4 \leq \tau \leq 0.6$, and outperforms or matches the mean of them for other τ. We also note that the optimal τ is around 0.7 for the single predictions, and moves to 0.6 for the ensemble.

In figure 3 we compare the F-measure achieved on the few-shot validation set for an ensemble over five predictions for each time-frequency transform with using an ensemble over both time-frequency transforms and the five predictions for each time-frequency transform (an ensemble over 15 predictions). We do not observe a significant increase in F-measure when comparing the ensemble to the ensembles using the PCEN (bioacoustic) time-frequency transforms, but it outperform the one using the PCEN (speech) and log Mel transform. The optimal threshold τ varies around 0.6 to 0.7 for the ensembles using a single transform, and is

at 0.6 for the ensemble.

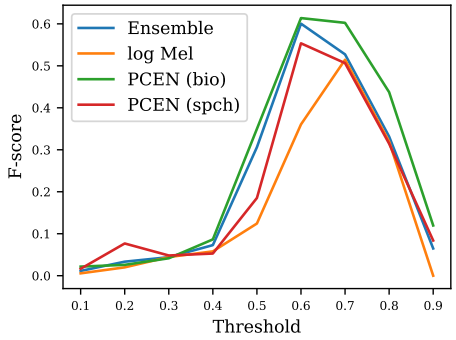


Figure 3: Comparing ensembles using embedding functions trained and tested on log Mel, PCEN (bioacoustics), or PCEN (speech), with an ensemble prediction of using all these time-freuency transforms.

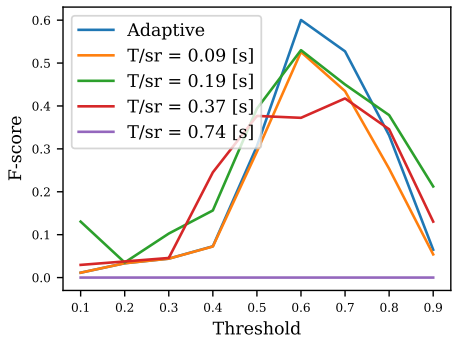


Figure 4: Comparing the adaptive embedding function with using each of the fixed size embedding functions respectively. sr denotes the sample rate, and T denotes the length of the audio segment.

In figure 4 we compare the F-measure achieved on the few-shot validation set when using the adaptive embedding functions in the ensemble with using any of the fixed $T \in \{T_1, 2^1T_1, 2^2T_1, 2^3T_1\}$. Adapting the embedding function increases performance from 53.0% (using best $T = 4096$) to 60.0% F-score for $\tau = 0.6$.

In table 2 we show an ablation where performance of a prototypical network using a embedding function (no ensemble) which has been trained on PCEN (speech) and a fixed segment length of 4096 is compared to using an adaptive embedding function, and then performing an ensemble prediction over the time-frequency transforms and random initialization of network weights and training/validation split (3×5). Adapting the embedding function increases the F-score on average with 8.3 percentage points, and adding the ensemble increases the F-score an additional 11.4 percentage points.

In table 3 we compare the results of the baselines with the F-score on the few-shot validation set for each of the systems for which we have submitted predictions on the test data for the challenge.

Method	Ensemble	Adaptive	F-score [%]
Ours	No	No	41.3 ± 3.8
Ours	No	Yes	49.6 ± 5.3
Ours	Yes	Yes	60.0

Table 2: An ablation of our system where we add adaptive embedding functions and ensemble.

Submission	τ	Ensemble + Adaptive	F-score [%] (valid)
Baseline (TM)	-	No	4.28
Baseline (PN)	0.5	No	29.59
Martinsson_1	0.6	True	60.0
Martinsson_2	0.5	True	30.6
Martinsson_3	0.6	False	44.6
Martinsson_4	0.5	False	13.3

Table 3: The validation scores for the baselines provided by the challenge organizers: template matching (TM) and prototypical networks (PN), and the validation score for the systems which have been submitted for test evaluation in the challenge.

5. DISCUSSION AND CONCLUSIONS

During development of this method we observed that random sampling of S_n, the set of negative examples, as in previous work does not work well for validation files with high event densities, which is why we chose to use the gaps between the first five annotated events instead.

We further observed that using one single fixed audio segment size T can be problematic. If T is much larger than the events we want to detect, the predictions will become too long to be counted as matches with the true events. Conversely, if T is much smaller than the events we want to detect, it may not cover the semantics which we want to detect. Therefore, choosing a T which matches the lengths of the events seem to be important. The adaptation of the embedding function could possibly be even stronger if based on both the event length statistic and the gap length statistic.

We observed that the optimal threshold was different for different validation files. We therefore try to calibrate the pseudo-probability in the predictions using an ensemble so that we get the best results by setting $\tau = 0.5$ as intended. We see from figure 2 and figure 3 that the optimal threshold τ moves from around $0.7 - 0.8$ to 0.6 which is what we want to achieve.

The ensemble improves performance by still correctly predicting most true positives, while no longer predicting as many false positives. This could intuitively be thought of as the ensemble being in agreement for true positive predictions, the average of which still yields a high pseudo-probability, while being in disagreement when predicting false positives, the average of which would be closer to 0.5. This effect can be seen in figure 1, where some of the false positives predicted when not using an ensemble (dashed blue line) are removed by using an ensemble of the predictions (solid blue line), leading to a reduction in false-positives.

In conclusion, we have shown that adapting the embedding function to the event lengths we want to detect can increase performance, and that false-positives can be reduced by an ensemble of predictions. We have also shown that out of the three time-frequency transforms we have studied, PCEN (bioacoustics) perform best for this bioacoustics task, followed by PCEN (speech) and then by log Mel.

6. REFERENCES

[1] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 4078–4088, 2017.

[2] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 6403–6414, 2017.

[3] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 1, pp. 5670–5674, 2017.

[4] V. Lostanlen, J. Salamon, M. Cartwright, B. Mcfee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-Channel Energy Normalization: Why and How," *IEEE SIGNAL PROCESSING LETTERS*, no. September, pp. 1–6, 2018. [Online]. Available: http://www.justinsalamon.com/uploads/4/3/9/4/4394963/lostanlen_pcen_spl2018.pdf

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Arxiv.Org*, vol. 7, no. 3, pp. 171–180, 2015. [Online]. Available: http://arxiv.org/pdf/1512.03385v1.pdf

[6] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," pp. 1–15, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

from the user to the producer (or seller) and enable higher utilization of products for the users.

Tukker (2004) argues that the economic potential of PaaS business models can be evaluated in terms of (i) tangible and intangible value for the user, (ii) tangible costs and risk premium for the provider, (iii) capital/investment needs, and (iv) issues such as the providers' position in the value chain and client relations. PaaS models therefore increase the incentive of the product company to capture value from product preservation (where the producer wants to keep the product attractive and in circulation for as long as possible) rather than product flow (where the aim is to sell as many products as fast as possible) (Stahel, 2010). However, selling functions rather than products can be perceived as a double-edged sword. On the one hand, it offers great potential for the product company to improve resource value preservation. On the other hand, it puts exhaustive demands on the balance sheet and cash-flows. Such challenges may inspire the perception that PaaS models are "just too difficult to implement" and therefore put a demand on more empirical evidence on PaaS as well as practical advice on how to make CBMs work (Kirchherr & van Santen, 2019).

As such, for businesses to transition from product-based to service-based CBMs, lack of access to financing and risk assessment tools to support change-in-ownership models is observed to be a key obstacle (Rizos et al., 2016). This highlights the importance of financial actors in facilitating such a transition by understanding and correctly assessing risks and potential of the new business models (ING bank, 2015; Toxopeus et al., 2021). However, *how* alternative risk assessment and financial solutions could look like to support CBMs remains unclear. Financial risk assessment models in CBMs would have to take into consideration a combination of factors regarding the long-term product value and market conditions and therefore should have the ability to collect large quantities of product and customer data. Digital technologies such as AI have the potential to make such models become feasible (Ellen McArthur Foundation, 2019), hence accelerate transition to CBMs.

The purpose of this paper is to explore financing solutions and innovations when moving from product-dominant business models to PaaS or function-based CBMs. In particular, two research questions that this paper addresses are as follows:

> **RQ1.** What different financial actors and solutions could enable PaaS-based CBMs in circular business ecosystems?
>
> **RQ2.** How can better predictions of residual values of products improve risk assessments for CBM?

We set out by describing our frame of reference in Section 2 and the method used in Section 3. We then present the results regarding financial solutions for circular business ecosystems in Section 4 and the results regarding financial risk assessments through AI-based predictions of asset residual value in Section 5. Section 6 offers a discussion followed by managerial implications in Section 7. Finally, Section 8 consists of some concluding remarks.

2 | THEORETICAL FRAMING

CBMs aim to improve resource efficiency—by extending the life spans of products and parts, by increasing the utilization of products, and by assuring recirculation of products and materials either through sourcing recycled material in production of a new item or by returning the product for recycling after its lifetime (Boyer et al., 2021). Previous literature often defines and categorizes CBMs in relation to how they improve resource efficiency by implementing the circular economy principles and strategies (Bocken et al., 2016; Linder & Williander, 2017; Nußholz, 2017). As such, Bocken et al. (2016, P.317) defines CBMs as "business model strategies suited for the move to a circular economy" based on the taxonomy of slowing, closing, and narrowing resource loops, or Lathi et al. (2018, P.3) propose a CBM definition "to explain how an established firm uses innovations to create, deliver, and capture value through the implementation of circular economy principles, whereby the business rational are realigned between the network of actors/stakeholders to meet environmental, social, and economic benefits."

For many major manufacturing and product selling companies, transition to a circular economy requires business model innovation as they need to rethink how they create, deliver, and capture value. CBM innovation therefore refers to the process of conceptualization, experimentation, and implementation of a new logic for creation, delivery, and capturing value (i.e., new CBM), which enables realizing environmental, social, and economic benefits (Chen et al., 2020; Frishammar & Parida, 2019; Lahti et al., 2018; Neligan et al., 2022).

Moving from the traditional, linear model of take-make-dispose to a circular model of make-use-reuse-remake-recycle, means that the firm creates value-in-use rather than in transaction and by bundling its products with advanced services to allow the products to be shared, repaired, upgraded, reused, refurbished, optimized, and eventually recycled (Frishammar & Parida, 2021). Kanda et al. (2021) discuss that business model as a firm-level unit of analysis has shortcomings in analyzing the industrial implementation of CBMs and argue for application of an ecosystem perspective as an appropriate concept to understand the high level of coordination required between different stakeholders necessary to implement circular systems. An ecosystem perspective can thus support innovation in the context of the circular economy where value is delivered through enhanced and new partnerships with ecosystem actors such as financial actors or service and technology providers. Moreover, this new logic increases the incentive for manufacturers to retain the ownership of their products and capture more long-term value from recurring revenues from leasing or rental fees combined with service contracts, instead of upfront payments.

Despite the sustainability and long-term economic benefits that the circular logic for value creation and capture (i.e., the CBM) posits, it poses challenges to standard financing solutions by (a) changing the

nature of the cash flow of the firm in that the cash flow is delayed and more long-term cost-effective financing is required to achieve scalability and (b) increasing capital volume needs to prefinance the products that will have a long-term and hence riskier return-on-investment if they fail in retaining value over time (ING bank, 2015). Therefore, financial barriers are identified as an important category of barriers to circular economy transitions (Adams et al., 2017; Govindan & Hasanagic, 2018; Grafström & Aasma, 2021; Kirchherr et al., 2018; Rizos et al., 2016; Tura et al., 2019; Vermunt et al., 2019). Henry et al. (2020) recognize that start-ups are more flexible and less path-dependent than incumbents and thus are relatively well-positioned to adopt CBMs. They also find, however, that start-ups with service-based model constitute only a small fraction (9%) of their start-up sample and that one of the reasons for this is the asset-heavy innovation needed in combination with lack of financial resources. Linder et al. (2022) found that PaaS-based CBMs are at a disadvantage in terms of bank financing due to significant challenges related to both collateral-based and business case-based credit assessments.

While previous literature mostly discusses how manufacturing companies seek alliances with specialized service companies, digital actors, and sub-suppliers for value creating and delivery in CBMs (e.g., Frishammar & Parida, 2019; Lieder & Rashid, 2016; Reim et al., 2021; Urbinati et al., 2020), not much has been discussed in terms of the nature of relationships required between financial actors and product companies in future ecosystems that enhance financing CBMs. To overcome the financial barriers, both product companies and financial actors need to reconfigure their roles and strategies in future business ecosystem. Financial institutions can contribute to transition to CBMs in two ways; first by helping manufacturers to make the transition to a circular economy on a financial level by providing the appropriate financial structure and services; and second by embodying the principles of the circular economy in their own thinking and updating their way of doing business and assessing risks (Accenture, 2018).

Toxopeus et al. (2021) suggest three financing strategies for product companies seeking CBMs: (1) reducing the uncertainty around the CBM by signaling the future cash flow through customer contracts and pre-orders (Frishammar & Parida, 2019; Linder & Williander, 2017), (2) Building relationships with financial actors, suppliers and customers to co-create financeable value proposition and delivery (Brown et al., 2020; Veleva & Bodkin, 2018); and (3) overcoming the difficulty of lending based on firm-specific assets by enabling asset-based lending for the CBM through standardization and modularity (Kirchherr et al., 2018) and creating secondary markets to allow for better pricing of the residual value of circular assets for banks.

While Toxopeus et al. (2021) offers one of the most forward-looking contributions in terms of financing strategies, they do not provide guidance for how these strategies can be implemented in practice and what circular financial solutions could look like. This paper further investigates required financing solutions for CBMs by collecting both the perspectives of the product companies and financial actors. It further illustrates through an experimental model how residual value of circular products can be predicted in alternative ways, improving the collateral-based risk assessments and the pricing of circular assets in future financial ecosystems. More particularly, the paper illustrates a model enabled by AI, which draws on open data from second-hand markets and predicts the second-hand price of a product that can be one indicator for residual value.

3 | METHOD

The paper draws on results from an empirical research project conducted between 2019 and 2022, which aimed to reduce uncertainties regarding future value of products and thereby increase the willingness among financiers to be part of the development of new CBMs. The aim of the study was set based on knowledge from previous studies that identified financing as a barrier to CBMs due to risks associated with predicting the residual value of products in a circular economy (Linder et al., 2022).

We build on analytical frameworks from strategy and business model literature. Besides the business model as a construct for understanding the logic of a firm for creating and capturing value (Fallahi, 2017; Teece, 2010), we employ business ecosystem as a complementary construct to understand the relationship between product companies and financial actors, service providers or customers that build the foundation for successful CBMs. We draw on Adner's definition of a business ecosystem as configurations of strategies and activities across a multilateral set of partners that need to interact in order for a circular value proposition to materialize (Adner, 2012, 2017).

To tackle the first RQ on what financial actors and solutions can enable PaaS-based CBMs, a series of 25 interviews were conducted (during 2020, 2021 and beginning of 2022) with actors along the business ecosystem, including eight financial actors, two OEMs from clothing and white goods industries, and six "circularity-enablers" offering digital and platform-based services such as insurance, sharing or second-hand marketplaces, and subscription financing solution, see Table 1.

We first selected two OEMs from two separate sectors, one from clothing and one from home goods, which have already released a PaaS CBM to gain perspective on challenges, developments, and already existing solution they have for financing their CBMs. Moreover, we selected two banks, one insurance company and one public credit institute, interested in working with CBMs. The initial semi-structured interviews with these seven actors focused on financing problem description from their perspective, opportunities they perceived and wishes they had with alternative solutions, how they were influenced by other actors in the business ecosystem as well as data access and data needs they had for financial risk assessment in CBMs.

To assure triangulation (Jick, 1979) later we expanded on number and type of financial actors to include also retail financiers, financiers with more experience of assessing CBMs and even specialized CBM financing firms, and we included six circularity enablers covering different parts of the circular business ecosystem. The focus of the interviews was on understanding circularity visions and strategies,

TABLE 1 Overview of interviews

Focal company	Circular economy focus	Role of respondent	Nr of interviews
Financier A	Merchant bank wanting to explore circular financing opportunities	Sustainability expert	3
Financier B	Merchant bank wanting to explore circular financing opportunities	Product manager asset finance	3
Financier C	Merchant bank with interest in circular financing	Head of sustainable finance	1
Financier D	Retail bank providing financing for subscription models	Sales manager	1
Financier E	Product insurance company with interest in circular risk and financing	Business developer	1
Financier F	Public credit institute for growth companies interested in exploring circular financing	Credit counselor	2
Financier G	Financing company offering subscription financing	Sales manager partner financing B2B	1
Financier H	Start-up financing company specializing in circular financing	Co-founder and CEO	1
OEM A	Outdoor garment with leading sustainability profile	CFO	2
OEM B	White goods provider, testing PaaS models and acquiring PaaS company	Business developer	2
		Director environmental & EU affairs	1
Enabler A	Financing provider for hardware-as-a-service	Co-founder and CEO	1
Enabler B	Platform for sharing of garments	Founder	1
Enabler C	Platform for sharing of children's clothing	Founder	1
Enabler D	Circular insurance provider	Founder andbusiness developer	2
Enabler E	Second-hand marketplace for wedding dresses	Founder	1
Enabler F	"Future price provider" for financing of IT equipment	Innovation lead	1

valuation of products in CBMs, financial risk assessment, and relevant data needed in future business ecosystems. Interviews were conducted face-to-face or through online platforms and lasted between 30 and 60 min. After each interview, interview notes were provided by the researcher(s) present at the interview and reviewed by the rest of the research team.

A compilation of interview notes and summaries using open coding was made afterwards and preliminary results were complemented with group discussions at a workshop in the beginning of 2021, with three OEMs and three financial actors (two banks and one public credit institute). The workshop used the business model canvas (Osterwalder & Pigneur, 2010) as a framework to map new value creation and value capture logic for a hypothetical financial actor that will offer financing services to manufacturing companies that sell PaaS in B2B or B2C contexts (for a detailed agenda of the workshop, see Appendix A).

Afterwards, the results from the interviews and the workshop were expanded to involve a broader range of PaaS companies through a survey about financing PaaS models. The survey was launched in spring 2021 and investigated the product characteristics and company's type of PaaS offer (as subscription, function sales, short term/long term rentals, and/or performance-based), ownership setup, turnover, maturity of the PaaS business model, presence of the Paas model in different markets, as well as needs for financing and financing solutions available. Moreover, the survey included questions on the following:

- Company's role in the business ecosystem (supplier, manufacturer, retailer, platform owner, service provider, or other)
- Company's financing situation in general (e.g., "How have you financed your PaaS model so far?" and "On a scale of 0-5, how big a problem is financing for your PaaS business?"),
- What type of solutions they need and what type of actors they could imagine working with (e.g., "If you think you need external funding for the next two years, what type of external funding would you prefer to use?" and "How do you view sharing financing risk with other players in your value chain, downstream or upstream? (e.g., customers, subcontractors, platform players)")
- Their view on how data and digital solutions could help them (e.g., "What kind of information/data do you think could make it easier for you to get financing and is this data available (yes or no)?").

The survey was designed to take 10–15 min to fill in and it was sent out (via e-mail or via LinkedIn) to 39 companies that had a PaaS business model in at least one product or category of products. Companies were chosen to ensure high sample diversity by representing different (a) industries, (b) type of business (B2B or B2C), (c) company size, and (d) type of PaaS model. Out of the 39 companies, 24 responded to the survey, which provided a high internal validity of the results (see Table 2 for an overview of the survey respondents). Besides structured analysis of the responses to the close-ended questions, responses from the open-ended questions generated further

TABLE 2 Overview of survey respondents

Industry	Type of business	Company turnover (MSEK 2019)	Type of PaaS model	PaaS as % of total sales
Garden, DIY, and home appliances	B2C	0	Subscription	100
Bicycles	B2C	0	Subscription, long-term rentals	90
Lighting	B2B	0.32	Functional sales, long-term rentals	100
Furniture	B2B	50	Long-term rental	20
Sports gear	B2C	190	Subscription, functional sales, short-term rental	1
Clothes	B2C	1	Short-term rental	50
Entrance mats	B2B	51	Functional sales	100
Home appliances	B2C	120,038	Subscription, functional sales	<1
Clothes	B2C	0.033	Subscription	100
Measurement systems	B2B	85,000	Functional sales, performance sales	2
Handheld tools	B2B	1200	Subscription, functional sales	22
IT equipment	B2B	2500	Functional sales, short-term rental, long-term rental	65
Packaging	B2B	0	Subscription, performance sales	100
Furniture	B2B	4.7	Subscription	100
Software	B2B	2	Subscription	10
Aquaponic equipment	B2B	0.5	Subscription, long-term rental	100
Software	B2B	2.5	Subscription	100
Signs	B2B	28	Functional sales	<10
Furniture	B2B	113	Functional sales, short-term rental, long-term rental	20
Coffee machines and vending machines	B2B	900	Functional sales	100
Sports	B2C	0	Short-term rental	40
Batteries	B2B	40,000	Subscription, long-term rental	Confidential
Camera and video equipment	B2B	2.6[a]	Subscription, short-term rental, long-term rental, rent-to-own	10
Housing	B2C	0.33[a]	Long-term rental	100

[a]Converted to SEK from DKK at the exchange rate 1,3.

input for categorizing alternative financial solutions relative to the type of PaaS company and its sector. Our sample is not large enough for claiming high external validity of our findings. However, the diverse set of actors in our sample in relation to industry/sector, size, and maturity of the PaaS businesses and the triangulation method used gives cause to claim a high level of applicability of our results.

To tackle the second RQ on how better predictions of residual value or products can improve financial risk assessment, we draw on arguments provided by Toxopeus et al. (2021) that uncertainties around the CBM can be reduced by signaling the future cash flows and in presence of secondary markets to allow for better estimation of residual values of circular assets. To better understand how residual value of products can be estimated in CBMs, we first (in the beginning of 2020) held a workshop with two OEMs, four financial actors (two banks, a public credit institute and an insurance company), and three technology enablers where the "Six thinking hats" method (de Bono, 1985) was applied. The focus of the workshop was on developing a vision for future circular and digital business ecosystems by asking the following questions:

Financier point of view:

- What type of information is critical for banks to be willing to take risks in new CBMs with new collateral?
- What are the most important aspects that digital technologies such as machine learning can add to risk assessment of collateral in CBMs, when historical data is missing?

OEM point of view:

- What type of product information could/should product companies share to support banks' risk assessments of the CBMs?
- How do the OEMs assess risk and opportunities in their own business?

- How would OEMs like the financiers to think and act to be able to expand their CBMs?

Enabling companies' point of view:

- What technologies are critical to enable financing of a transition to CBMs?
- How can "intermediary technology-based companies"/"innovation enablers" support this process?
- How could future business ecosystems look like?
- What other technology or enabling roles are there to fill?

A more detailed workshop agenda can be found in Appendix A.

Finally, to approach the question of estimating residual value of circular assets to reduce the risk in PaaS financing, experiments were set up to model residual value in used items. For this, data was collected regarding second-hand sales of used items in online auctions. The dataset contained 88,511 ads for items in the clothing categories of an online auction site in Sweden. The dataset was split into training, validation, and testing sets, and a number of machine learning models were trained and evaluated on the data. The ending price of auctions were used as the target predictive value. The aim was to obtain a trained model that could take information about used items (such as images, text descriptions and seasonal trends) and give accurate estimates for unseen items, to help estimating the value and risk of the PaaS business. As targets were rather sparse, and presumably containing substantial amounts of noise, the target values were discretised into different bins. These bins can be interpreted as price categories of the investigated products, ranging from low price items to high price items. See Appendix B for further description of the AI model and Table 6 for a summary of the price classes.

4 | FINANCIAL SOLUTIONS FOR CIRCULAR BUSINESS ECOSYSTEMS

Our results confirm the challenges and highlight opportunities and potential solutions for financing CBMs and the effect on roles and actors in the business ecosystem. The results from the survey show that the companies that struggle the most with financing (responding 4 or 5 on a scale from 0 to 5 representing difficulty with financing) are companies that want to scale relatively low-valued product, such as clothes and sports articles, and/or products that are new in the PaaS space and that do not have established second-hand markets, such as cameras, lighting, and aquaponics equipment.

The survey results also show that the majority of the respondents look for traditional financial actors to collaborate with, when scaling PaaS business models. The smaller start-ups often look for venture capital and other types of owner investments, and companies that sell (access to) relatively high-value products, such as coffee machines, IT equipment or office furniture, often see leasing companies as a natural partner. Several of the respondents also wish that their bank could offer a flexible solution for "PaaS scale up credit." A few of the respondents already use less traditional financing solutions involving actors in different part of the value chain (mainly suppliers, but also customers and retailers), and a clear majority of the respondents are positive to these kinds of collaborations.

Based on the lending technologies employed by banks to assess credit risk (Berger & Black, 2011; Berger & Udell, 2006) and earlier studies in the PaaS financing space (Linder et al., 2022; Toxopeus et al., 2021), we identify and categorize our results on financial solutions and opportunities for PaaS-based CBMs into three groups:

1. asset-/collateral-based, where the asset (product or contract) used as collateral can be liquidated by the financier in case of default of the product company, and the residual value thus realized.
2. business-case-based, where the loan repayment capacity of the product company is assessed through future business projections.
3. relationship-based, where the trustworthiness of the team behind the business is assessed, together with its collaboration partners.

Asset-based solutions can be enabled through standardized, modular, and adaptive product designs (Kirchherr et al., 2018), which keep the product attractive and retain its value over a longer period, hence allowing for better estimation of the residual value of circular assets for banks (Toxopeus et al., 2021). For products to be continuously attractive and thus attract customers with a sufficient willingness to pay, it is required that over time they are not only technically sustainable (do not break down) but also functionally (can be upgraded to new needs), esthetically (can withstand fashion fluctuations), and socially (what is acceptable and what works with current policies) sustainable (Nyström, 2019). This facilitates the dialogue with investors, where maintaining value over time for the products is important with regard to stable financial security.

Moreover, PaaS companies should preferably not grow faster than a second-hand market with residual value statistics would have time to be built up. Financiers find it easier both to assess the value of and—in case of default of the product company—to liquidate the products when an aftermarket exists. This indicates that there is a need to build an aftermarket, and that this might be easier in the presence of other industrial players offering similar products and services.

> "A large secondary market increases the opportunities for borrowing, even if the values are relatively low individually. But everything that can be sold on a secondary market is good and can in principle be mortgaged. Here, the residual values can be very important. It is important to understand these residual values over time."— Financier B

Contract lending based on large contracts and long contract periods can be an alternative to object financing. Short notice periods, on the other hand, are critical for customers in some service models, and would deteriorate the loan case. In those cases, the mass of

customers, for example in a subscription model, could, however, constitute a redundant inertia and thus a sufficiently secure mass of contracts for the financier. If the dropouts begin to exceed the number of customers, or the consumer behavior declines among existing customers, the company has time to gradually sell a corresponding proportion of the capital-binding objects. Contract lending is particularly suitable when the service is based on low-value products or where the services themselves are what create value rather than the hardware. Also, this scenario is simplified if there exist (several) other industrial players that could potentially take over the contracts of the product company, in case of default.

Leaseback is a solution where the product company sells the product to a finance company and then leases them back with interest and with a repurchase clause. The customer relationship stays with the product company, and the balance sheet value and the risk are then moved to the leaseback financier. A financier specializing in such credit solutions could be considered a more secure and less risky debtor than each product company by itself.

Lease-on-lease. Transfer the PaaS model in the value chain—either to suppliers or customers. Creating "leasing chains" is a way of transferring the risk along the value chain to where it could most easily be incorporated into (the balance sheet of) a running business, or to the final user (in B2C cases).

Stepwise loans within a larger loan frame, can be an alternative solution where a successive scale-up of credit is needed. A gradual upscaling of the business can allow a stepwise increase of the credit based on, for example, the total amount of subscriptions from subscribers, without needing to perform a full-scale credit check each time.

Business-case-based solutions are enabled by the product company showcasing the profitability and growth potential of the CBM (Frishammar & Parida, 2019; Toxopeus et al., 2021). The business model should preferably show greater potential than the linear model if the latter is still in operation. The depreciation rate and the time period for comparison should be picked in order to secure this. It is also good to present the risks associated with the linear business model, such as increasing concern for supply of raw material and inputs, changing customer preferences and more focus on sustainability and business solutions that tackle climate change. New ventures are always considered more risky than existing ones, but in a changing environment the risk of inertia is often underestimated.

The product company can signal high-quality forecasted revenue streams as they can estimate the year's sales figures in advance (existing monthly revenue from existing customer base plus new customers minus dropouts, so-called churn). This is an important aspect in dialogue with financiers so that a positive inertia in revenue streams or early warning signals such as a downward trend can be detected. The product company can also choose more in advance to use the future margin to increase growth (e.g., by investing more in communication) or to consolidate and reduce the growth ceiling (increase profit). A large number of customers (as in B2C) also creates redundancy and thus financial stability, which can be a considerable advantage, especially for business models with low-value products.

> "Some customers say that they wish these models with stable revenue streams were already in place. 'Too bad they were not in place before covid-19', they say."— Financier B

Moreover, it is important to reduce financial risks with controlled growth. Excessive exponential growth of business models that have their best profitability in the latter part of the product life cycle (when the product cost has been fully depreciated but the product is still attractive) leads to the potentially negative margins of new, yet unprofitable units overshadowing those that have become profitable. This is a risk particularly when a fast depreciation rate of the assets is applied, and where it could be mitigated with a balanced growth rate.

> "The problem with start-ups that want to grow fast: If they accumulate capital faster than they will earn the profit of older depreciated garments, then they will never reach profitability."— Financier F

A solution that is fundamentally different from the two solutions above, is to use the so-called LTV/CAC ratio (Lifetime Value/Customer Acquisition Cost) to convince the financier of coming profitability. Instead of relying on trustworthy forecasts based on the existing customer base, this solution speculates on the profitability of future customers. This could be particularly useful for small companies in the early stages, with highly attractive service offerings. The total future value from a new customer (LTV) can be, for example, three times greater than the cost of bringing in that customer (CAC) and then the customer is considered profitable even if the revenue comes later. If the ratio is positive, the company theoretically becomes richer the more it grows. For the financier, the risk should be reasonable if the ratio is positive, and even smaller if the company can ongoingly repay the capital plus interest, based on future income.

Relationship-based solutions include opportunities for collaboration and defining new roles in the business ecosystem, where credit services are developed together with and for new actors and for solutions combining several actors. Access to finance can be facilitated directly through relationship building with customers and banks (Toxopeus et al., 2021). Financing from customers can be in the form of crowdfunding or pre-payments, especially in the case of having an engaged and loyal customer base. In part, customer financing provides capital, but it also gives a signal of stability to future financiers if the company has a broad customer base with many owners crowdfunding the business.

> "Rental customers could be co-owners/micro-investors with kickbacks, mouth-to-mouth-method, references etc. (like when Uber started out in Sweden). This could also be a good basis for a bank loan 'on top'."— Financier A

Besides customers, building a close collaboration with the financier can also facilitate relationship-based financial solutions. In

addition to better understanding the business, revenue streams, and the value of the products over time, this provides opportunities for the financiers to explore and understand the roles they can play in the new circular business ecosystem, such as taking a position as a strategic advisor and to develop new credit products and services based on deep industry expertise. Involving the customer of the PaaS company in these dialogues could strengthen the case further. Closer collaboration between the PaaS company and the bank could allow for small-scale "in blanco loans", where the financier takes a deliberate risk with the purpose of developing the business and learning. A potential solution that has been identified in our study—and that could be seen as a next step after building competence and collaborating closely with customers—is to start industry-specific financing vehicles. As financiers gain a deeper understanding of an industry with its customers, products, and trends, specializing in financing solutions for that industry is a risk minimizing strategy.

> "There is a possibility that the development will be more towards niche financiers, who know their industry and/or their objects. It has always been like that, but it could develop even more. It could even develop towards a role that is not only a financier, but also part of the value chain."— Financier B

Customer relationships as well as specialized partner relationships can make the entire business case stronger and more stable. Today, there are companies that specialize in special insurance, recycling, repair as well as in providing (or realizing) the residual value of the product. A lot can be achieved in the short term in collaboration with them instead of trying to do everything in-house, which often requires a lot of time, focus, and resources and which may still not become equally good in the end. Table 3 summarizes the 15 different financial solutions presented in this paper.

5 | FINANCIAL RISK ASSESSMENT THROUGH AI-BASED PREDICTIONS OF ASSET RESIDUAL VALUE

Our workshop findings suggest that data on asset residual value are, together with data on the customer's customers (payment statistics, churn, lifetime value), the most important type of data needed to assess the risk of the credit case. The financial actors identify that both customer and contract data, which reflect the state of the business rather than the value of an asset, will become increasingly important as PaaS-based businesses become more common. There is still, however, a strong focus on asset value, which points at the importance of the AI-based predictions also carried out in the project. A thorough technical description of the AI-based modeling work can be found in Listo Zec et al. (2022). Our data-based research shows that with the help of AI it is possible to predict residual value data (e.g. price levels and how quickly the products can be sold) but that it takes time to train the intelligence for each product and industry and that the open source range of information is often limited to open second-hand market platforms, such as on-line auction sites.

> "Resale value of equipment can change over the term of the contract. Here, AI can help to continuously assess resale value."— Financier B

Results from the data-driven experiments carried out in this project to model residual value in used items showed that the online auction ads contained sufficient signal in user uploaded images and text descriptions to make coarse-grained predictions of the residual value of used items based on the existing data. We trained a multilayer perceptron (MLP) and a logistic regression and compared them using different data as input (text and/or images). The results are summarized in Tables 4 and 5. See Appendix B for more details of the modelling.

TABLE 3 Overview of financing solutions per category

Solutions/categories	Asset-/collateral-based	Business case-based	Relationship-based
Adaptive product design	X		
Build an aftermarket	X		
Contract-based lending	X		
Leaseback	X		
Leasing chains	X		
Stepwise loans	X		
Compare with and outperform the linear model		X	
Consider the linear risk		X	
Show high-quality revenue streams		X	
Controlled growth		X	
Show LTV/CAC ratio		X	
Customer financing			X
Closer collaboration with financier			X
Industry-specific financial actors			X
Specialized partner relationships			X

TABLE 4 Test accuracies for different machine learning models and representations of the four price classifications task

Model	Accuracy (%)
MLP (clip image)	49.32
MLP (clip text)	53.18
MLP (clip text + clip image)	54.37
Logistic reg (unigram)	54.12
Logistic reg (bigram)	56.11
MLP (bigram)	57.03
MLP (bigram + clip image)	57.40

TABLE 5 Test accuracies for different machine learning models and representations of the nine price classifications task

Model	Accuracy (%)
MLP (clip text + clip image)	33.86
Logistic reg (unigram)	34.33
Logistic reg (bigram)	36.08
MLP (bigram)	36.96
MLP (bigram + clip image)	37.2

Further, our analysis shows that in the rare event of a failure of price predictions, the model will still be close to the correct price range and does not over- or underestimate to a large extent. This can be seen in Figure 1, which shows a confusion matrix of the best performing model's predictions, and it is further emphasized in Figure 2, which shows the difference between true classes and predicted classes (see Table 6 for description of classes and price ranges). Each row in the confusion matrix represents the true class, while each column represents the predicted class by the model. It shows how often the model predicts the correct label (the diagonal) and how often it performs a classification error. A perfect model would have a score of 1.0 in the diagonal.

We also wanted to evaluate if the model could estimate the value of a stock of items. To do that, we ran the best performing model (i.e., the MLP) (bigram + image) through the 17,704 items in the test set and calculated the sum of the predicted price ranges. We discarded all 1773 items in class 8 (401+ SEK) since they did not have an upper range. The resulting estimate of the stock was 1,263,419–1,928,556 and the true value of these items is 1,711,444 SEK, which lies in the predicted price range. Our results thus show that the model is able to estimate also the value of a stock of items.

The results from the AI-based modeling experiments thus suggest that there is sufficient signal in the collected data to make coarse-grained predictions about price categories. While the data may have significant differences compared to data that can be envisioned being used in PaaS companies' stock inventories, the results indicate that given the correct data as input, it is possible to predict price ranges of used clothing items, both at individual and aggregated level. These results are particularly relevant to strengthen the asset-/collateral-based financial solutions (as per the categories in Section 4).

An evaluation comparing the AI model's performance to humans was also carried out. A questionnaire was created and answered by 37 humans, where 10 random images from the test set were chosen, and the question posed was "What do you think the end price of this auction was?" The alternatives were the nine different price categories. The results can be seen in Figure 3. The AI model accuracy, shown in orange, achieved a score of 40%. This can be compared to the mean human accuracy of 18.75% (green bar). Only two humans were equally good as the AI, and only one beat it. Seven humans got a score of 0%. Moreover, a majority human vote (red bar) only achieved 10%, not performing better than random chance (11.11%). These results indicate how hard it is for humans to estimate price ranges of used clothing items, and that an AI model is able to better asses this value.

6 | DISCUSSION

There is an array of different PaaS scenarios, including both subscription models, long- and short-term rentals and functional and performance sales. The types of companies that operate (and want to operate) PaaS business models are diverse and different from each other (for example in terms of size and planned speed of scaling), as are the products and services involved, implying that there is a need for a varied set of financing solutions. Earlier studies have listed the challenges of financing such PaaS-based CBMs (Linder et al., 2022) as well as the solution strategies (Toxopeus et al., 2021) along the lines of the lending technologies used by financiers, grouping them in asset-/collateral-based, business case-based, and relationship-based.

We position our findings in line with these three categories and further both confirm earlier research on challenges (Adams et al., 2017; Govindan & Hasanagic, 2018; Grafström & Aasma, 2021; Henry et al., 2020; Kirchherr et al., 2018; Tura et al., 2019; Vermunt et al., 2019) and add more concrete solution suggestions:

The ***asset-based solutions*** we propose include product design for value retention, actively building an aftermarket for products and services, using contracts as collateral and using credit products, such as leaseback, lease-on-lease, and stepwise loans. The ***business case-based solutions*** proposed include making direct comparisons with the linear model over a relevant period of time to show the advantage of the circular business case, stressing the linear risk and the high-quality revenue projections of the circular case. In addition to this, the growth rate might need to be controlled so that profitability can be made visible, and a more opportunistic approach is to convince the financier of future profitability through the so-called CAC-LTV ratio. We also suggest ***relationship-based solutions***, emphasizing the relationship between the PaaS-company and the financier, but also pointing at risk sharing with customers through prepayments and crowdfunding, as well as with actors in the supply chain. A truly deep understanding of a customer segment might develop into industry-specific financial actors. Involving other actors that can provide repair and refurbishing services as well as the realization of the residual value of the product, will also help providing trustworthiness in your case toward the financier.

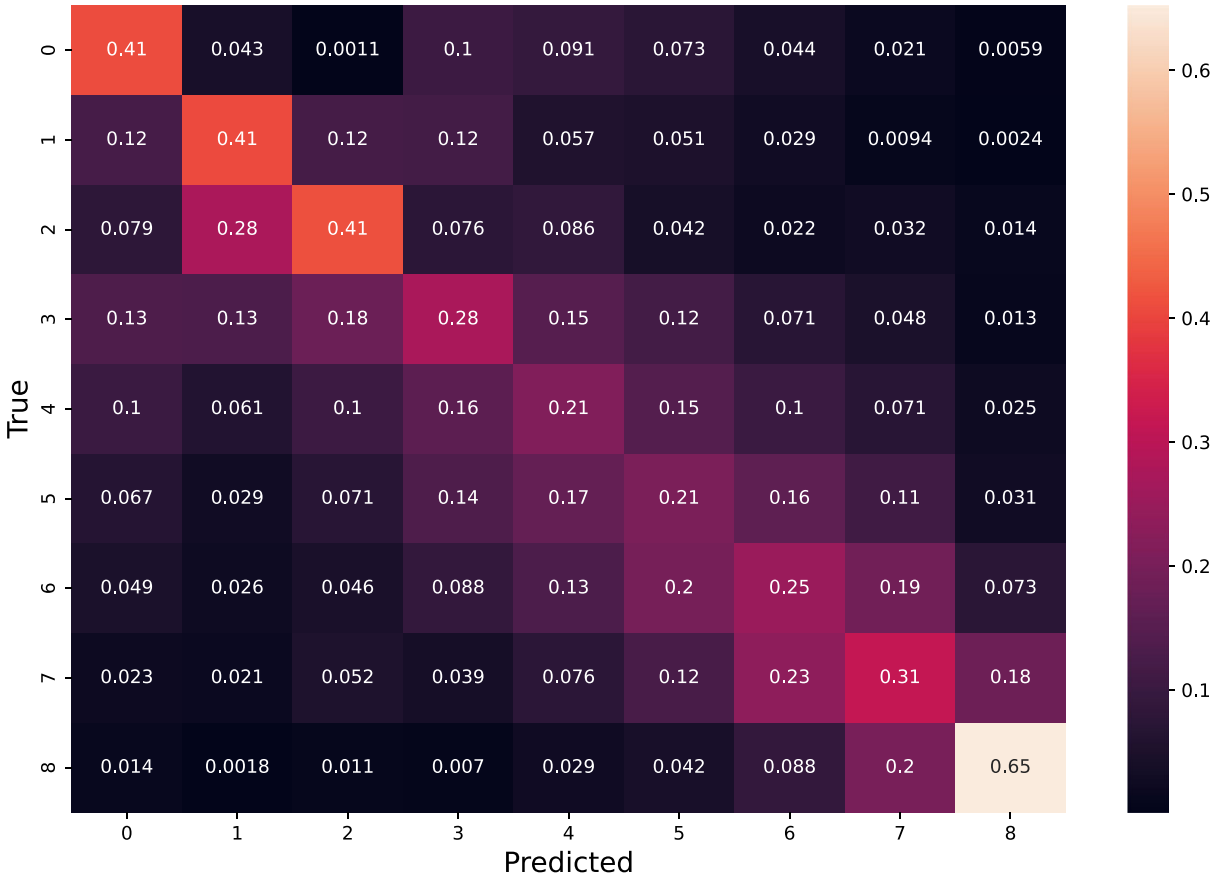


FIGURE 1 Confusion matrix normalized over the predictions for the nine-class task for the best performing classification model, that is, the MLP (bigram + clip image) model

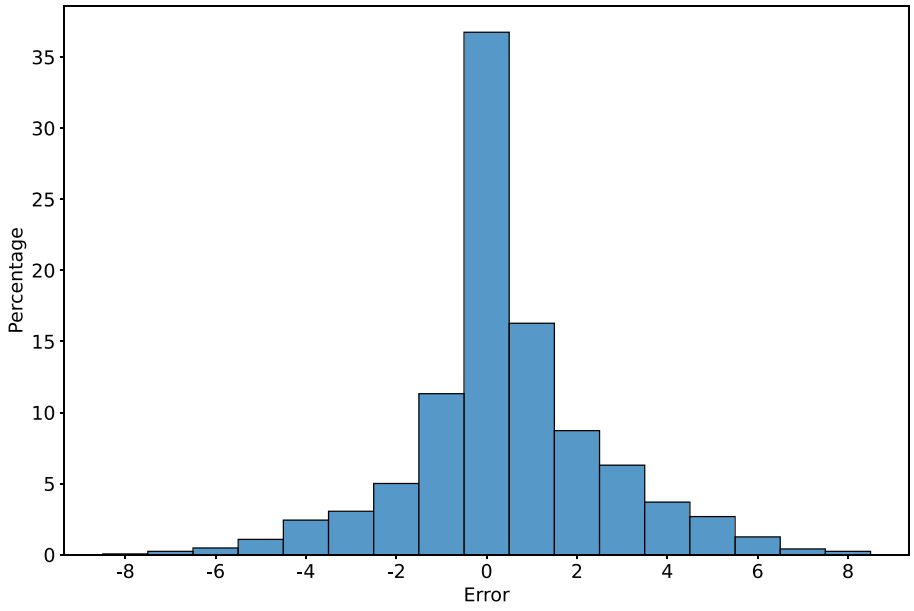


FIGURE 2 Test errors (true-predicted) for the nine-class task of the best performing classification model, that is, the MLP (bigram + clip image) model

Furthermore, we have shown that it is possible to some extent to predict auction end prices for categories clothing. Our results show that image representations of auction items can be used to train a small neural network to model the residual value. Together with text representations from CLIP, the performance can be boosted. However, in the end the simplicity of only using unigram and bigram representations gave the best results, combined with the image representations. This is a promising result, indicating that AI-

predictions could be used to better assess risk in asset- or collateral-based risk assessment scenarios. An objective value calculated by an AI-model could be used to strengthen the arguments for both asset-based and business case-based assessments, since they would potentially be more trustworthy than manually estimated values.

TABLE 6 Description of classes and price ranges (for four and nine classes, respectively)

Class	Price range (SEK)
0	1–50
1	51–75
2	76–150
3	151+
Class	**Price range (SEK)**
0	1–34
1	35–49
2	50
3	51–79
4	80–103
5	104–154
6	155–249
7	250–400
8	401+

These suggested solutions for financing of CBMs sometimes overlap, and they can be combined with each other. The solutions are focused on how to solve the need for bank credit when scaling PaaS models. Depending on the situation and development phase of the PaaS firm, financing solutions could of course also include equity investments and different forms of venture capital, especially in earlier phases. It is also possible to solve some of the challenges of transitioning an existing linear company to a PaaS-based business through placing the circular business in a separate business unit or even company. This could enable the use of more precise and suitable key figures and financial ratios for the benchmark of business cases. Our conclusion is, however, that at some point in the scaling of the PaaS business, whether a start-up or an existing company, bank credit will be a necessary prerequisite for most companies.

While residual value and the possibility to realize collateral through the liquidation of the asset used as security is one of the key aspects of credit assessment frameworks today, it should be noted that in a future more circular world, where the value of products is retained for as long as possible, it is possible—even likely—that there will be no aftermarket for products in PaaS-based CBMs. The products will be kept by the PaaS provider and deliver value for a very long time, until they can no longer be used for their purpose, and need to move into a less value-preserving circular loop such as recycling. This means that residual value as a concept will become less important. This will also possibly blur the lines between the two credit

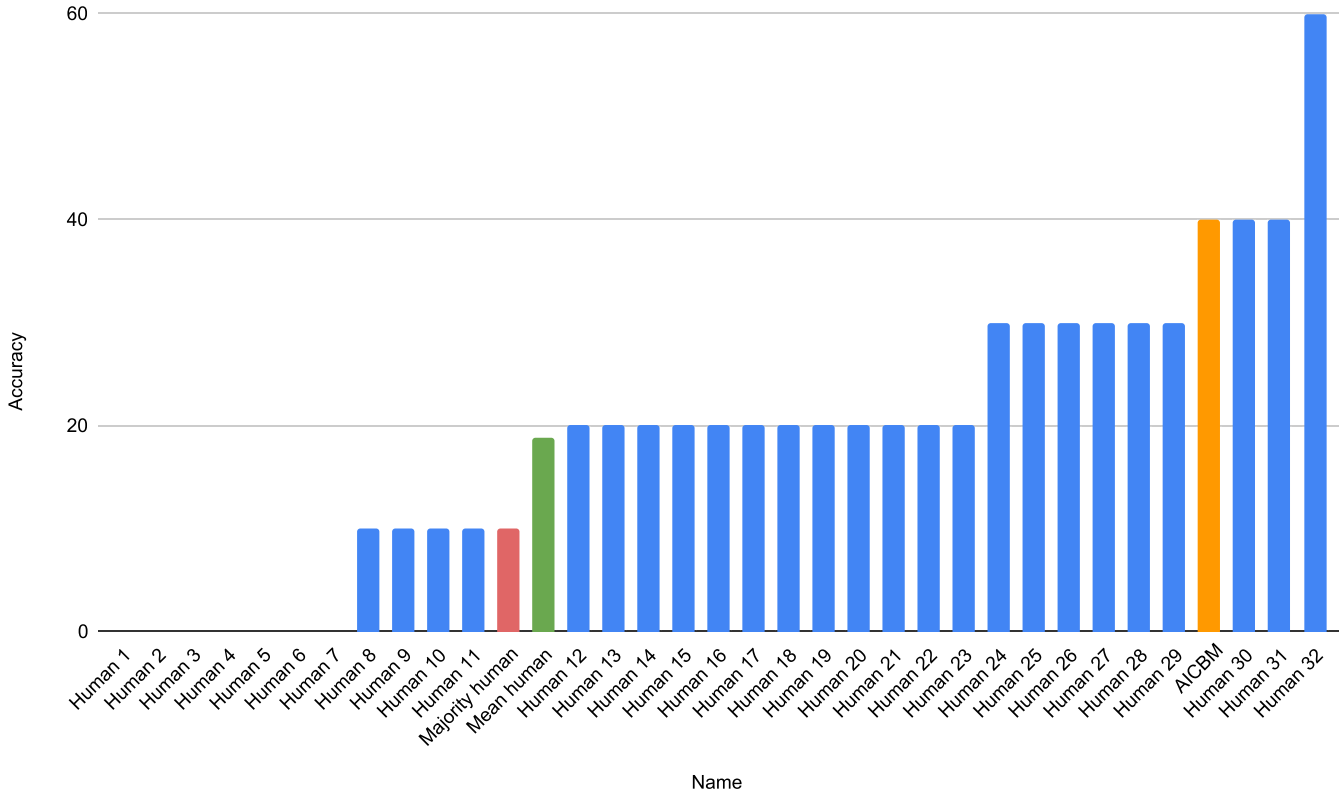


FIGURE 3 Results from the human evaluation. Accuracies for 32 humans (blue) and the proposed vision model (AICBM; orange) on 10 random images from the test set. We also report the achieved accuracy of the mean human (green) and majority human voting (red).

assessment categories, asset-based and business-case based. For the time being, however, it seems likely that solutions trying to predict residual value in an objective way, such as the AI-model in our study—and asset-based credit assessment in general—are important enablers for the circular transition through PaaS CBMs.

7 | MANAGERIAL IMPLICATIONS

For the PaaS-companies, our research results have some important practical implications. Firstly, the business case-based financial solutions indicated above, are all directed toward the PaaS-company, giving them hands-on tips on how to present and package the circular business case in the dialogue with potential financial partners. The asset-based solution on future and modular product design also indicates the importance of combining product and business model design both for optimal circular results and for addressing the risk averse financier. Solutions pointing at the need for new types of relationships also give an indication to the PaaS-company of the importance of establishing and developing networks and relationships beyond the traditional ones, for example, to help establishing value retention through repair and refurbishment partners and to establish residual value points and aftermarkets, through more collaborations with—potentially competitive—actors in the same sector.

Our AI results indicate that there are potential AI-based solutions to tackle financing based on "hard-to-value" assets. And if residual value predictions are combined with monitoring and predicting customer, contract, and payment data, there is an opportunity for a "risk monitor" that could potentially strengthen both the internal management decisions and the dialogue with the financier. Moreover, AI modeling of residual value can be used for other purposes than financial assessment. The quality of production and materials used in the product (offered for sale or as a service) is of crucial value also for strategic decision making in terms of product and business model design decisions, for example, identifying frequent points of failures of a product may give invaluable signals to improve the value retention of products.

For the banks and financial actors, this study points at several concrete solutions to be able to take on the financing challenge of CBMs. The suggested list of credit instruments (leaseback, lease-on-lease and stepwise loans) in the asset-based solutions, points at opportunities that might not be new, but still seem under-explored and under-used in relation to CBMs. Moreover, the suggested relationships-based solutions have strong implications for banks and other financiers, since they go beyond the normal bank–company relationships of today, involving both closer collaboration and deeper business understanding, but also involving other partners that have the potential to share the risk burden. This is a huge opportunity for learning and business development for the bank and might even develop into realization of new financing instruments and vehicles to better serve the financing market for PaaS CBMs.

Further our results indicate that a machine learning model is a much better predictor of residual values of used clothing than humans. This is an important result, since collateral-based risk assessment based on residual values is still important for banks, and if those residual values were assessed by a machine instead of a human being, they could be considered more neutral, trustworthy, and valid. Moreover, the possibility for an AI model to also assess "time before sell" could further decrease the risks associated with taking over inventory/assets in case of default.

8 | CONCLUSIONS

This study extends the short body of empirical literature on managing transition to CBMs by paying particular attention to innovations needed in financial risk assessment and financial instruments for CBMs. The current study is the first to concurrently examine the cross-section of CBMs, business ecosystems, finance, and artificial intelligence by discussing the future of circular business ecosystems and the nature of collaborations required between incumbents and financial actors when moving from mainly linear to innovative CBMs.

This paper provides Circular Economy practitioners with recommendations and insights related to potentials and challenges for financing CBMs. Furthermore, it demonstrates how AI modeling can be incorporated in financial risk assessment, presenting a novel AI solution, which will be made openly available, and a thorough experimental evaluation of its properties. This suggests that AI-based solutions are applicable in the setting of CBMs and motivates further work in this direction. Understanding what alternative financial solutions in new circular business ecosystems could look like will in turn decrease the perceived uncertainties and risks associated with practice of circular economy and can accelerate the transition toward CBMs.

8.1 | Theoretical contributions

This article makes theoretical contributions to the literature on CBMs, sustainable finance and servitization in the following ways:

First, our results contribute to CBM literature by particularly responding to previous literature highlighting financial barriers to CBM when firms transition from product-based to service-based business models. We provide empirical solutions for sustainable financing of CBMs from multiple stakeholders' viewpoint by focusing on both product companies and financial actors needs and uncertainties. The 15 financial solutions provide concrete examples for how the circular financing strategies suggested by Toxopeus et al. (2021) can be implemented by product companies and financial actors.

Second, we show how application of cutting-edge digital technologies such as AI can facilitate modeling the residual value of products and thereby calculating financial risks in circular economy. Awan et al. (2021) found that among empirical studies of digital technologies in the context of circular economy, artificial intelligence was discussed only in a few works, while IoT was more prevalent. Rusch et al. (2022) even revealed that the frequent occurrence of AI as a keyword in this

setting did not reflect the prevalence of an AI-related research. Instead, the keyword AI was in most cases assigned wrongly to papers that did not even mention AI technology but only use references that have the word "Artificial Intelligence" in the title. The experimental model developed in the current article therefore fills this existing gap and makes a novel and hand-on contribution to our understanding of the importance of cross-section of digital technologies and circular economy previously highlighted in the CBM literature (Chauhan et al., 2022; Ellen McArthur Foundation, 2019).

8.2 | Limitations and suggestions for future research

Our study has certain limitations that should be acknowledged when interpreting the results and findings:

First, our study focused on financing solutions from the viewpoint of large international banks and product companies that already have developed PaaS-offerings and have a CBM in operation. This was to be able to gain insights on financing solutions already in place and possible learnings from previous experiences. These insights, however, are limited to specific products in the Swedish market and adopting a broader case selection to test viability and feasibility of the different financing solutions in different industries and markets would provide more generalizable results. We recommend that future research explores how market characteristics and customers' willingness to pay can affect financing solutions available in different markets. From the financial point-of-view other types of risks than residual value of assets are also interesting to explore, for example risks related to different lengths and terms of contracts and risks related to customer behavior and payment history.

Second, the AI-model and experimentation presented in this paper was developed based on data from one of the largest second-hand auction markets available in Sweden. The model was developed based on one product category, which had the largest number of transactions at the time data was exported. By extending the model to other product categories, more insights can be generated for better cross-case analysis. An interesting question for future research is to investigate whether predictions of residual value is more critical and generates more effect in risk assessment for specific product and price classes. The proposed AI approach can be adapted to new settings and other data sources, to enable such investigations.

Third, results from the AI model are based on predictions of residual values drawn from existing peer-to-peer transactions in an auction-based second-hand market, which might not reflect the long-term future residual values of products in a circular economy where items are maintained, repaired, and reconditioned to retain value over a longer period. The model can therefore be further trained with other types of data provided by product companies that retain ownership of the products to prepare them better for circulation between multiple users. Moreover, it is important to remember that the existing peer-to-peer transactions are still bound by the limits of a linear economy that is dominating on the markets. These values are not equally representative in reflecting the residual value of products in a future scenario where circular business models are up and running. We therefore propose that the technology should be continuously updated and retrained to mirror the changes through the transition from linear to circular business models. This will decrease the potential error from transferring the model from the linear domain to the circular domain.

Fourth, as more businesses embrace circular business models, an AI model trained on residual values in a linear economy will no longer be valid and may not reflect the dynamics present in the new setting. We need more investigation into this and hope to be able to return to it in future work. In this setting, other properties of items may be of more interest. How much remaining life does an item have? What are the detailed properties of an item at this point in time? What uses are suitable for an item with a specific remaining life and specific properties? All these questions should be possible to model using AI techniques, and as we further transform into circular economy, we will get the data needed to start tackling them.

Finally, the collected data contains used items put online for sale by individuals. The advertisements contain misspellings, varying formatting, and photographs produced by amateurs without editing. This puts a cap on the accuracy achievable by a predictive model trained on the data. Further investigation should be put into working with data that was more curated, or more uniformly produced. Such data may be available from online retailers, who run second-hand stores for different brands, or from the brand owners themselves. For other future work, it would be interesting to collect and use larger data sets in the modeling. If data is collected over several months, or even years, seasonality and trends could be used to further optimize when the best time to sell an item is, and to estimate the profit. More frequent collection and analysis of data, for example, on a daily basis, could also potentially add value to risk assessments based on residual value, especially for products in fast changing markets.

ACKNOWLEDGMENTS

We are grateful to Vinnova (Sweden's Innovation Agency) for financial support (grant number 2019-03166) through the research project AID-CBM: AI Driven financial risk assessment for Circular Business Models.

CONFLICT OF INTEREST

The authors declare no conflict of interest in this work.

ORCID

Sara Fallahi https://orcid.org/0000-0003-4820-5104
Ann-Charlotte Mellquist https://orcid.org/0000-0002-3462-5987
Olof Mogren https://orcid.org/0000-0002-9567-2218
Lukas Hallquist https://orcid.org/0000-0001-9037-1486

REFERENCES

Accenture. (2018). *Sustainable banking: Finance in the circular economy*. Technical Report. Accenture. Retrieved from: https://www.accenture.com/ma-en/case-studies/banking/sustainable-banking-circular-economy

Adams, K., Osmani, M., Thorpe, A., & Thornback, J. (2017). Circular economy in construction: Current awareness, challenges and enablers. *Waste and Resource Management*, *170*(1), 1–11. https://doi.org/10.1680/jwarm.16.00011

Adner, R. (2012). *The wide lens: A new strategy for innovation*. Penguin Books.

Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of Management*, *43*(1), 39–58. https://doi.org/10.1177/0149206316678451

Awan, U., Sroufe, R., & Shahbaz, M. (2021). Industry 4.0 and the circular economy: A literature review and recommendations for future research. *Business Strategy and the Environment*, *30*(4), 2038–2060. https://doi.org/10.1002/bse.2731

Berger, A. N., & Black, L. K. (2011). Bank size, lending technologies, and small business finance. *Journal of Banking & Finance*, *35*(3), 724–735. https://doi.org/10.1016/j.jbankfin.2010.09.004

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance*, *30*(11), 2945–2966. https://doi.org/10.1016/j.jbankfin.2006.05.008

Bocken, N. M., De Pauw, I., Bakker, C., & Van Der Grinten, B. (2016). Product design and business model strategies for a circular economy. *Journal of Industrial and Production Engineering*, *33*(5), 308–320. https://doi.org/10.1080/21681015.2016.1172124

Boyer, R. H. W., Mellquist, A., Williander, M., Fallahi, S., Nyström, T., Linder, M., Algurén, P., Vanacore, E., Hunka, A. D., Rex, E., & Whalen, K. A. (2021). Three-dimensional product circularity. *Journal of Industrial Ecology*, *25*(4), 824–833. https://doi.org/10.1111/jiec.13109

Brown, P., Bocken, N. M. P., & Balkenende, R. (2020). How do companies collaborate for circular oriented innovation? *Sustainability*, *12*(4), 1648. https://doi.org/10.3390/su12041648

Chauhan, C., Parida, V., & Dhir, A. (2022). Linking circular economy and digitalisation technologies: A systematic literature review of past achievements and future promises. *Technological Forecasting and Social Change*, *177*, 121508. https://doi.org/10.1016/j.techfore.2022.121508

Chen, L. H., Hung, P., & Ma, H. W. (2020). Integrating circular business models and development tools in the circular economy transition process: A firm-level framework. *Business Strategy and the Environment*, *29*(5), 1887–1898. https://doi.org/10.1002/bse.2477

De Bono, E. (1985). *Six thinking hats*. Penguin Books.

Ellen McArthur Foundation. (2019). *Artificial intelligence and the circular economy AI as a tool to accelerate the transition*. Cowes, UK. Retrieved from https://emf.thirdlight.com/link/dl06eujbcbet-wx40o7/@/preview/1?o

Fallahi, S. (2017). *A process view of business model innovation*. (Doctoral dissertation). Chalmers University of Technology.

Frishammar, J., & Parida, V. (2019). Circular business model transformation: A roadmap for incumbent firms. *California Management Review*, *61*(2), 5–29. https://doi.org/10.1177/0008125618811926

Frishammar, J., & Parida, V. (2021). The four fatal mistakes holding back circular business models. *MIT Sloan Management Review*, *62*(3), 68–72.

Ghisellini, P., Cialani, C., & Ulgiati, S. (2016). A review on circular economy: The expected transition to a balanced interplay of environmental and economic systems. *Journal of Cleaner Production*, *114*, 11–32. https://doi.org/10.1016/j.jclepro.2015.09.007

Govindan, K., & Hasanagic, M. (2018). A systematic review on drivers, barriers, and practices towards circular economy: A supply chain perspective. *International Journal of Production Research*, *56*(2), 1–34. https://doi.org/10.1080/00207543.2017.1402141

Grafström, J., & Aasma, S. (2021). Breaking circular economy barriers. *Journal of Cleaner Production*, *292*(2021), 126002, ISSN 0959-6526. https://doi.org/10.1016/j.jclepro.2021.126002

Henry, M., Bauwens, T., Hekkert, M., & Kirchherr, J. (2020). A typology of circular start-ups: An analysis of 128 circular business models. *Journal of Cleaner Production*, *245*, 118528. https://doi.org/10.1016/j.jclepro.2019.118528

ING bank. (2015). *Rethinking finance in a circular economy—Financial implications of circular business models*. ING Bank Report. Retrieved from: https://www.ing.nl/media/ING_EZB_Financing-the-Circular-Economy_tcm162-84762.pdf

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, *24*(4), 602–611. https://doi.org/10.2307/2392366

Kanda, W., Geissdoerfer, M., & Hjelm, O. (2021). From circular business models to circular business ecosystems. *Business Strategy and the Environment*, *30*(6), 2814–2829. https://doi.org/10.1002/bse.2895

Kirchherr, J., Piscicelli, L., Bour, R., Kostense-Smit, E., Muller, J., Huibrechtse-Truijens, A., & Hekkert, M. (2018). Barriers to the circular economy: Evidence from the European Union (EU). *Ecological Economics*, *150*(2018), 264–272, ISSN 0921–8009. https://doi.org/10.1016/j.ecolecon.2018.04.028

Kirchherr, J. W., & van Santen, R. (2019). Research on the circular economy: A critique of the field. *Resources, Conservation and Recycling*, *151*, 104480. https://doi.org/10.1016/j.resconrec.2019.104480

Lahti, T., Wincent, J., & Parida, V. (2018). A definition and theoretical review of the circular economy, value creation, and sustainable business models: Where are we now and where should research move in the future? *Sustainability*, *10*(8), 2799. https://doi.org/10.3390/su10082799

Lieder, M., & Rashid, A. (2016). Towards circular economy implementation: A comprehensive review in context of manufacturing industry. *Journal of Cleaner Production*, *115*, 36–51. https://doi.org/10.1016/j.jclepro.2015.12.042

Linder, M., Mellquist, A., Vanacore, E., Hallquist, L. & Whalen, K. (2022). *Financing circular business models: The case of product-as-a-service. Submitted for publication*.

Linder, M., & Williander, M. (2017). Circular business model innovation: Inherent uncertainties. *Business Strategy and the Environment*, *26*, 182–196. https://doi.org/10.1002/bse.1906

Listo Zec, E., Mogren, O., Mellquist, A., Fallahi, S., & Algurén, P. (2022). *Residual value prediction using deep learning for circular economy*. 2nd International Workshop on Big Data Analytics for Sustainability (BDA4S 2022). Osaka, Japan.

Neligan, A., Baumgartner, R. J., Geissdoerfer, M., & Schöggl, J. P. (2022). Circular disruption: Digitalisation as a driver of circular economy business models. *Business Strategy and the Environment*, 1–14. https://doi.org/10.1002/bse.3100

Nußholz, J. (2017). Circular business models: Defining a concept and framing an emerging research field. *Sustainability*, *9*(10), 1810. https://doi.org/10.3390/su9101810

Nyström, T. (2019). *Adapative Design for Circular Business Models in the automotive manufacturing industry*. [Licenciate]. University of Gothenburg.

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers*. John Wiley & Sons.

Reim, W., Sjödin, D., & Parida, V. (2021). Circular business model implementation: A capability development case study from the manufacturing industry. *Business Strategy and the Environment*, *30*(6), 2745–2757. https://doi.org/10.1002/bse.2891

Rizos, V., Behrens, A., Van der Gaast, W., Hofman, E., Loannou, A., Kafyeke, T., Flamos, A., Rinaldi, R., Papadelis, S., Hirschnitz-Garbers, M., & Topi, C. (2016). Implementation of circular economy business models by small and medium-sized enterprises (SMEs): Barriers and enablers. *Sustainability*, *8*(11), 1212. https://doi.org/10.3390/su8111212

Rusch, M., Schöggl, J. P., & Baumgartner, R. J. (2022). Application of digital technologies for sustainable product management in a circular economy: A review. *Business Strategy and the Environment*, 1–16. https://doi.org/10.1002/bse.3099

Stahel, W. (2010). *The performance economy*. Palgrave Macmillan Hampshire.

Teece, D. J. (2010). Business models, business strategy and innovation. *Long Range Planning*, *43*(2–3), 172–194. https://doi.org/10.1016/j.lrp.2009.07.003

Toxopeus, H., Achterberg, E., & Polzin, F. (2021). How can firms access bank finance for circular business model innovation? *Business Strategy and the Environment*, *30*(6), 2773–2795. https://doi.org/10.1002/bse.2893

Tukker, A. (2004). Eight types of product–service system: Eight ways to sustainability? Experiences from SusProNet. *Business Strategy and the Environment*, *13*(4), 246–260. https://doi.org/10.1002/bse.414

Tura, N., Hanski, J., Ahola, T., Ståhle, M., Piiparinen, S., & Valkokari, P. (2019). Unlocking circular business: A framework of barriers and drivers. *Journal of Cleaner Production*, *212*(2019), 90–98. https://doi.org/10.1016/j.jclepro.2018.11.202

Urbinati, A., Rosa, P., Sassanelli, C., Chiaroni, D., & Terzi, S. (2020). Circular business models in the european manufacturing industry: A multiple case study analysis. *Journal of Cleaner Production*, *274*, 122964. https://doi.org/10.1016/j.jclepro.2020.122964

Veleva, V., & Bodkin, G. (2018). Corporate-entrepreneur collaborations to advance a circular economy. *Journal of Cleaner Production*, *188*, 20–37. https://doi.org/10.1016/j.jclepro.2018.03.196

Vermunt, D. A., Negro, S. O., Verweij, P. A., Kuppens, D. V., & Hekkert, M. P. (2019). Exploring barriers to implementing different circular business models. *Journal of Cleaner Production*, *222*(2019), 891–902, ISSN 0959-6526. https://doi.org/10.1016/j.jclepro.2019.03.052

How to cite this article: Fallahi, S., Mellquist, A.-C., Mogren, O., Listo Zec, E., Algurén, P., & Hallquist, L. (2023). Financing solutions for circular business models: Exploring the role of business ecosystems and artificial intelligence. *Business Strategy and the Environment*, *32*(6), 3233–3248. https://doi.org/10.1002/bse.3297

APPENDIX A

Workshop 1 World Café/Six Thinking Hats (Stockholm, Sweden, January 30, 2020)

World café instructions:

- Three groups/tables, split based on role (see next slide)
  - Banks and financiers
  - Product companies
  - Innovation Enablers
- Discussion 2–3 questions per table/topic.
- One host per table, taking notes. Everyone else moves (after the first round) to a table of choice. 20 + 10 + 10 min/table.
- Each discussion starts with the host doing a short "recap" of previous discussion(s)
- Summary and presentation 3 * 5 min by the hosts.

Groups and questions:

Banks and financiers (Financier A, Financier B, Financier E, Financier F)

What type of information is critical for banks to be willing to take risk in new circular business models and new types of collateral? (How important are residual value and existing second-hand markets compared to other aspects—persons, business case, cash flow?)

What are the most important aspects that AI and Machine learning can add to risk assessment of collateral, where historical data is needed/lacking?

Product companies (OEM A, OEM B)

What type of product information can/should product companies share to support the banks' risk assessments of CBMs? How do product companies value risk and opportunities of CBM in their business?

How should banks and financiers preferably think and act to enable the expansion of CBMs by product companies?

Innovation enablers (Enabler G, Enabler H, Enabler I)

What technologies are critical to enable financing of a transition to CBM?

How can "intermediary technology-based companies"/"innovation enablers" support this process?

What does the future business ecosystem (where banks and producers work in closer collaboration to scale up CBM) look like? What tech- or other supporting roles are there to fill?

Workshop 2 (On-line format facilitated by a Mural canvas, 28th January 2021)

Purpose:

Identify and sharpen the understanding of how the business model of the bank could support financing of PaaS models.

Workshop scenario:

Industry will go through an extensive transition to circular business models and will—for example—transition from linear product sales to keeping ownership of the products and selling their function. We have resigned from earlier positions to start a new company—The Function Bank AB—which will offer financial services to circular companies offering function (or product-as-a-service). In the same way that Omocom realized the lack of a specific service offer for functional sales companies in the insurance industry, we see the potential of competing with existing financiers with a service offer for functional sales-based circular businesses. Today we will meet some potential customers, both start-ups and established larger companies, to discuss and collaborate on how to find a win-win solution.

Instructions—Part 1:

Describe the business model for Function Bank AB in the BMC canvases—for the case of your group (B2C or B2B). Focus on the value proposition (why would the customer buy your service) and your key resources and key activities. Please also note which actor will be the owner of the product, that is, have the booked value on its balance sheet.

The 30-min discussion in breakout-rooms. After that, each group presents its business model, and we will comment and discuss each other's results.

Instructions—Part 2:

Reflect on and answer the following questions. Work with your case, or with both—free of choice.

a. What other actors (outside the product company and the bank) could enable the cases, and what roles could they take?
b. How would the business model/value proposition of Function Bank AB be affected if their customers sell functions/pay-per-use instead of subscriptions?
c. How can data and AI support your model?

The 15-min individual brainstorming. Note the answers on sticky notes in the canvases.

The 45-min presentation and joint discussion.

APPENDIX B

In this work we are training two different models: a logistic regression model and a neural network. Logistic regression is a simple algorithm that can be trained to learn a linear mapping $\mathbf{y} = \mathbf{Wx}+\mathbf{b}$ from some input data **x** to classes **y**. **W** is called a weight matrix and b a bias term, both learned from the training data. A neural network (or multi-layer perceptron, MLP) can be viewed as an extension of this where we add layers of more weight matrices $\mathbf{y} = \mathbf{W_2}\,(\sigma\,(\mathbf{W_1x} + \mathbf{b}))$. This gives the model more capacity to learn more complex patterns in the data. σ is called an activation function, and is a non-linear function added to make it possible to learn non-linear patterns in the data.

In order to solve the auction end price classification task, we have used the title and the description of each item as well as an image of the item to make the prediction. For the text descriptions, we have experimented with three different types of representations: unigrams, bigrams, and Swedish CLIP embeddings. A *n*-gram representation is a simple way of representing text and consists of a sequence of n items (in our case words). For unigrams ($n = 1$), this means that we for each sentence (or description of an item) count which unique words are present. For a bigram ($n = 2$), we instead count unique pairs of words, which gives us spatial information of the sentence.

CLIP is a large deep learning model that is trained to predict which images were paired with which texts in a dataset. We use the Swedish language model in CLIP to create text representations of item descriptions, and we use the vision model (a ResNet RN50x4) to create image representations.

tions for short or long periods of time. One way to mitigate the consequences of unwanted fires is to ensure their detection at an early stage, thereby increasing the potential for successful intervention.

Fire detection systems are common in most industrial facilities, assembly premises, hotels and health care facilities. A recent study, however, indicated that the number of false alarms on automatic alarm systems in Germany could be as high as 87% [2]. Similar data from Sweden indicates that false alarms could actually be as high as 97% in certain applications [3, p. 81]. Independent of the actual size of the problem, false alarms are a serious problem for the owner of a facility, for people in the building and for the fire and rescue services called to the building to respond to the fire. A false fire alarm creates unnecessary interruptions to business operations, forces people to evacuate and introduces a high unnecessary traffic risk. The reasons for false alarms vary depending on the type of detector, its application and position; but, reasons may include non-fire particles in a dirty industrial environment or produced by cooking in a kitchen (whether domestic or industrial), or steam produced in industrial or domestic situations. In essence, there are two solutions to this problem, either false alarms are stopped by organisational measures, i.e. a fire must be confirmed by a complementary means before activating the detection system to initiate a response; or the reliability of the detector is increased through a variety of technical measures. In the latter category, some effort has been made to study multi-sensor fire detection to improve the reliability of detection and reduce the number of false alarms [4, 5]. Such systems typically rely on a combination of traditional sensors and data treatment to reinforce detection reliability by confirmation of detection through several fire characteristics such as, e.g. smoke, temperature, CO-emissions and CO_2-emissions (see e.g. [6, 7]). While such efforts have been successful in improving the level of detection compared to single sensor detectors [6]), they typically rely on a range of chemical (e.g. species) detection methods, heat and particle detection [8].

In recent years, papers have been published concerning the use of various types of machine learning to improve the development of algorithms to analyse these multi-sensor signals (see e.g. [9] and references therein). However, the authors have not been able to find recent papers that refer to the use of audio signals to detect fires. The closest study in the literature concerns a recent article where the authors detect the position of a fire and its characteristics using sound emitted by the detector rather than the fire, in an effort to improve data collection as input to tactical response to the fire [10]. In this application, Xiong et al. [10], assume that the fire itself has been detected by other means and the sound is produced by the alarm itself. But using acoustic signals as a way to detect fires is not addressed in the paper.

Clearly, there is a need to improve the capability for simple detectors to perform with a high level of reliability and, in the long term, this needs to be solved in a cost effective manner [11]. Two strategies can be identified in the literature to solve the issue: improve existing detectors with respect to their sensitivity by signal filters, or find alternative fire characteristics to detect an incipient fire [12]. The current study aims to investigate the novel use of machine learning for fire detection based on acoustic measurements (FORMAS Contract# 2019-00954) in an

effort to improve reliability while using simple detection methods, i.e. the focus is on using the second method with an alternative fire characteristic being applied to detect the fire. The advantage of using sound as the fire characteristic is that it can rapidly reach the detector (more rapidly than smoke dispersion or convective heat transfer through air) and it is less hindered by physical barriers such as walls. Indeed, the concept of using acoustic measurements to detect fires was first investigated in the 1990s at the National Institute of Standards and Technology (NIST) in the US [13], although acoustic flame characteristics have been investigated since the 1960s [14], albeit without reference to fire detection. In a more modern application of acoustics to combustion phenomenon, Nair [15] investigated the use of sound to identify flame blow-off. While interesting from a combustion point of view, his methodology has not been applied to the detection investigation presented in this article as there is no assumption of a steady flame to detect burning.

The initial work by Grosshandler and Jackson [13] provided proof-of-concept for using acoustic detection, but was not pursued due to difficulties with signal to noise ratio and acoustic measurement technologies which could not detect signals for large distances at that time. In the initial study, the efficacy was limited since the detection algorithm was based on hard-coded algorithms. Sound sources are often characterized only by the sound power they are emitting. Sometimes, characteristics of the frequency domain are also studied, with e.g. high-pass or low-pass filters, but it may not be possible to differentiate between vastly different sources, such as music and construction noise, with analytical or numerical methods unless the time domain is considered. When the time signal is considered, however, typically either very simple relationships are considered, e.g. counting the number of events over a threshold level, or complicated processes are studied that depend on well-defined and stable conditions. One example is the Minor Component Analysis (MCA) based method to detect signatures in the time domain presented by Kwan et al. [16].

Recently, machine learning has made great progress for many applications, due to algorithmic developments together with progress in computational capacity and the availability of large datasets with labelled data. Indeed, in a review by Naser [17], the application of machine learning and artificial intelligence in fire engineering and sciences was explored. Naser identified the use of machine learning in enhancing fire detection in domestic applications and wildland fires, but relied on traditional sensors or picture information. In no case that we have found has machine learning been applied to acoustic signals. One of the most successful approaches is deep-learning, which applies artificial neural networks (ANN) with many layers to problem solving. To date, deep learning has been used in such diverse applications as to achieve computer vision for self-driving cars, speech recognition, automatic translation, and text summarization [18]. In particular, a method called convolutional neural networks has completely redefined the state-of-the-art for processing images, video, and audio. In the area of fire safety, several attempts have been performed using computer vision techniques based on deep learning for fire detection [19–21]. Deep learning models have the capacity to discriminate complex patterns in high-dimensional data, potentially overcoming the limitations in early approaches to sound-based fire detection mentioned above.

This paper provides a proof-of-concept for using acoustic measurements together with machine learning for rapid fire detection. The aim is that this proof-of-concept will provide the foundation for more applied research into early fire detection in real fire scenarios where traditional fire detectors are prone to false alarms such as dirty industrial environments, or domestic environments with confounding issues making detection difficult. The novelty of this article is the combination of machine learning techniques with those of acoustic measurements of a simple fire. The aim is to simplify the fire scenario by using a standardised fire test methodology, i.e. the cone calorimeter (ISO 5660). This simplified application has been chosen to limit the research question to whether it is possible to use machine learning to teach a system to recognise whether there is a fire or not. Future studies will explore such questions as, which scenarios are most relevant to acoustic fire detection and whether additional acoustic complications (e.g. additional background noise) invalidate the method. When conducting this work, the authors were in agreement that there is a clear need to simplify the application in this first step and to add layers of complexity in the next step.

2. Theory

This section contains theory for acoustic emissions from fires and machine learning for sound event detection which are relevant to understand the contents of the paper.

2.1. Sound Generation Mechanisms for Fire

Acoustic emission is an essential element of the fire detection algorithms developed in the current study. Fire detection using acoustic emission has been evaluated by Grosshandler and Braun [22] (who actually measured surface vibrations) and by Kwan et al. [16]. These initial studies have, to date, not been pursued further, particularly due to the high risk of false alarms and the problems associated with formulating threshold rules, based on analytical or signal process approaches that can function in noisy environments.

Acoustic emission is defined in ISO 2007 [23] as the range of phenomena that results in the generation of structure-borne and fluid-borne (liquid or gas) propagating waves due to the rapid release of energy from localized sources within and/or on the surface of a material. The sounds are typically either very short transient signals of a wide frequency range, or more continuous signals with narrower frequency distribution due to e.g. leaking heated fluids. There are several types of sound associated with different stages of fire development, from heating and ignition to flaming combustion.

For flaming combustion, sound is generated by hydraulic instabilities and turbulence in the flame and fire plume and is typically located in the infrasound frequency range, although the resulting generated sound may be within the audible range. Detriche and Lanore [24] investigated the pulsation characteristics of small pool fires in 1980 and concluded that the signal was very sensitive to surrounding conditions, making it difficult to use analytical sound characterisation for detect-

ing a fire. It should be noted that improved sensor technology has been developed since the 1980s, and progress in signal processing and data analysis techniques might motivate these studies to be revisited.

Both during heating and flaming combustion, sound is typically emitted by the fuel itself. Indeed, the sound that is generally associated with the moniker "fire sound", is that generated and emitted by heated material. For solid materials, the sounds originate from internal stress due to the physical decomposition and deformation of the material during the heating, pyrolysis and burning phases. A typical example is the crackling sound from burning a log of wood originating from evaporation of small pockets of trapped water in the material. Liquid fuels, as well as some thermoplastics which melt before burning (e.g. PMMA), sound can also be emitted due to boiling of the fuel.

Fires can also induce sounds not directly linked to the combustion process. One example is a paper recently published by Thompson et al. [25] where they use the sound of firebrands as they impact on a steel box to both detect the location of the flame front as well as the fire intensity at that location.

2.2. Machine Learning for Sound Event Detection

Machine learning techniques such as deep neural networks have revolutionized many fields, including computer vision [18]. Recently, learnings from the field of computer vision have been transferred to sound event detection. The goal of sound event detection is computerized analysis of acoustic signals for detection of sound events, i.e. what is heard and when does a specific event occur [26]. Deep neural networks contain a sequence of simple transformations which are usually trained together end-to-end, and learn a hierarchy of representations for the input signal, in each step transforming the data into a space more suitable to solve the end task (see Fig. 1). Similar to two-dimensional optical images, the time-fre-

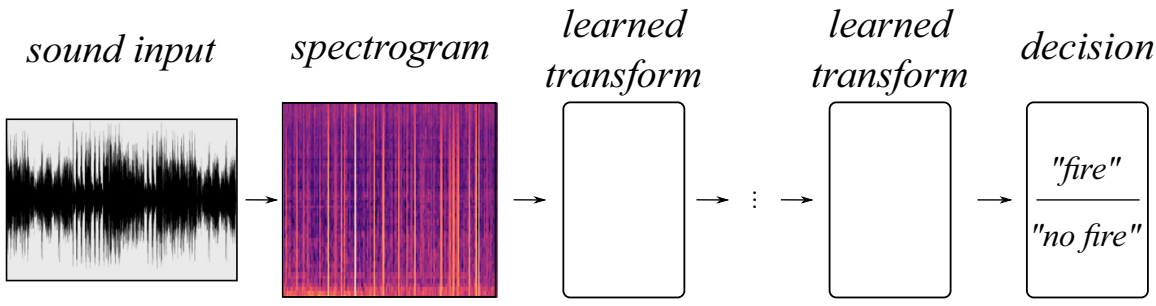


Figure 1. In this work, we consider machine learning models based on deep convolutional neural networks, consisting of a series of transformations or layers. Each layer consists of a linear transformation (matrix-vector multiplication), and an element-wise nonlinearity. Convolutional layers also incorporate a convolution operation which provides spatial invariance. The input to the neural network is a sound signal which is transformed into a Mel spectrogram.

quency representation of audio signals have successfully been modelled using a version of neural networks known as convolutional neural networks [27].

Convolutional neural networks learn to detect complex patterns in a spatially organized input. This includes the potential of spatial invariance, which allows a vision model to detect patterns at different spatial locations, and an acoustic model to detect patterns at different locations in the time-frequency domain. Compared to other machine learning approaches (including Multi-layer Perceptrons), a convolutional network is relatively parameter efficient, and has a larger field-of-view. This class of models obtains state-of-the-art results in many tasks within both vision and acoustics.

3. Experimental Methodology

This section contains a description of how the sound data from the fire experiments were produced and collected, as well as a description of the experimental setup for training and designing a deep convolutional neural network for sound event detection.

3.1. Fire Experiments

Fire experiments were performed using the cone calorimeter [28], which is one of the most widely used instruments for ignition tests within fire science and known to develop repeatable and reproducible results. The method has been standardized in ISO 5660 and ASTM E1354 (see Fig. 2) but basically, the apparatus was applied as a radiative panel without any measurements of effluent gases, heat release rate or weight loss. The apparatus was slightly modified, i.e. the hood was removed to reduce the sound generated from the fan and instruments in the cone calorimeter on the recordings.

Figure 2. The cone calorimeter setup used in the experiments.

Typically, the experimental procedure was to heat up the cone, mount the sample in the sample holder and prepare the recording devices, i.e. the recorder and the microphone. The next step was to initiate the recording, place the sample holder under the pre-heated cone and open the aperture shielding the radiative cone from the sample, thus starting the experiment. The timing started as the aperture opened. The ignition process is inherently unsteady and, therefore, no distinction between the different stages in the burning process (see Sect. 2) has been made. Each experiment was terminated when the sample material stopped burning (see Table 2). This was deemed appropriate for the current study which is a proof-of-concept, future studies should investigate possible differences in acoustic signatures for different stages. Even as the purpose with the experiment was to record sound during the preheating time, i.e. before the material ignited, the recording continued until the material stopped burning to collect also sound during the burning phase. A set of six materials or material combinations were chosen for the initial investigation. Approximately half the tests included wood: softwood (spruce), hardwood (oak) and chipboard, with the remaining containing plastic: polymethylmethacylate (PMMA), polyurethane (PUR) and a PUR/fabric combination. The choices were made to explore a range of common material and fire performance, i.e. charing and melting, see Table 1. Cone calorimeter samples are typically preconditioned according to ISO 5660-1 to minimize sample variability between tests. However, in this application, variability was desirable to prevent overfitting of the model. Therefore, the samples were not preconditioned. To make sure that the signal detected is the fire event, and not the noise from the cone, the isolated sounds emitted by the cone without any sample material was also collected (see Table 2).

Table 1
Description of The Sound Data Recorded for the Fire Event Class Detailing the Different Material Types, the Radiation Used During The Heating Phase, the Thickness of the Material, the Number of Trials (Whole Experiment, Including the Heating, Pyrolysis and Burning Phases), and the Total Amount of Recorded Time for Each Material Type

Sample (–)	Radiation (kW/m^2)	Thickness (mm)	Number of trials (–)	Total recorded time (min)
Oak	35	45	3	33
Oak	30	45	1	15
Oak	35	10	4	22
Spruce	30	43	1	18
Spruce	35	43	15	178
PMMA	30	10	5	61
PMMA	35	10	1	8
PUR	35	50	1	2
PUR/fabric	35	50	1	5
Chipboard	35	10	3	19

Table 2
Description of the Sound Data Recorded for the Non-fire Event Class Detailing the Presence of Fan Noise, Radiation Noise, the Number of Trials and the Total Recorded Time

Fan (–)	Radiation (kW/m^2)	Number of trials (–)	Total recorded time (min)
On	0	1	5
Off	0	1	5
On	35	3	17
Off	35	2	20
Varying	30	1	15
Varying	35	1	15

Even as the hood over the cone was removed, there was still some background noise in the room, mainly emitted by the ventilation system in the lab. To reduce the influence of noises due to the ventilation system in the machine learning phase, the fan was arbitrarily turned on and off during some trials. Timing for when the fan should be turned on or off was sampled randomly between 30 s and 50 s (see Table 2). Further, acoustic damping using mineral wool was mounted on nearby rigid steel surfaces (see Fig. 2).

The sound was recorded using a Zoom H2n, with a sampling frequency of 96 kHz/24 bit, connected to an external microphone of type Earthworks Audio M23. The microphone was placed approximately 100 mm from the sample. The distance between the microphone and the sample was chosen short enough to be able to detect sound from the material decomposition and at the same time with a safe enough distance from the radiative heat source to not damage the microphone. It is desirable to position the microphone as close to the sample as possible as the sound pressure reduces by the square of the distance. The samples were all 100 x 100 mm. Sample material, sample thickness, incident radiation and the data collected is presented in Table 1.

All recordings where a sample material and radiation is present are considered as fire events (see Table 1), and recordings without either a sample or radiation are considered as non-fire events (see Table 2). These acoustic data recordings were used to train a machine learning model to distinguish between fire events and non-fire events, which is further explained in Sect. 3.2.

3.2. Machine Learning for Acoustic Fire Detection

This section presents the way the model has been trained to distinguish between fire and non-fire events on acoustic recordings of fires, and gives a description of the model architecture.[1]

[1] A complete description of the training setup and the model, as well as instructions on how to reproduce the main results of this study can be found in the Git-repository: https://github.com/johnmartinsson/fire-event-detection-dataset/.

3.2.1. Training Setup The acoustic recordings of fire events and non-fire events were first split into training, validation, and test sets. The training set was used to train the model, the validation set was used to validate the model *during training*, and the test set was used to evaluate the performance of the final model. The recordings were down-sampled to 32 kHz to reduce the computational cost and further split into 5 s long segments without overlap. The segments were then uniformly and independently sampled, without replacement, into the training (70%), validation (10%) and test (20%) set respectively. The resulting training, test and validation all have a class imbalance and consists of 16% to 20% non-fire events and 80% to 84% fire events.

The training data was split into batches of 16 segments each and these were used together with a loss-function and the model to compute the gradients used to update the model parameters. The model parameters were optimized using the optimization method Adam [29] which is an extension of the optimization method stochastic gradient decent. A loss function is used to guide the optimizer. Since there are two classes, fire event and non-fire event, this was modeled as a binary classification problem. Binary cross-entropy was the loss function used. An epoch of training is one iteration through the whole training dataset. The model was trained until no more improvements in the loss were observed on the validation data during the previous 100 epochs after which the model with the lowest validation loss was chosen. The training and validation loss curves are shown in Fig. 3, where the model has been trained for a total of 218 epochs meaning that the model with the lowest validation loss was observed at epoch 118 which is the chosen model.

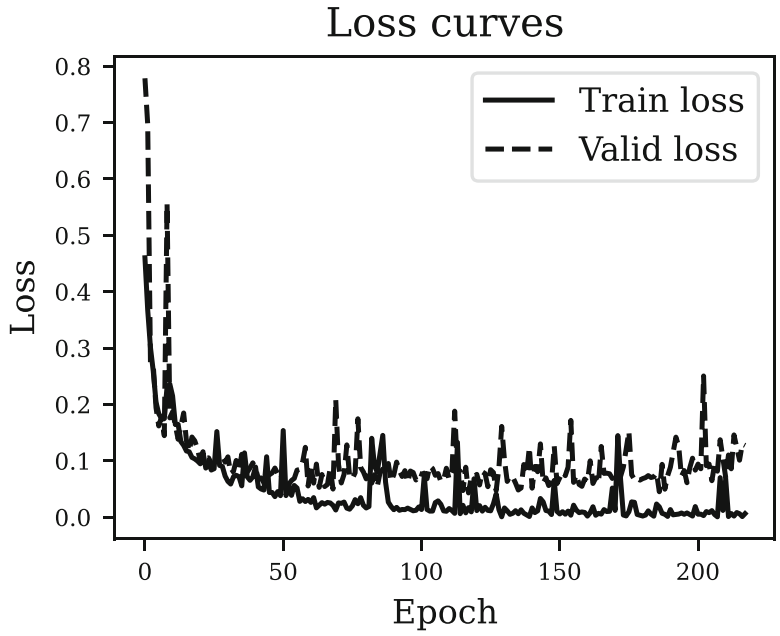


Figure 3. The training and validation loss curves observed during model training.

Table 3
The 14 Layer Convolutional Neural Network Architecture, Consisting of 12 Convolutional Layers with a Kernel Size of 3 × 3 and Dierent Feature Map Sizes According to the Table

Model architecture	CNN14
Input	Log-mel spectrogram, 64 Mel bins
Layers	(3 × 3 @ 64, BN, ReLU) x 2
	(3 × 3 @ 128, BN, ReLU) x 2
	(3 × 3 @ 256, BN, ReLU) x 2
	(3 × 3 @ 512, BN, ReLU) x 2
	Average pooling 2 x 2
	(3 × 3 @ 1024, BN, ReLU) x 2
	Average pooling 2 x 2
	(3 × 3 @ 2048, BN, ReLU) x 2
	Global average pooling
	FC 2048, ReLU
Output	FC 1, Sigmoid

All followed by a batch normalization layer and a rectified linear unit (ReLU) activation function. The last two layers are fully-connected layers of size 2048 and 1 with a ReLU activation and a sigmoid activation respectively

3.2.2. Model The state-of-the-art convolutional neural network model introduced by Kong et al. [30] was used to model the acoustic data. The architecture was designed for classification of sound events, and has been shown to transfer well between different problem domains. The architecture has 14 layers (see Table 3) and takes as input a time-frequency representation of the audio waveform. The time-frequency representation is a Mel spectrogram [31] which is a series of short-time Fourier transforms on sequences of the input data followed by a Mel filter-bank which projects them onto Mel bins. While designing a filter-bank specifically for this task may be beneficial, the development and evaluation of this is left for future work. In this work, the Mel filter-bank is used because of its general applicability and for being the standard choice in the machine listening literature [26].

Any audio segment which is assigned a sigmoid output score of more than a threshold τ is considered as a fire event, otherwise a non-fire event. This threshold can be adjusted, a higher threshold means that the network needs to assign a higher score for an event to be considered as a fire event, which is a way to adjust how sensitive the model is.

3.2.2.1. Input Representation The input to the model is a 5 s waveform with a sample rate of 32 kHz, resulting in 160, 000 samples. A window of size 1024 is moved over the waveform with a hop length of 320, and a short-time Fourier transform is applied to each windowed segment of the waveform to compute the periodogram for each windowed segment. The result is a sequence of periodograms, which is called a spectrogram. The spectrogram is then processed by a Mel filter-bank, which were chosen as a set of 64 triangular filters used to map a decibel-scaled power spectrogram onto the Mel scale [31] (see Table 4 for a summary of the parameters).

Table 4
Parameters for the Mel Spectrogram

Window length	1024
Hop length	320
Window	Hanning
Mel bins	64

3.2.2.2. Model Layers The Mel spectrogram passes through the convolutional neural network which consists of several different layers (see Table 3). In the table "(3 × 3 @ 64, BN, ReLU) x 2" denotes a convolutional block, which consists of a convolutional layer with a kernel of size 3 × 3 wich outputs 64 feature maps (3 × 3 @ 64), followed by a batch normalization layer (BN) and a rectified linear unit (ReLU), applied twice (×2) in that order. Standard average pooling layers are used to reduce the dimensions of the representation, and finally a global average pooling layer is used to take an average over the time-dimension before applying two fully-connected (FC) neural network layers to the final representation of the input. During the training phase a dropout layer with a dropout fraction of 0.2 is applied after each convolutional block.

Dropout [32] is used to prevent over-fitting during training, which is when the model learns the training data too well, and starts performing worse on validation and test data. Batch normalization [33] is used to reduce *internal covariate shift*, and is a way to stabilize the training of the neural network and to speed up convergence.

The rectified linear unit (ReLU) is a non-linear activation function:

$$f(x) = max(x, 0), \tag{1}$$

which has become a standard activation function in the deep learning literature. Compared to e.g. the sigmoid function, the ReLU function requires little computation, and it is argued to reduce the problem of vanishing gradients.

3.2.2.3. Output Representation A fire event is modeled as a 1 and a non-fire event as a 0 using the logistic function:

$$f(x) = 1/(1 + e^{-x}), \tag{2}$$

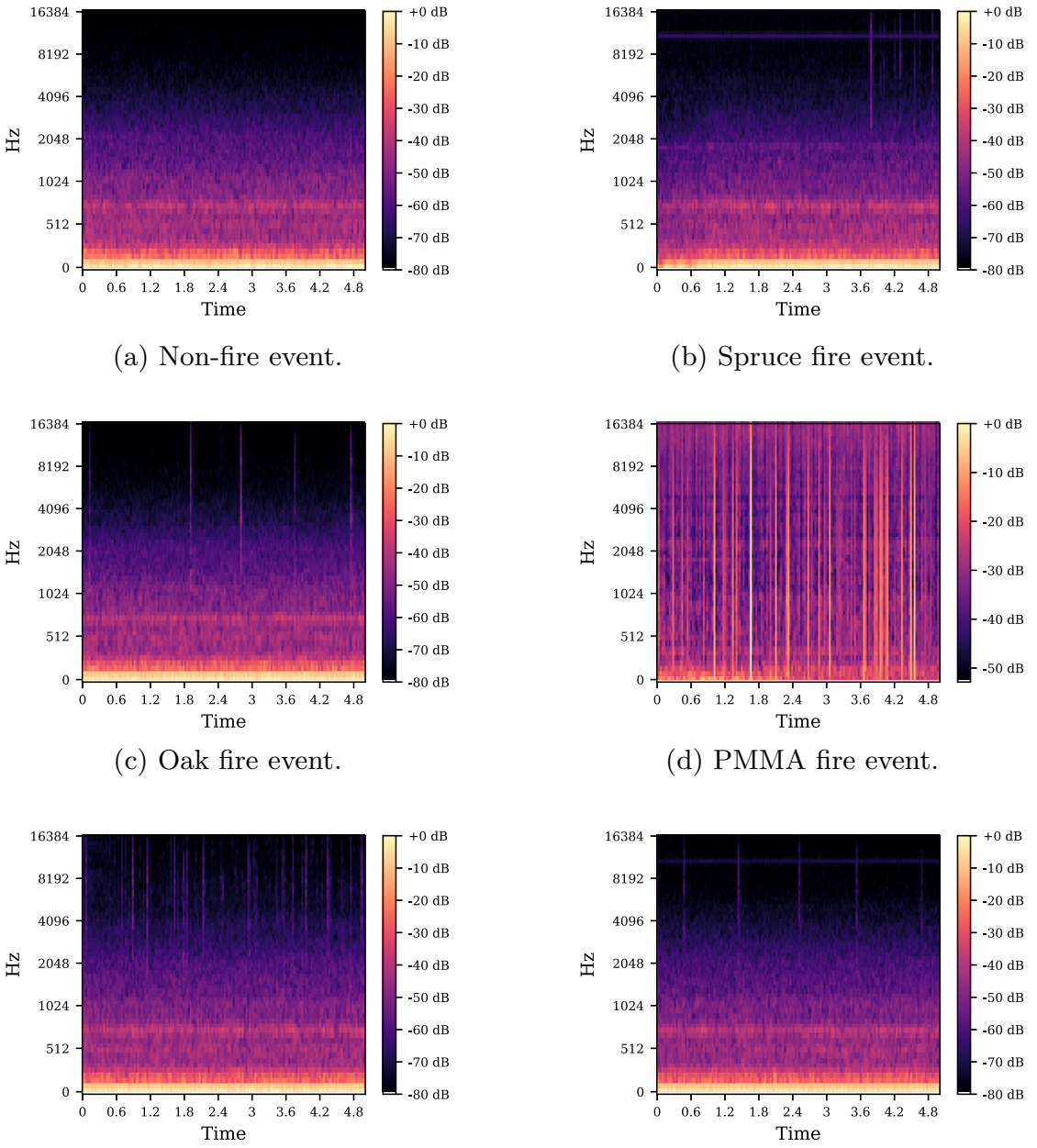
where x is the output of the last fully-connected layer in the deep convolutional netural network.

4. Results

This section contains the results from the analysis of the acoustic signals collected from the different fires, and presents the evaluation results of the final sound event detection model when applied to the test data.

4.1. Acoustic Recordings of Fire Events

A dataset was collected as described in Sect. 3.1. Details of the data can be seen in Tables 1 and 2.



(a) Non-fire event.

(b) Spruce fire event.

(c) Oak fire event.

(d) PMMA fire event.

(e) PUR fire event.

(f) Chipboard fire event.

Figure 4. Mel spectrogram visualization of the waveform collected from each different material type.

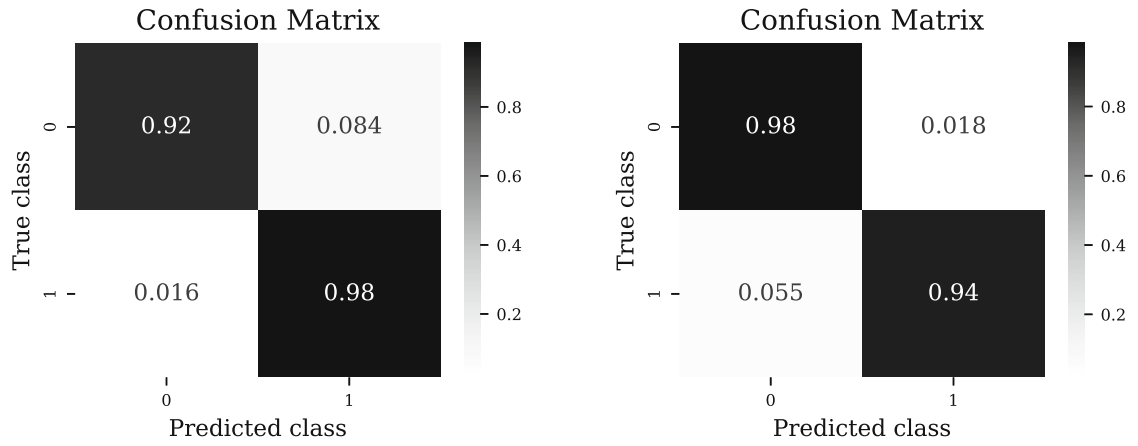
Figure 4 shows an example Mel spectrogram for a 5 s segment for each material type. These are arbitrary examples which have been chosen to provide a visual understanding of the difference between the sound events that occur for the different materials. For the human observer it is easy to distinguish PMMA fire event from a non-fire event, however, for the other materials, the distinction is not as clear. There are clear transient sounds in the recordings from all materials, and, by manual inspection of many of these Mel spectrograms, these transient sounds are the least visually prominent for the recordings of oak fire events.

4.2. Fire Event Detection Using a Convolutional Neural Network

This section presents the results from the analysis of using a deep convolutional neural network for acoustic fire event detection. All results are presented for two different values of τ (see Sect. 3.2) where $\tau = 0.5$ is the default choice, and $\tau = 0.97$ is chosen such that the number of false positives using the validation data is zero. The effect of τ can be seen in Fig. 5.

The main results which demonstrate the effectiveness of the method on the collected data are presented in Table 5. The model achieves a 97.1% accuracy on the test set for the default value of τ and a precision, recall and F-score all equal to 98.4%, which means that there are equally many false positives as false negatives, in this case 14 of each. Choosing $\tau = 0.97$ means that the model becomes less sensitive towards detecting false positive fire events at the cost of becoming more sensitive towards detecting false negative events. That is, trading precision for recall. The overall performance of the model decreases, but maintains a high accuracy and F-score.

The fire event class consists of recordings of fire events from five different material types: spruce, oak, PMMA, PUR and chipboard, and the non-fire event class consists of recordings when there is no material present, i.e. the material type "none". A further analysis of the model performance on each different material



(a) $\tau = 0.5$ (b) $\tau = 0.97$

Figure 5. The normalized confusion matrix for τ = 0.5 (a) and τ = 0.97 (b).

Table 5
The Accuracy, Precision, Recall and F-score for the CNN14 Model on the Test Set with Two Different Threshold Values

Metric	$\tau = 0.5$ (%)	$\tau = 0.97$ (%)
Accuracy	97.3	95.1
Precision	98.4	99.6
Recall	98.4	94.5
F-score	98.4	97.0

A fire event is considered as the positive class, and a non-fire event is considered as the negative class

Table 6
Model Accuracy for Each Respective Material in the Test Set, with $\tau = 0.5$ and $\tau = 0.97$

Material	Accuracy ($\tau = 0.5$) (%)	Accuracy ($\tau = 0.97$) (%)
None	91.6	98.2
Spruce	99.8	98.1
Oak	92.9	78.2
PMMA	99.4	98.9
PUR	100	100
Chipboard	100	100

type is shown in Table 6. The effect of τ is apparent in this table which shows that the less sensitive model achieves a higher accuracy on the non-fire events at the cost of achieving a lower accuracy for in particular the oak material, but also slightly lower for spruce and PMMA.

In Fig. 5 the accuracy of the model on fire events (lower right) and non-fire events (upper left) is presented, as well as the false positive (upper right) and false negative (lower left) rate, for different values of τ.

5. Discussion

The current study presents a setup and method for data collection of acoustic signals from fire events. The collected acoustic signals are used to define a classification task for fire event detection. A convolutional neural network is used to model the acoustic signal and to detect the fire event. These fire events are shown to be detectable with an accuracy of 97.3%, a precision of 98.4%, a recall of 98.4%, and an F-score of 98.4% when the threshold τ is set to 0.5. That is, the fire events, as defined in this study, are shown to be detectable from the acoustic signal using a convolutional neural network.

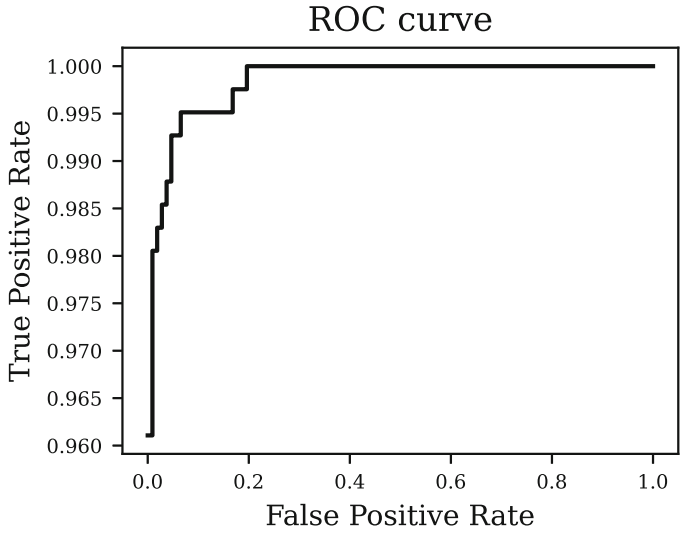


Figure 6. The effect of different threshold values on the true positive and false positive rate.

Note that the class imbalance in this dataset does not reflect what is expected in most real settings where non-fire events would be expected to greatly outnumber fire events. The presented accuracy of the model should therefore be read with that in mind. The F-score and ROC curve are presented as a complement which are suitable metrics for imbalanced datasets (Fig. 6).

The accuracy of the fire event detector varies depending on the material being exposed to the heating condition. The materials that give rise to a very distinct acoustic signal, such as PMMA, are detectable with very high accuracy, and the materials that give rise to a less distinct acoustic signal, such as oak, are harder to detect. Of the wood samples tested, spruce is detected with the highest accuracy and it can be hypothesised that this is due to the more pronounced crackling sound associated with spruce compared to oak. However, the sound produced by the flame and fire plume during the combustion phase could also have an effect. Also, the external conditions like initial temperature and moisture content may also have an influence on the acoustic characteristics, especially for wooden based samples. The sensitivity of the model to variations in temperature and moisture was somewhat decreased by using non-preconditioned samples, and thus a variability in this respect in the training set. However, the ability of the method to be generalized to other materials and conditions than those present in this study is not known, and to take this work from a proof of concept stage to a realistic task, more materials and fire scenarios are needed in the data set.

It should also be noted that the heating conditions used in this study are not necessarily representative of how most actual fire starts, but were chosen as a way to isolate the acoustic fire event signal of interest from other potential acoustic signals to demonstrate that it is feasible to use acoustic measurements for fire detection. The potential influence of heating conditions is not known at this time although efforts were made to compensate for sounds emitted from the heating cone. A benefit of the chosen experimental setup is that it is well known in the fire community and known to deliver results that are repeatable and reproducible.

The strength of a data driven method is that it can be adapted to a new environment, either by training the model using data collected from such an environment, or using data which has been augmented to resemble such an environment. A limitation in the data collection setup in this study is that there was not much variance in the acoustic environment. In a real setting there may be other noise sources present such as talking humans and driving vehicles, and the impulse-response of the acoustic room may also vary depending on e.g. the size of the room and the material of the walls.

To make the model more robust against varying acoustic environments the training and test data need to capture this variance. A way to mitigate the need for such costly data collection efforts is to augment the already existing data with other noise sources by simply mixing multiple acoustic signals together. To emulate different acoustic rooms the impulse-response of such environments could also be taken into account when mixing the signals.

The distance between the fire and the microphone will have an effect on the performance of the system. In this study, we collected data where the sound source was 100 mm from the microphone (see Sect. 3.1). A number of sources of noise is present in this data; most notably ventilation and electrical interference. At a greater distance, the increased signal-to-noise ratio will make fire prediction harder but we hypothesize that the solution will have potential if the training data is extended to cover this variance through further collection or data augmentation. We leave it to future work to study this effect in detail.

Another promising way of reducing the need for extensive data collection is transfer learning. The neural network architecture used in this study has been developed and shown to transfer well between different acoustic tasks, and pre-training the network on similar acoustic data is an interesting way forward. Transfer learning and data augmentation could therefore be two important ways forward to take this a proof of concept to a method applicable in a more realistic setting.

The data collected in this study, together with the annotations, have been made publicly available to facilitate further research on fire event detection using acoustic signals. Instructions on how to download the data can be found in the supplementary material.

Interesting future work would be to treat this as a regression problem and, e.g., study if it is possible to predict more detailed characteristics of the flame such as flame size or heat release rate from the acoustic signal during the kindling phase, or the time until and after the kindling phase.

6. Conclusions

This study investigates the use of acoustic sensors for early fire detection. Microphones are a relatively inexpensive form of sensor, and using the acoustics from a fire event as a complementary signal in current fire event detection methods can make them more robust and reliable. The results presented shows that the acoustic signal from a fire event can be used to detect fires in the setting proposed in

this study. The acoustic vibrations of the materials exposed to heat are used to train a machine learning method to detect such vibrations. The results show that the machine learning method can detect fire events from measurements of the acoustic signals being emitted from the materials when heated. The analysis suggests that performance of the convolutional neural network varies depending on the material which is being exposed to the heating condition.

The proposed method provides proof-of-concept only and further research is needed to investigate, e.g. the impact of different acoustic environments and different materials on the predictive qualities of the method. Transfer learning, domain adaptation, and data augmentation are suggested as potential methods for further investigation.

7. Supplementary information

All the raw data used in this study can be found at the following Git-repository: https://github.com/johnmartinsson/fire-event-detection-dataset. The repository contains instructions on how to download and pre-process the data, and how to train and evaluate the machine learning model presented in this study on the data.

Acknowledgements

The work presented in this article was funded by FORMAS, the Swedish Research Council for Sustainable Development (Contract Number: 2019-00954).

Funding

Open access funding provided by RISE Research Institutes of Sweden.

References

1. McNamee M, Meacham B, van Hees P, Bisby L, Chow WK, Coppalle A, Dobashi R, Dlugogorski B, Fahy R, Fleischmann C, Floyd J, Galea ER, Gollner M, Hakkarainen T, Hamins A, Hu L, Johnson P, Karlsson B, Merci B, Ohmiya Y, Rein G, Trouvé A, Wang Y, Weckman B (2019) IAFSS agenda 2030 for a fire safe world. Fire Saf J. https://doi.org/10.1016/j.firesaf.2019.102889
2. Festag S (2016) False alarm ratio of fire detection and fire alarm systems in Germany—a meta analysis. Fire Saf J 79:119–126. https://doi.org/10.1016/j.firesaf.2015.11.010
3. Hjort B (2001) Automatiskt brandlarm - onödiga larm. Technical report, Räddningsverket, Karlstad
4. To H, Fong N (2013) Investigation of the performance and improvement of optical smoke detectors. Procedia Eng 62:985–993. https://doi.org/10.1016/j.proeng.2013.08.152
5. Chen X, Bu L (2010) Research of fire detection method based on multi-sensor data fusion. In: 2010 International conference on computational intelligence and software engineering, CiSE 2010. https://doi.org/10.1109/CISE.2010.5677271
6. Milke JA, Hulcher ME, Worrell CL, Gottuk DT, Williams FW (2003) Investigation of multi-sensor algorithms for fire detection. Fire Technol 39(4):363–382. https://doi.org/10.1023/A:1025378100781
7. Davis WD, Cleary T, Donnelly M, Hellerman S (2003) Using sensor signals to analyze fires. Fire Technol 39(4):295–308. https://doi.org/10.1023/A:1025322015802
8. Baek J, Alhindi TJ, Jeong Y-S, Jeong MK, Seo S, Kang J, Heo Y (2021) Intelligent multi-sensor detection system for monitoring indoor building fires. IEEE Sensors J. https://doi.org/10.1109/JSEN.2021.3124266
9. Yu C, Fang J, Wang J, Zhang Y (2010) Video fire smoke detection using motion and color features. Fire Technol. https://doi.org/10.1007/s10694-009-0110-z
10. Xiong C, Wang Z, Huang Y, Shi F, Huang X (2022) Smart evaluation of building fire scenario and hazard by attenuation of alarm sound field. J Build Eng. https://doi.org/10.1016/j.jobe.2022.104264
11. Khan F, Xu Z, Sun J, Khan FM, Ahmed A, Zhao Y (2022) Recent advances in sensors for fire detection. Sensors. https://doi.org/10.3390/s22093310
12. Fonollosa J, Solórzano A, Marco S (2018) Chemical sensor systems and associated algorithms for fire detection: a review. Sensors. https://doi.org/10.3390/s18020553
13. Grosshandler W, Jackson M (1994) Acoustic emission of structural materials exposed to open flames. Fire Saf J 22(3):209–228. https://doi.org/10.1016/0379-7112(94)90012-4
14. Thomas A, Williams GT (1966) Flame noise: sound emission from spark-ignited bubbles of combustible gas. Proc R Soc Lond Ser A Math Phys Sci 294(1439):449–466
15. Nair S (2006) Acoustic characterization of flame blowout phenomenon. PhD thesis, Georgia Institute of Technology
16. Kwan C, Zhang X, Xu R (2003) Early fire detection using acoustic emissions. IFAC Proc Vol (IFAC-PapersOnline) 36(5):351–355. https://doi.org/10.1016/S1474-6670(17)36516-3
17. Naser MZ (2021) Mechanistically informed machine learning and artificial intelligence in fire engineering and sciences. Fire Technol. https://doi.org/10.1007/s10694-020-01069-8
18. Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

19. Muhammad K, Ahmad J, Baik SW (2018) Early fire detection using convolutional neural networks during surveillance for effective disaster management. Neurocomputing 288(C):30–42. https://doi.org/10.1016/j.neucom.2017.04.083
20. Dung NM, Ro S (2018) Algorithm for fire detection using a camera surveillance system. In: Proceedings of the 2018 international conference on image and graphics processing (ICIGP 2018). Association for Computing Machinery, New York, pp 38–42. https://doi.org/10.1145/3191442.3191450
21. Lin G, Zhang Y, Xu G, Zhang Q (2019) Smoke detection on video sequences using 3d convolutional neural networks. Fire Technol 55(5):1827–1847
22. Grosshandler W, Braun E (2019) Fire safety science. In: Proceedings of the fourth international symposium, pp 773–784
23. ISO-22096:2007 (2007) Condition monitoring and diagnostics of machines—acoustic emission. Standard, International Organization for Standardization, Geneva
24. Detriche P, Lanore JC (1980) An acoustic study of pulsation characteristics of fires. Fire Technol 16(3):204–211. https://doi.org/10.1007/BF02476759
25. Thompson DK, Yip DA, Koo E, Linn R, Marshall RG, Refai Schroeder D (2022) Quantifying firebrand production and transport using the acoustic analysis of in-fire cameras. Fire Technol. https://doi.org/10.1007/s10694-021-01194-y
26. Mesaros A, Heittola T, Virtanen T, Plumbley MD (2021) Sound event detection: a tutorial. IEEE Signal Process Mag 38(5):67–83. https://doi.org/10.1109/msp.2021.3090678
27. LeCun Y, Bengio Y (1998) Convolutional networks for images, speech, and time series. MIT Press, Cambridge, pp 255–258
28. Babrauskas V (1982) Development of the cone calorimeter: a bench-scale heat release rate apparatus based on oxygen consumption. NISTInteragency/Internal Report (NISTIR), National Institute of Standardsand Technology, Gaithersburg
29. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings. http://arxiv.org/abs/1412.6980
30. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD (2020) PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans Audio Speech Lang Process 28(1):2880–2894. https://doi.org/10.1109/TASLP.2020.3030497
31. Stevens SS, Volkmann JE, Newman EB (1937) A scale for the measurement of the psychological magnitude pitch. J Acoust Soc Am 8:185–190
32. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(56):1929–1958
33. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D (eds) Proceedings of the 32nd international conference on machine learning. Proceedings of machine learning research, vol 37. PMLR, Lille, France, pp 448–456. https://proceedings.mlr.press/v37/ioffe15.html

1 Introduction

Our future will be recorded and quantified in unprecedented temporal resolution. A rapidly increasing variety of variables gets stored, describing activities we engage in as well as physiological and medical phenomena. One example is the increasingly wide adoption of continuous blood glucose monitoring systems (CGM) which has given type-1 diabetics (T1D) a valuable tool for closely monitoring and reacting to their current blood glucose levels and trends. CGM data helps patients manage their insulin distribution by providing an informative source of data to act upon. CGM availability has also been of crucial importance for the development and use of closed loop systems such as OpenAPS [15]. Blood glucose levels adhere to complex dynamics that depend on many different variables (such as carbohydrate intake, recent insulin injections, physical activity, stress levels, the presence of an infection in the body, sleeping patterns, hormonal patterns, etc) [4, 9]. This makes predicting the short term blood glucose changes (up to a few hours) a challenging task, and developing machine learning (ML) approaches an obvious approach for improving patient care. However, acquiring domain expertise, understanding sensors, and hand-crafting features is expensive and not easy to scale up to further applications. Sometimes natural, obviously important and well-studied variables (e.g. caloric intake for diabetics) might be too inconvenient to measure for end-users. On the other hand deep learning approaches are a step towards automated machine learning, as features, classifiers and predictors are simultaneously learned. Thus they present a possibly more scalable solution to the myriad of machine learning problems in precision health management resulting from technology changes alone.

In this paper, we present a neural network model trained to predict blood glucose levels from CGM history, and demonstrate that

- it is feasible to predict future glucose levels from glucose levels alone,
- appropriate models can be trained by non-experts without feature engineering or complicated training procedures, and
- the proposed model can quantify the uncertainty in its predictions to alert users to the need for extra caution or additional input.

Our method was trained and evaluated on the Ohio T1DM dataset for blood glucose level prediction; see [16] for details.

2 Modeling blood glucose levels using recurrent neural networks

A recurrent neural network (RNN) is a feed forward artificial neural network that can model a sequence of arbitrary length, using weight sharing between each position in the sequence. In the basic RNN variant, the transition function at time t is a linear transformation of the hidden state $\mathbf{h_{t-1}}$ and the input, followed by a point-wise non-linearity:

$$\mathbf{h}_t = \tanh(W\mathbf{x}_t + U\mathbf{h}_{t-1} + \mathbf{b}),$$

where W and U are weight matrices, $\mathbf{b}$ is a bias vector, and tanh is the selected nonlinearity. W, U, and $\mathbf{b}$ are typically trained using some variant of stochastic gradient descent (SGD).

Basic RNNs struggle with learning long-range dependencies and suffer from the vanishing gradient problem. This makes them difficult to train [12,1], and has motivated the development of the Long short term memory (LSTM) architecture [13], that to some extent solves these shortcomings. An LSTM is an RNN where the cell at each step t contains an internal memory vector $\mathbf{c}_t$, and three gates controlling what parts of the internal memory will be kept (the forget gate $\mathbf{f}_t$), what parts of the input that will be stored in the internal memory (the input gate $\mathbf{i}_t$), as well as what will be included in the output (the output gate $\mathbf{o}_t$). In essence, this means that the following expressions are evaluated at each step in the sequence, to compute the new internal memory $\mathbf{c}_t$ and the cell output $\mathbf{h}_t$. Here "$\odot$" represents element-wise multiplication and $\sigma(\cdot)$ is a logistic sigmoid function.

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i), \\
\mathbf{f}_t &= \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + \mathbf{b}_f), \\
\mathbf{o}_t &= \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1} + \mathbf{b}_o), \\
\mathbf{u}_t &= \tanh(W_u \mathbf{x}_t + U_u \mathbf{h}_{t-1} + \mathbf{b}_u), \\
\mathbf{c}_t &= \mathbf{i}_t \odot \mathbf{u}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1}, \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t).
\end{aligned}
$$

We model the blood glucose levels using a recurrent neural network (see Figure 1), working on the sequence of input data provided by the CGM sensor system. The network consists of LSTM cells. The whole model takes as input a stream of blood glucose measurements from the CGM system and outputs one prediction regarding the blood glucose level after time T (we present experimental evaluation for $T \in \{30, 60\}$ minutes). An RNN is designed to take a vector of inputs at each time step, but in the case of feeding the network with blood glucose measurements only, the input vectors are one-dimensional (effectively scalar valued).

The output vector from the final LSTM cell (see $\mathbf{h}_t$ in Figure 1) in the sequence is fed through a fully connected neural network with two hidden dense layers and one output layer. The hidden layers consist of 512 and 256 neurons respectively, with rectified linear activations and a dropout of 20% and 30% respectively. The dropout layers mitigate over-fitting the model to the training data. The output layer consists of two neurons: one with a linear activation and one with an exponential activation.

The output is modeled as a univariate Gaussian distribution [3], using one value for the mean, μ, and one value for the standard deviation, σ. This gives us an estimate of the confidence in the models' predictions.

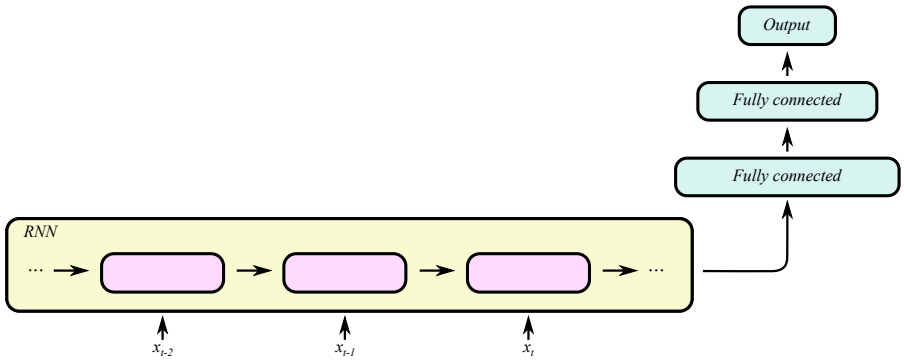


Fig. 1: High-level illustration of the RNN model used in this work. Each RNN cell processes the blood glucose level at one time step, and at prediction time t, the RNN output h_t is used as input to a stack of fully connected layers finally outputting the parameters for the predicted distribution of the future glucose level. Boxes represent neural network layers (processing), and each arrow represents a vector fed from a layer to the next.

$$\mu = W_1 \mathbf{h}_{fc} + \mathbf{b}_1 \tag{1}$$

$$\sigma = \exp(W_2 \mathbf{h}_{fc} + \mathbf{b}_2) \tag{2}$$

where $\mathbf{h}_{fc}$ is the output of the last hidden dense layer. As in [3], we use a linear activation for the mean (see Equation 1), and an exponential activation for the standard deviation (see Equation 2) to ensure that the output is positive since standard deviation is not defined for negative values.

The negative log-likelihood (NLL) loss function is derived from the Gaussian probability density function,

$$\mathcal{L} = \frac{1}{k} \sum_{i=0}^{k} - \log \left(\mathcal{N}(y_i | \mu_i, \sigma_i^2) \right),$$

where y_i is the target value from the data, and μ_i, σ_i are the network's output given the input sequence $\boldsymbol{x}_i$. This way of modeling the prediction facilitates basing decisions on the predictions, by providing an estimate of the prediction uncertainty.

Physiological loss function: We also trained the model with a glucose-specific loss function [10], which is a metric that combines the mean squared error with a penalty term for predictions that would lead to contraindicated interventions possibly leading to clinically critical situations.

2.1 Preliminary study

Preliminary results from this study was presented at *The 3rd international workshop on knowledge discovery in healthcare data* at ICML/IJCAI 2018 [17]. However, since the preliminary workshop paper, the proposed model has been

further refined by a more thorough exploration of hyperparameters and changes to the model design (such as the activation functions), and the results have consequently improved. This paper also includes a more thorough analysis, such as surveillance error grid analysis and an investigation of the variance predictions using controlled synthetic data. The model in the current study is trained on all available training data whereas the preliminary study considered models trained specifically for one patient at a time.

2.2 Experimental setup

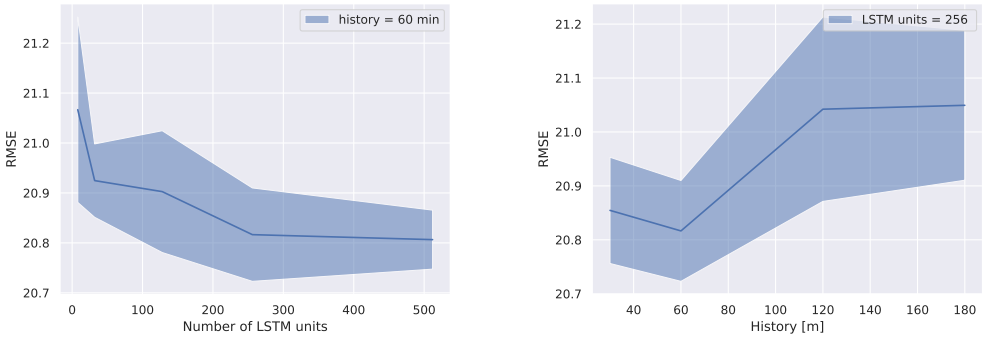
We trained and evaluated our method on the Ohio T1DM dataset for blood glucose level prediction [16]. The data consists of blood glucose level measurements for six people with type 1 diabetes (T1D). A continuous glucose monitoring (CGM) device was used to collect eight weeks of data, at five minute intervals, for each of the six patients. There were two male patients and four female patients between 40 and 60 yeras old. All patients were on insulin pump therapy. There are roughly the same number of blood glucose level observations for each patient in the training and testing data (see Table 1). The patients have been de-identified and are referred to by ID numbers. Patients 563 and 570 were male, and patients 559, 575, 588 and 591 were female.

Table 1: The number of blood glucose level measurements that are used as training and testing data for each patient in the Ohio T1DM dataset for blood glucose level prediction. The table also show the gender for each patient.

Patient ID	Training examples	Test examples	Gender
559	10796	2514	F
563	12124	2570	M
570	10982	2745	M
575	11866	2590	F
588	12640	2791	F
591	10847	2760	F

There are other data self-reported by the patients such as meal times with carbohydrate estimates; times of exercise, sleep, work, stres, and illness; and measures of heart rate, galvanic skin response, skin temperature, air temperature, and step count. In this work we consider the problem of predicting future blood glucose levels using only previous blood glucose level measurements. The only preprocessing done on the glucose values is scaling by 0.01 as in [19] to get the glucose values into a range suitable for training.

Dataset split: For all patients, we take the first 60% of the data and combine it into a training set, we take the following 20% of the data and combine it into a validation dataset used for early stopping, and we choose the hyperparameters by the root mean squared error performance on the last 20% of the data.



(a) (b)

Fig. 2: Mean RMSE and standard deviation (shaded region) for the validation data over 30 different random initializations for each hyperparameter configuration. A history of 60 minutes means that the LSTM use the blood glucose measurements taken during the last 60 minutes to make a prediction 30 minutes into the future.

Hyperparameter selection: The hyperparameters for the model are chosen using grid search over different parameter configurations. The size of the LSTM state was selected from the range $\{8, 32, 128, 256, 512\}$ and the amount of history from $\{30, 60, 120, 180\}$ minutes. We use the Adam optimizer with a batch size of 1024 and a learning rate of 10^{-3} and set the early stopping criterion to 20 epochs. That is, if no improvement is observed on the validation data for the last 20 epochs we terminate the training. For each hyperparameter configuration we train with 30 different random seeds and choose a model configuration with a low mean RMSE score while keeping the model complexity low. The results are shown in Figure 2. Using a glucose level history of 60 minutes to make a prediction results in the lowest RMSE on the validation data. The difference in RMSE between using 256 and 512 LSTM units is very small, and we choose 256 LSTM units to keep the model complexity low.

We then choose the learning rate and the batch size by fixing the number of LSTM units and the amount of history used and instead vary the learning rate between 10^{-3} and 10^{-5} and the batch size between 128 and 1024. The converged models give approximately the same validation loss for different learning rates and batch size, but a learning rate of 10^{-3} and a batch size of 1024 leads to faster convergence and is therefore chosen.

Final models: The final models were trained using 60 minutes of glucose level history for predictions 30 and 60 minutes into the future. The setup for the final training was to train on the first 80% of the glucose level training data from all patients, and do early stopping on the last 20%. The final models were trained with the Adam optimizer with a learning rate of 10^{-3}, a batch size of 1024, a maximum of $10,000$ epochs, and an early stopping criterion set to 200 epochs. We train 100 models with different random initializations of the parameters and report the mean evaluation score for all 100 models on the test data.

Evaluation: The final models were evaluated on the officially provided test partition of the dataset. Root mean squared error (RMSE) and surveillance error scores are reported. Each CGM value in the test set is considered a prediction target provided that it is preceeded by enough CGM history. The number of missing predictions depends on the number of gaps in the data, i.e., the number of pair-wise consecutive measurements in the glucose level data where the time-step is not exactly five minutes. We do not interpolate or extrapolate to fill the missing values since it is unclear how much bias this would introduce, but instead only use data for which it is possible to create the (x, y) pairs with a given glucose history, x, and regression target, y, for a given prediction horizon. As a result, we make predictions for approximately 90% of the test data. The discarded test-points are not counted in the evaluation.

Computational requirements: In our experimental setup training of the model could be performed on a commodity laptop. The model is small enough to fit in the memory of, and be used on mobile devices (e.g. mobile phones, blood glucose monitoring devices, etc). Training could initially be performed offline and then incremental training would be light enough to allow for training either on the devices or offline.

3 Results

Table 2: Mean and standard deviation of the root mean squared error (RMSE) per patient over 100 different random initializations and the mean over all patients for predicting glucose levels 30 respectively 60 min into the future. t_0 denotes the naive baseline of predicting the last value.

	30 min horizon		**60 min horizon**	
Patient ID	LSTM	t_0	LSTM	t_0
559	18.773 ± 0.179	23.401	33.696 ± 0.365	39.404
570	15.959 ± 0.374	18.809	28.468 ± 0.834	31.577
588	18.538 ± 0.106	21.893	31.337 ± 0.210	35.928
563	17.961 ± 0.192	20.786	29.012 ± 0.169	34.032
575	21.675 ± 0.218	25.452	33.823 ± 0.268	39.164
591	20.294 ± 0.107	24.249	32.083 ± 0.182	38.219
μ	18.867	22.432	31.403	36.387
σ	± 1.794	± 2.217	± 2.078	± 2.860

The results presented in Table 2 are the mean RMSE and the standard deviation on the test data for 100 models with the same hyper parameter configuration but with different random initializations presented for each patient individually and as a mean over all patients. The baseline, t_0, is just naively predicting the last known glucose value.

(a) (b)

Fig. 3: Prediction (orange), predicted standard deviation (shaded orange) and the ground truth glucose concentration (dashed blue) for patient 570 (a) and 575 (b). The plot shows eight hours of predictions starting from an arbitrarily chosen time for each patient in the test data. The predictions are 30 minutes into the future.

The glucose level of patient 575 is harder to predict than the glucose level for patient 570, as seen in Table 2 where the mean RMSE for patient 570 is 15.959 and the mean RMSE for patient 575 is 21.675. We observe that patient 575 has higher glucose variability than patient 570. The percentage of first differences greater than 10 mg/dl/5m or lower than -10 mg/dl/5m are 7.3% for patient 575 and 3.0% for patient 570 in the test data. Abnormal rates of change are potentially harder to predict, which may partially explain why the performance is lower on patient 575 than on patient 570.

Figure 3a and Figure 3b show the predicted glucose concentrations and the corresponding ground truth glucose concentrations for patient 570 and 575. We see that the predictions follow the ground truth well in most regions, but that there is a lag in the predicted values for quickly increasing regions.

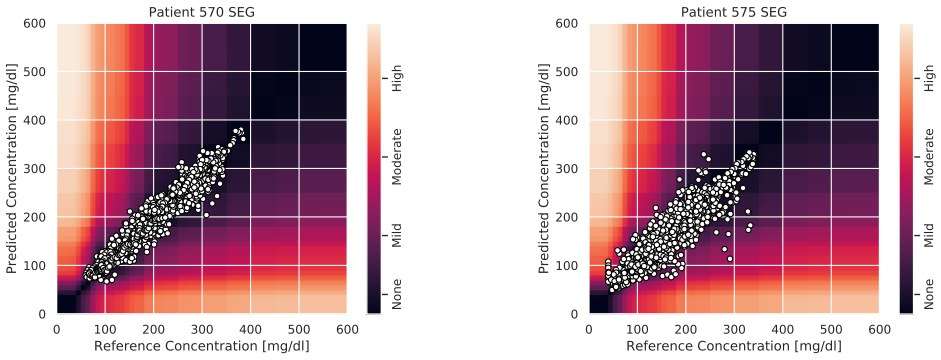
Surveillance error grid: In addition to the RMSE metric it is informative to know how well the model performs in a clinical scenario. We therefore use the surveillance error grid [14] to define an evaluation criterion that accounts for the clinical risk of making an incorrect prediction. The criterion is defined by a bilinear interpolation of the 600×600 surveillance error grid and is denoted by $e(y, \hat{y}) \in [0, 4]$, where $e(y, \hat{y})$ is the estimated clinical risk of predicting the blood glucose concentration $\hat{y} \in [0, 600]$ (in mg/dl) given that $y \in [0, 600]$ is the ground truth concentration. Let $\{\hat{y}_t | t \in \{1, \ldots, T\}\}$ be the predictions for a patient at each discrete time step t, and let $\{y_t | t \in \{1, \ldots, T\}\}$ be the corresponding ground truth reference concentrations. The criterion is then given by

$$SE = \frac{1}{T} \sum_{t=1}^{T} e(y_t, \hat{y}_t).$$

Note that the criterion is only defined for blood glucose concentrations up to 600 mg/dl, which is the limit of most CGMs and any model that predicts values outside of this region should be discarded or constrained.

Table 3: Results individually per patient and averages in predicting glucose levels with a 30 respectively 60 min prediction horizon. The table shows the surveillance error (SE) of the LSTM model trained with NLL. t_0 refers to the naive baseline of predicting the last value.

	30 min horizon		**60 min horizon**	
Patient ID	LSTM	t_0	LSTM	t_0
559	0.178 ± 0.003	0.224	0.331 ± 0.003	0.386
570	0.105 ± 0.002	0.141	0.195 ± 0.004	0.244
588	0.177 ± 0.002	0.214	0.291 ± 0.002	0.349
563	0.176 ± 0.002	0.222	0.293 ± 0.002	0.360
575	0.224 ± 0.004	0.272	0.389 ± 0.005	0.434
591	0.256 ± 0.003	0.299	0.396 ± 0.003	0.478
μ	0.186	0.229	0.316	0.375
σ	± 0.047	± 0.050	± 0.068	± 0.073



(a) (b)

Fig. 4: The surveillance error grid overlayed with each model prediction concentration and reference concentration for patient 570 (a) and patient 575 (b). The predictions are for all the test data points preceded by 90 minutes of consecutive glucose level measurements without missing values. That is, 60 minutes of history and a 30 minute prediction horizon. The predicted concentrations and the corresponding reference concentrations are illustrated with white circles, and the estimated risk of a predicted concentration given the ground truth reference concentration is illustrated by color in the plot. The risk zones are divided into four main risk categories: none, mild, moderate and high.

In Table 3 we present the mean surveillance error and the standard deviation on the test data for the 100 different random seeds for each patient individually and a mean and standard deviation for all patients. We can see that the performance is worse for patient 575 than for patient 570, but according to this metric the model performs worst on patient 591.

In Figure 4 we see that the predictions for patient 570 are mostly concentrated to the none and mild risk regions, but for patient 575 we can see that there are a few predictions in the moderate to high risk regions as well.

Noise experiments: To get insight into what uncertainty the model is able to learn we have conducted three experiments to isolate different types of noise

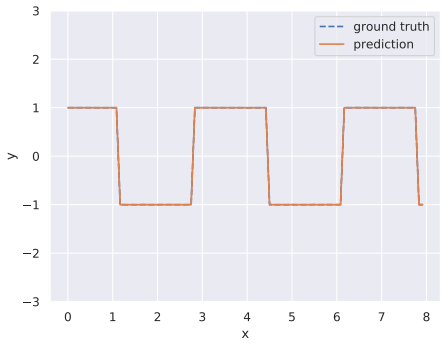


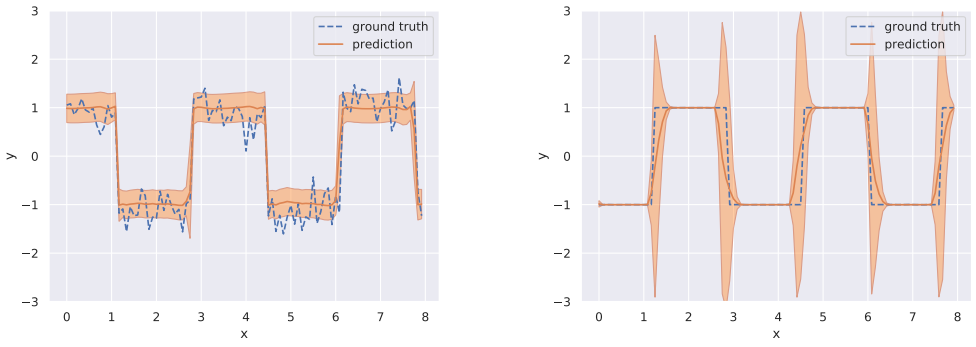
Fig. 5: The predictions from the proposed model trained on a deterministic squared waveform with step length 20 and states in -1 and 1. The predicted mean is plotted in orange and the predicted standard deviation is plotted in shaded orange. The signal we train on is plotted in blue. The ground truth signal is not visible in the plot since the model solves the problem and the predictions occlude the ground truth.

added to a deterministic signal. The deterministic signal is a simple squared waveform with a step length of 20 and two state values of -1 and 1 (see Figure 5). We add two types of noise which we will call measurement noise and state length noise. The measurement noise is drawn from a normal distribution with a zero mean and a standard deviation of 0.3 and is simply added to the state value (see Figure 6a). The state length noise is drawn from a normal distribution with a zero mean and a standard deviation of 3 and is added to the step length of the waveform, i.e., the length we stay in each state is normally distributed with a mean of 20 and a standard deviation of 3 (see Figure 6b). The experiment with measurement noise indicate that the model learns to attribute a higher uncertainty to the prediction, when the CGM is giving readings with higher noise levels. The experiment with noisy state length is set up in such a way that the model can not know when the state change will occur, and that this uncertainty gets higher the longer we have stayed in a state. We can see that the model learns to attribute high uncertainty to predictions that are made close to a state change.

4 Discussion

In this paper, we have proposed a recurrent neural network model that can predict blood glucose levels in type-1 diabetes for horizons of up to 60 minutes into the future using only blood glucose level as inputs. We achieve results comparable to state-of-the-art methods on the standard Ohio T1DM dataset for blood glucose level prediction.

End-to-end learning: Our results suggest that end-to-end machine learning is feasible for precision health management. This allows the system to learn all internal representations of the data, and reduces the human effort involved — avoiding labor-intensive prior work by experts hand-crafting features based on extensive domain knowledge.



(a) (b)

Fig. 6: The predictions from the proposed model trained on a waveform signal with step length of 20 and states -1 and 1 with an added noise drawn form a normal distribution with mean 0 and standard deviation 0.3 (a), and a waveform with a step length of 20 with an added noise to the step length drawn from a normal distribution with mean zero and standard deviation 3.0 (b). The predicted mean is plotted in orange and the predicted standard deviation is plotted in shaded orange. The signal we train on is plotted in blue.

Quantifying uncertainty: Our model gives an estimate of the standard deviation of the prediction. This is a useful aspect for a system which will be used by CGM users for making decisions about administration of insulin and/or caloric intake. The predicted standard deviation can also be a useful signal for downstream components in a closed loop system, making automatic decisions for a patient. The results in Figure 3 show the predicted standard deviation for patient 570 and patient 575, the ones where the model is the most and the least successful in prediction accuracy, respectively. One principal problem is that disambiguating between intra-patient variation and sensor errors is unlikely to be feasible.

Physiological loss function: To our surprise we did not see improvements when using a physiologically motivated loss function [10] for training (results not shown). This is essentially a smoothed version of the Clarke error grid [7]. Of course our findings are not proof that such loss functions cannot improve results. Possibly a larger-scale investigation, exploring in particular a larger area of the parameter space and different training regimes might provide further insights. Penalizing errors for hypo- or hyper-glycemic states should lead to better real-world performance, as we observed comparatively larger deviations in minima and maxima. One explanation for that is the relative class imbalance, as extrema are rare. This could be countered with data augmentation techniques.

Model selection: Even though the different patients pose varying challenges for the prediction task (see Figure 2), we obtain the best result when training our model on the training data from all patients at once. This suggest that there are patterns governing blood glucose variability that can generalize between different patients, and that the model benefit from having access to more data.

Missing data: There are gaps in the training data with missing values. Most of the gaps are less than 10 hours, but some of the gaps are more than 24 hours. The number of missing data points account for roughly 23 out of 263 days of the total amount of patient data or 9% of the data. The gaps could be filled using interpolation, but it is not immediately clear how this would affect either the training of the models, or the evaluation of the models since this would introduce artificial values. Filling a gap of 24 hours using interpolation would not result in realistic data. Instead we have chosen not to fill the gaps with artificial values and limit our models to be trained and evaluated only on real data. This has its own limitations since we can only consider prediction targets with enough glucose level history, and therefore not predict the initial values after a gap, but the advantage is that model training and evaluation is not biased by the introduction of artificial values.

Additional patient data: As mentioned in the description of the dataset there are other data self-reported by the patients such as meal times with carbohydrate estiamtes, times of exercise, sleep, work, stress, and illness; and measures of heart rate, galvanic skin response, skin temeperature, air temperature and step count. From the results in this work, we conclude that a simple setup using only CGM history obtains results that are on par with more complex solutions that do incorporate more features. It is well documented that the additional features do affect blood glucose dynamics but the dependencies may be more subtle and complex and thus harder to learn. This motivates further work to develop models that can leverage the additional information and make more accurate predictions.

5 Related work

Early work on predicting blood glucose levels from CGM data include Bremer, et.al. [4], who explored the predictability of data from CGM systems, and showed how you can make predictions based on autocorrelation functions. Sparacino, et.al. [23] proposed a first-order auto-regressive model.

Wiley [24] proposed using Support vector regression (SVR) to predict blood sugar levels from CGM data. They report RMSE of 4.5 mg/dl, but this is using data that was aggressively smoothed using a regularized cubic spline interpolation. Bunescu, et.al. [5] extended this work with physiological models for meal absorption dynamics, insulin dynamics, and glucose dynamics to predict blood glucose levels 30 and 60 minutes into the future. They obtained a relative improvement of about 12% in prediction accuracy over the model proposed by Wiley. The experiments in [5] is performed on non-smoothed data.

There have been approaches using neural networks to predict blood glucose levels. Perez, et.al. [22] presented a feed forward neural network (FFNN) taking CGM history as input, and predicting the level 15, 30, and 45 minutes into the future. RMSE accuracy for 30 minute predictions are similar to those of [24]. Mougiakakou, et.al. [20] showed that RNNs can be used to predict blood glucose levels from CGM data. They evaluated their method on four different

children with type-1 diabetes, and got some promising results. On average, they reported an RMSE accuracy of 24.1 mg/dl.

Some papers have incorporated additional information (e.g. carbohydrate/meal intake, insulin injections, etc). [21] proposed an FFNN taking as input CGM levels, insulin dosages, metered glucose levels, nutritional intake, lifestyle and emotional factors. Despite having all this data at its disposal, the model makes predictions 75 minutes into the future with an RMSE score of 43.9 mg/dl. [26] proposed a neural network approach in combination with a first-order polynomial extrapolation algorithm to produce short-term predictions on blood glucose levels, taking into account meal intake information. The approach is evaluated both on simulated data, and on real data from 9 patients with Abbott Freestyle Navigator. None of the above mentioned approaches have the ability to output a confidence interval.

A problem when modeling continuous outputs trained using least squares as a training criterion is that the model tends to learn a conditional average of the targets. Modeling a distribution over the outputs may limit this problem and make training more stable. Mixture density networks were proposed by [3]. By allowing the output vector from a neural network model to parameterize a mixture of Gaussians, they manage to learn a mapping even when the targets are not unique. Besides enabling learning stability, this also allows the model to visualize the certainty of its predictions. A similar approach was used together with RNNs in [11], to predict the distribution of next position for a pen during handwriting.

The release of the Ohio dataset [16] in combination with *The blood glucose level prediction challenge (BGLP)* at *The workshop on knowledge discovery in healthcare data (KDH)* 2018, spurred further interest on blood glucose prediction models. At the workshop, a preliminary version of this study was presented [17]. While a challenge was formulated, no clear winner could be decided, because of differences in evaluation procedure. The results listed below cannot directly be compared to the results in this paper due to these differences. However, they all refer to predictions made with a 30 minute horizon. While our study have focused on predicting the blood glucose levels using only the CGM history as input, all methods below use more features provided in the dataset such as carbohydrate intake and insulin distribution, and none of them gives an estimate of the uncertainty.

Chen, et.al. [6] used a recurrent neural network with dilations to model the data. Dilations allow a network to learn hierarchical structures and the authors chose to use the CGM values, insulin doses, and carbohydrate intake from the data, resulting in an average RMSE of 19.04 mg/dl. Xie, et.al. [25] compared autoregression with exogeneous inputs (ARX) with RNNs and convolutional neural networks (CNNs), and concluded that the simpler ARX models achieved the best scores on the Ohio blood glucose data, with an average RMSE of 19.59 mg/dl. Contreras, et.al. [8] used grammatical evolution (GE) in combination with feature engineering to search for a predictive model, obtaining an average RMSE of 24.83 mg/dl. Bertachi, et.al. [2] reported an average RMSE of 19.33 mg/dl by using physiological models for insulin onboard, carbohydrates onboard, and activity onboard, which are fed as features to a feed forward neural

network. Midroni, et.al. [18] employed XGBoost with a thorough investigation of feature importance and reported an average RMSE of 19.32 mg/dl. Zhu, et.al. [27] trained a CNN with CGM data, carbohydrate intake, and insulin distribution used as features and obtained an average RMSE of 21.72 mg/dl.

6 Conclusions

In this paper, we presented a deep neural network model that learns to predict blood glucose levels up to 60 minutes into the future. The model parameterize a univariate Gaussian output distribution, facilitating an estimate of uncertainty in the prediction. Our results make a clear improvement over the baseline, and motivate future work in this direction.

However, it is clear that the field is in desperate need of larger data sets and standards for the evaluation. Crowd sourcing from patient associations would be one possibility, but differences in sensor types and sensor revisions, life styles, and genetic mark-up are all obvious confounding factors. Understanding sensor errors by measuring glucose level in vivo, for example in diabetes animal models, with several sensors simultaneously would be very insightful, and likely improve prediction quality. Another question concerns preprocessing in the sensors, which might be another confounding factor in the prediction. While protection of proprietary intellectual property is necessary, there has been examples, e.g. DNA microarray technology, where only a completely open analysis process from the initial steps usually performed with vendor's software tools to the final result helped to realize the full potential of the technology.

Acknowledgments

The authors would like to thank Christian Meijner and Simon Persson who performed early experiments in this project.

Conflicts of interest

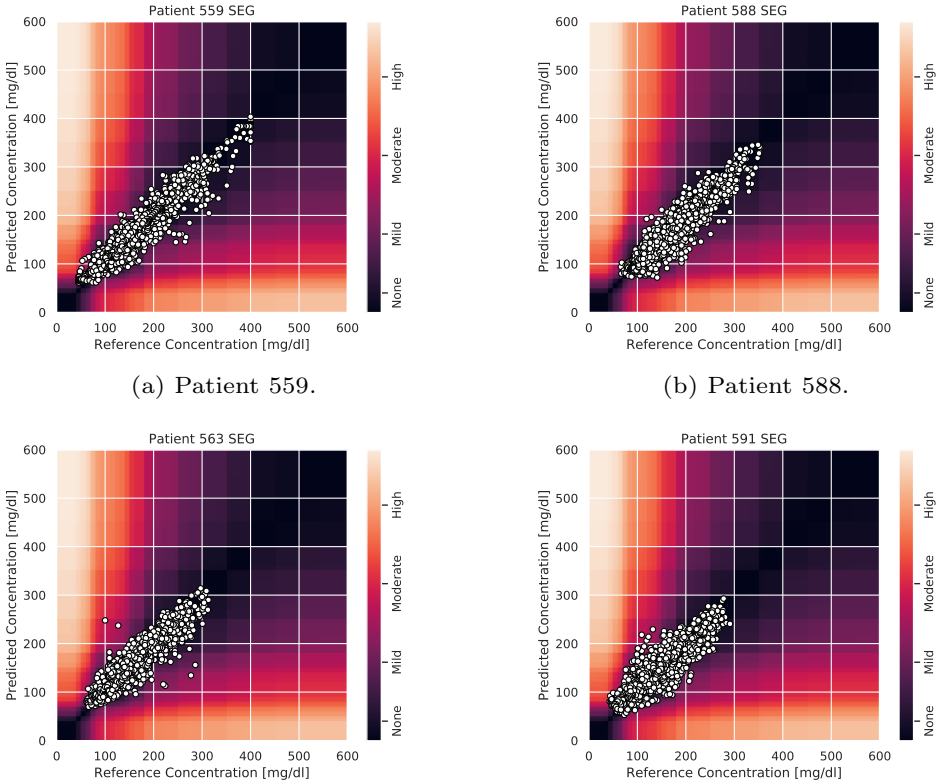
On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. Neural Networks, IEEE Transactions on **5**(2), 157–166 (1994)
2. Bertachi, A., Biagi, L., Contreras, I., Luo, N., Vehi, J.: Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 85–90 (2018)
3. Bishop, C.M.: Mixture density networks. Tech. rep., Citeseer (1994)

4. Bremer, T., Gough, D.A.: Is blood glucose predictable from previous values? a solicitation for data. Diabetes **48**(3), 445–451 (1999)
5. Bunescu, R., Struble, N., Marling, C., Shubrook, J., Schwartz, F.: Blood glucose level prediction using physiological models and support vector regression. In: Machine Learning and Applications (ICMLA), 2013 12th International Conference on, vol. 1, pp. 135–140. IEEE (2013)
6. Chen, J., Li, K., Herrero, P., Zhu, T., Georgiou, P.: Dilated recurrent neural network for short-time prediction of glucose concentration. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 69–73 (2018)
7. Clarke, W.L., Cox, D., Gonder-Frederick, L.A., Carter, W., Pohl, S.L.: Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes care **10**(5), 622–628 (1987)
8. Contreras, I., Bertachi, A., Biagi, L., Vehi, J., Oviedo, S.: Using grammatical evolution to generate short-term blood glucose prediction models. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 91–96 (2018)
9. Cryer, P.E., Davis, S.N., Shamoon, H.: Hypoglycemia in diabetes. Diabetes care **26**(6), 1902–1912 (2003)
10. Favero, S.D., Facchinetti, A., Cobelli, C.: A Glucose-Specific Metric to Assess Predictors and Identify Models **59**(5), 1281–1290 (2012)
11. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
12. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **6**(02), 107–116 (1998)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Klonoff, D.C., Lias, C., Vigersky, R., Clarke, W., Parkes, J.L., Sacks, D.B., Kirkman, M.S., Kovatchev, B., the Error Grid Panel: The surveillance error grid. Journal of Diabetes Science and Technology **8**(4), 658–672 (2014). DOI 10.1177/1932296814539589. URL `https://doi.org/10.1177/1932296814539589`. PMID: 25562886
15. Lewis, D., Leibrand, S., Community, .O.: Real-world use of open source artificial pancreas systems. Journal of diabetes science and technology **10**(6), 1411–1411 (2016)
16. Marling, C., Bunescu, R.: The OhioT1DM dataset for blood glucose level prediction. In: The 3rd International Workshop on Knowledge Discovery in Healthcare Data. Stockholm, Sweden (2018). CEUR proceedings in press, available at http://smarthealth.cs.ohio.edu/bglp/OhioT1DM-dataset-paper.pdf
17. Martinsson, J., Schliep, A., Eliasson, B., Meijner, C., Persson, S., Mogren, O.: Automatic blood glucose prediction with confidence using recurrent neural networks. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 64–68 (2018)
18. Midroni, C., Leimbigler, P., Baruah, G., Kolla, M., Whitehead, A., Fossat, Y.: Predicting glycemia in type 1 diabetes patients: Experiments with xg-boost. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 79–84 (2018)
19. Mirshekarian, S., Bunescu, R., Marling, C., Schwartz, F.: Using LSTMs to learn physiological models of blood glucose behavior. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS pp. 2887–2891 (2017). DOI 10.1109/EMBC.2017.8037460
20. Mougiakakou, S.G., Prountzou, A., Iliopoulou, D., Nikita, K.S., Vazeou, A., Bartsocas, C.S.: Neural network based glucose-insulin metabolism models for children with type 1 diabetes. In: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pp. 3545–3548. IEEE (2006)
21. Pappada, S.M., Cameron, B.D., Rosman, P.M., Bourey, R.E., Papadimos, T.J., Olorunto, W., Borst, M.J.: Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. Diabetes technology & therapeutics **13**(2), 135–141 (2011)

22. Pérez-Gandía, C., Facchinetti, A., Sparacino, G., Cobelli, C., Gómez, E., Rigla, M., de Leiva, A., Hernando, M.: Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes technology & therapeutics **12**(1), 81–88 (2010)
23. Sparacino, G., Zanderigo, F., Corazza, S., Maran, A., Facchinetti, A., Cobelli, C.: Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. IEEE Transactions on biomedical engineering **54**(5), 931–937 (2007)
24. Wiley, M.T.: Machine learning for diabetes decision support (pp. 55-72) (2011)
25. Xie, J., Wang, Q.: Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 97–102 (2018)
26. Zecchin, C., Facchinetti, A., Sparacino, G., De Nicolao, G., Cobelli, C.: Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration (6), 1550–1560 (2012)
27. Zhu, T., Li, K., Herrero, P., Chen, J., Georgiou, P.: A deep learning algorithm for personalized blood glucose prediction. In: 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ ICML/IJCAI 2018, 13 July 2018, pp. 64–78 (2018)



(a) Patient 559.

(b) Patient 588.

(c) Patient 563.

(d) Patient 591.

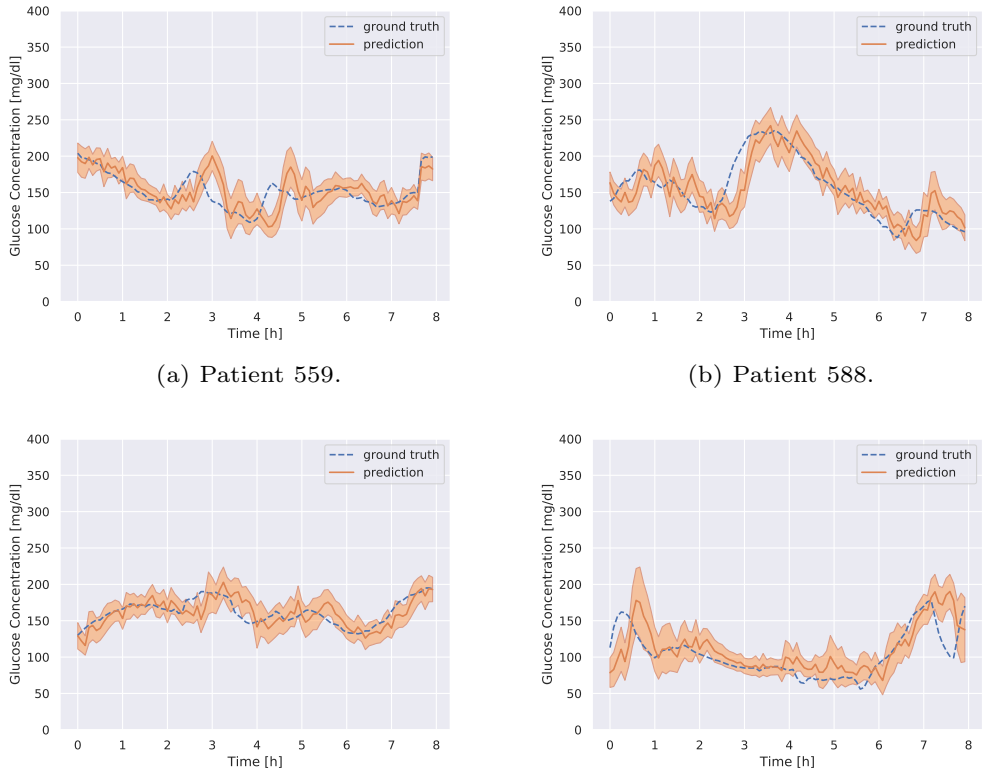
Fig. 7: The surveillance error grid overlayed with each model prediction concentration and reference concentration for all patients. The predicted concentrations and the corresponding reference concentrations are illustrated with white circles, and the estimated risk of a predicted concentration given the ground truth reference concentration is illustrated by color in the plot. The risk zones are divided into four main risk categories: none, mild, moderate and high.

A Software

The software including all scripts to reproduce the computational experiments is released under an open-source license and available from `https://github.com/johnmartinsson/blood-glucose-prediction`. We have used Googles TensorFlow framework, in particular the Keras API of TensorFlow which allows for rapid prototyping of deep learning models, to implement our model and loss functions.

B Additional figures

In this appendix we have included additional surveillance error grid plots (see Figure 7) and prediction plots (see Figure 8) for the four patients that are not presented in the results section.



(a) Patient 559.

(b) Patient 588.

(c) Patient 563.

(d) Patient 591.

Fig. 8: Prediction (orange), the predicted standard deviation (shaded orange) and the ground truth glucose concentration (dashed blue) for all patients. The plot shows eight hours of predictions starting from an arbitrarily chosen time for each patient. The predictions are 30 minutes into the future.