Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Accelerating Machine Learning in Health Care by Addressing the Labeling Bottleneck: Development and Evaluation Study

Spencer Vecile

spencer.vecile@sickkids.ca

Hospital for Sick Children

William Dixon

Hospital for Sick Children

Azadeh Assadi

Hospital for Sick Children

Michael Kim

Seattle Children's Hospital

Robert Greer

Hospital for Sick Children

Asad Siddiqui

Hospital for Sick Children

Daniel Ehrmann

University of Michigan Medical School

Andrew Goodwin

Hospital for Sick Children

Danny Eytan

Rambam Health Care Campus

Mjaye Mazwi

Seattle Children's Hospital

Anica Bulic

Hospital for Sick Children

Sebastian D. Goodfellow

Hospital for Sick Children

Research Article

Keywords: Arrhythmia detection, Machine Learning, Dataset Labeling, Electrocardiogram (ECG), active learning, rare class enrichment, clinician-in-the-loop annotation, pediatric critical care

Posted Date: October 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8194963/v3

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Accelerating Machine Learning in Health Care by Addressing the Labeling Bottleneck: Development and Evaluation Study

Spencer Vecile[1], William Dixon[1], Azadeh Assadi[1,2], Michael Kim[3], Robert Greer[4], Asad Siddiqui[5,6], Daniel Ehrmann[7], Andrew J. Goodwin[1,8], Danny Eytan[9,10], Mjaye Mazwi[3], Anica Bulic[11], Sebastian D. Goodfellow[1,12]

Affiliations

1. Department of Critical Care Medicine, The Hospital for Sick Children, Toronto, ON, Canada.
2. Lawrence S. Bloomberg Faculty of Medicine, University of Toronto, Toronto, ON, Canada
3. Division of Cardiac Critical Care, Department of Pediatrics, Seattle Children's Hospital, Seattle, Washington, USA
4. SickKids Artificial Intelligence, The Hospital for Sick Children, Toronto, ON, Canada.
5. Department of Anesthesia and Pain Medicine, The Hospital for Sick Children, Toronto, ON, Canada
6. Department of Anesthesiology and Pain Medicine, University of Toronto, Toronto, ON, Canada.
7. University of Michigan Medical School, Department of Pediatrics, Division of Cardiology, Ann Arbor, Michigan, USA
8. School of Biomedical Engineering, University of Sydney, Sydney, Australia
9. Pediatric critical care unit, Rambam Medical Center, Haifa, Israel
10. The faculty of medicine, The Technion, Haifa, Israel
11. Department of Pediatric Cardiology, The Hospital for Sick Children, Toronto, ON, Canada
12. Department of Civil & Mineral Engineering, University of Toronto, Toronto, ON, Canada

Abstract

Background: Timely detection of postoperative arrhythmias after pediatric cardiac surgery is essential for preventing hemodynamic compromise, yet machine learning models are constrained by the scarcity of labelled datasets reflecting real-world monitoring. Existing approaches rely on adult 12-lead electrocardiograms, which are rarely used continuously in pediatric intensive care units (ICUs) and fail to capture age-dependent waveform variability. Rare arrhythmias such as junctional ectopic tachycardia are especially difficult to find and label within large continuous monitoring streams.

Objective: We evaluated how efficiently a staged, self-bootstrapping labeling pipeline surfaces a rare arrhythmia class relative to naive random review of an unlabelled dataset, and how that efficiency evolves across the pipeline's dependent stages.

Methods: Leveraging a repository of over 1.6 million hours of unlabelled continuous lead II electrocardiogram (ECG) from more than 9,000 pediatric patients, we implemented four sequential sourcing stages, each enabling the next: retrospective MUSE (GE Healthcare Inc.) studies and real-time ICU observations produced an expert-labelled seed set, which trained a preliminary classifier that then guided two active-learning stages, uncertainty sampling and embedding-based retrieval. All labels were assigned by clinical experts; the classifier was used only to prioritize which segments they reviewed. To benchmark the pipeline, we labelled 200 randomly drawn segments and compared junctional yield using the Fisher exact test.

Results: Random sampling yielded junctional arrhythmia in only 2 of 200 segments (1.0%). The full pipeline produced a dataset that was 15.3% junctional by labelled hours and 24.7% by patient, a roughly 15-fold and 25-fold enrichment over random sampling, respectively (Fisher exact test, P<.001). Yield rose at each successive stage, from 7.6% of labelled hours (MUSE) to 35.7% (embedding retrieval), reflecting efficiency compounding as the pipeline matured. Of segments retrieved from the junctional embedding region, 60.2% were confirmed junctional once labelled. The final dataset comprised 181.2 hours of expert-labelled ECG from 1,447 unique patients.

Conclusions: A staged pipeline that bootstraps model-guided sampling from a small expert-labelled seed set enriches a rare arrhythmia class roughly 15-fold over random review, directing scarce clinician effort toward informative segments. We make no claim about the guiding model's diagnostic performance, which will be reported separately. This methodology is broadly applicable to clinical domains where expert annotation is limited and rare events dominate.

Keywords: Arrhythmia detection; Machine learning; Dataset labeling; Electrocardiogram (ECG); Active learning; Rare class enrichment; Clinician-in-the-loop annotation; Pediatric critical care

Introduction

Arrhythmias are a common complication following pediatric congenital heart surgery and are associated with increased morbidity and mortality[1]. Certain arrhythmias, such as junctional ectopic tachycardia (JET), can be particularly harmful, potentially resulting in hemodynamic instability and circulatory collapse if not identified and treated in a timely manner[1]. Delays in diagnosis may arise from a range of factors, including limited experience among front-line providers, systemic constraints such as provider availability, and other operational challenges.

Machine learning (ML)-based clinical decision support (CDS) systems have emerged as a promising tool to improve the efficiency of arrhythmia detection to facilitate prompt diagnosis and intervention. Most existing ML models have been developed using adult datasets and rely heavily on 12-lead electrocardiogram (ECG) recordings[2] which has limited utility in pediatric populations. These models are not well-suited to the pediatric intensive care unit (ICU) setting, where arrhythmia detection is typically initiated by clinicians using bedside monitors that display only one to three ECG leads. The 12-lead ECG is generally reserved for confirmation after an arrhythmia is suspected. Furthermore, age-dependent variations in ECG waveform characteristics, including heart rate, R-R intervals, and amplitude limit the applicability of adult data for pediatric ML-based CDS[3].

Existing research often depends on publicly available datasets like MIT-BIH[4], PTB-XL[5], PhysioNet 2021[6], and MIMIC-IV-ECG[7], which focus on adult populations and 12-lead recordings, making them unsuitable for detecting JET in children. Prior studies[8–10] have trained models using 12-lead or multi-lead ECG data, which typically requires prior clinical suspicion and manual acquisition. In contrast, our approach utilizes continuous lead II ECG monitoring, which reflects standard ICU practice where pediatric patients rarely have all 12 leads attached. This strategy also enables real-time arrhythmia detection without relying on clinician-initiated ECG acquisition.

A critical barrier to developing effective ML models for pediatric arrhythmia detection is the lack of large, labelled datasets that reflect the clinical reality of continuous three-lead monitoring[11]. Research often relies on 12-lead ECGs because they are closely associated with confirmed diagnoses, which simplifies retrospective labeling and makes it easier to identify rare arrhythmia classes. However, these datasets do not capture the continuous monitoring streams used in practice, where underrepresented rhythms are much more difficult to find and label. Clinicians who provide expert labels face substantial time constraints due to clinical responsibilities, and this limits the amount of manual annotation they can perform. The methodology presented in this paper addresses this challenge by increasing the likelihood that a sample selected for labeling contains an underrepresented class. Enabling more efficient mining of a dataset, which allows clinicians to focus their limited time on labeling underrepresented rhythms rather than on classes that are already well represented. Through this approach, we construct a labelled pediatric ECG dataset that is enriched with junctional arrhythmias, using a scalable and clinically informed strategy for deriving meaningful labels from an extensive unlabelled AtriumDB[12] dataset. With the goal of supporting the development of robust and deployable ML models for pediatric arrhythmia detection.

We frame this work as an evaluation of labeling efficiency. The four sourcing strategies described below are not competing alternatives but a dependent sequence: retrospective (MUSE) and ICU observation-based labels form an initial expert-labelled seed set, that seed set makes it possible to train a preliminary classifier, and that classifier in turn enables the two model-guided strategies. Our central question is therefore how efficiently this staged, self-bootstrapping pipeline surfaces the rare class relative to the alternative, naive random review of the dataset, and how that efficiency evolves across the pipeline's dependent stages. The classifier is used solely to prioritize candidate segments for expert review; we make no claim about its diagnostic performance, which is reported separately.

Methods

AtriumDB: Physiologic Database

We developed two integrated platforms, AtriumDB and AtriumStreams (see Figure 1), to enable both large-scale physiologic data capture and real-time model deployment. AtriumDB is a timeseries database specifically designed for use with physiological waveforms[12] and has been operational in 42 bedspaces at The Hospital for Sick Children (SickKids) ICU since 2016. It has supported the collection of over 1.6 million hours of continuous lead II ECG data from more than 9,000 critically ill children in the ICU and operating rooms leading to one of the largest pediatric physiologic waveform datasets in the world. AtriumDB includes a dedicated python library that enables rapid data extraction at scale, specifically designed to support high-throughput deep learning workflows. Complementing this, AtriumStreams is a real-time deployment engine that ingests real- time waveforms directly from bedside monitors, enabling real-time inference at the bedside. Together, AtriumDB and AtriumStreams function as the foundation of our broader translational pipeline. This system has now been deployed at 10 international hospitals in both pediatric and adult settings (as of September 2025).

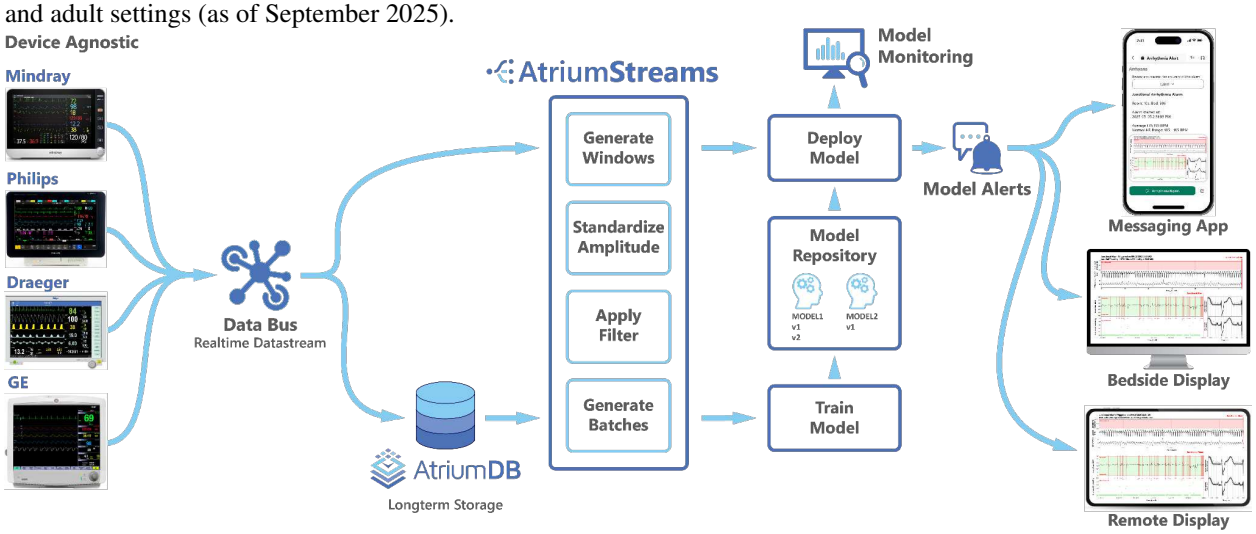


Figure 1. Overview of the AtriumDB Platform.

Ethics Approval

The ECG dataset used in this study is part of a clinical AtriumDB-based dataset, collected prospectively from a 42-bed ICU at The Hospital for Sick Children (SickKids) in Toronto, ON since 2016. Secondary use of this dataset is governed by the Hospital for Sick Children's (SickKids) Research and Ethics board approval (REB#1000068499) and has a waiver of consent.

Label Schema

To support dataset labelling, we developed a hierarchical label schema (Figure 2) grounded in pediatric electrophysiology practice and iteratively refined through expert input. Labels were applied to lead II ECG strips from the bedside monitor as described in more detail below. The schema organizes rhythm types into five clinically meaningful parent categories: Sinus Rhythms, Blocks, Paced Rhythms, Ventricular Arrhythmias, and Supraventricular (SVT) Arrhythmias. Each parent class includes a set of child labels representing specific rhythm diagnoses encountered in pediatric critical care. The "unknown" class is excluded from the training data and is available for when samples aren't exactly noise but clinicians also can't confidently determine the rhythm class.

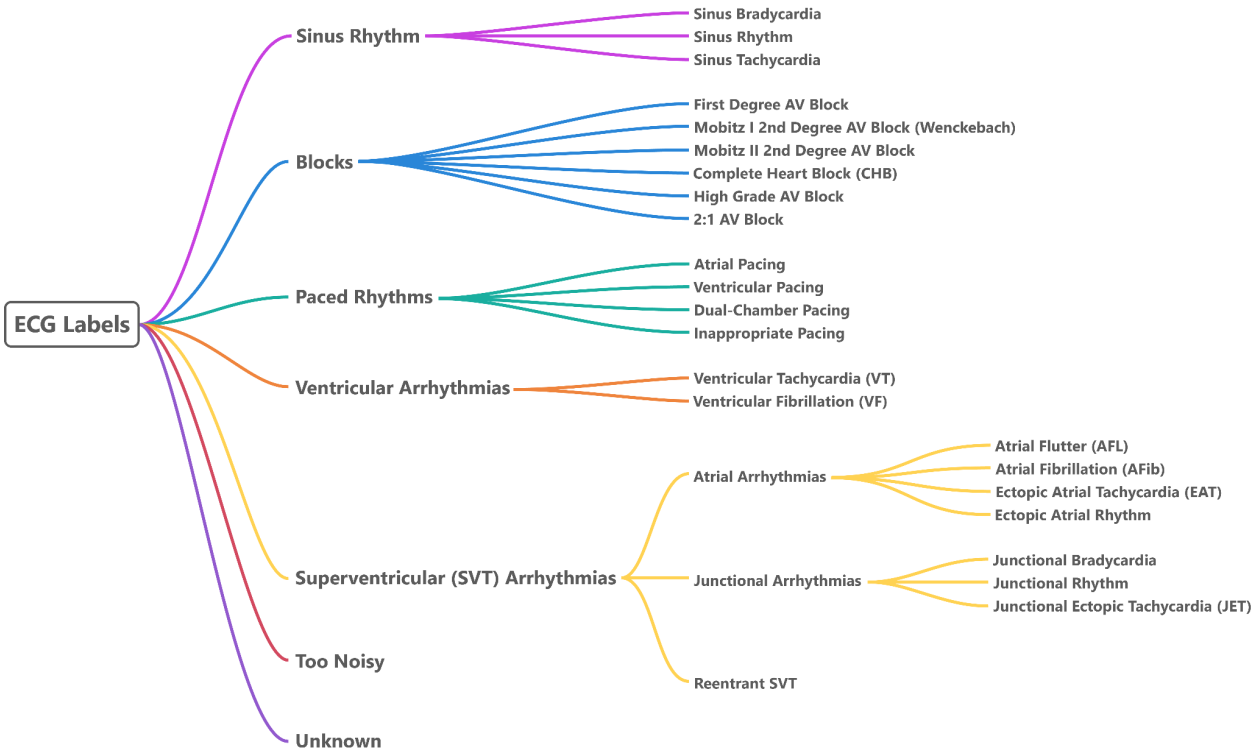


Figure 2. Hierarchical arrhythmia label schema used for dataset labeling. Going from left to right the labels go from more general parent classes to more specific child classes.

Initial Labelling Workflow

Ground Truth Labels

All ECG labels used for training and evaluation were annotated by a team of pediatric intensive care clinicians with specialized cardiac training operating under the direction of a senior pediatric electrophysiologist. Given the limited availability of expert clinicians, we adopted a distributed annotation strategy to maximize efficiency, limiting labelling of each individual ECG segment to a single labeller to avoid duplicating effort. When uncertainties arose, annotators maintained active communication with the senior electrophysiologist to resolve ambiguities and uphold labelling integrity. Before initiating large-scale annotation, we conducted a study on 200 two-minute ECG samples to assess whether single-annotator labeling was feasible under real-world resource constraints. The scarcity of clinician time is common in clinical environments, and this design reflects that practical reality. This assessment yielded an average Cohen's kappa of 0.65, indicating sufficient inter-rater agreement. On this basis, labels from individual clinicians were accepted as ground truth. Further details of this study can be found in Appendix 1.

Labelling Software

To train our arrhythmia classifier, which assigns rhythm labels to 5-second segments of lead II ECG, we developed a clinically integrated labelling pipeline built around Label Studio, an open-source annotation platform (Figure 3). The interface was custom-designed with direct input from pediatric electrophysiologists, enabling users to securely log in and label de-identified waveform data through a web-based environment. Label Studio's native support for time-series

data, combined with its customizable HTML framework, allowed us to tailor the interface for efficient ECG annotation at scale.

Our interface presented both a 2-5 minute lead II ECG waveform strip and synchronized heart rate signal (sampled at 1 Hz). If a 12-lead ECG was obtained with a confirmed diagnosis in the MUSE system with a timestamp corresponding to the telemetry strip in question, the annotation was included as contextual metadata on the labelling interface. Label selection occurs through a structured dropdown menu derived from the hierarchical schema (Figure 2).

The hierarchical schema supports both coarse and fine-grained annotation. Annotators can apply a parent label (e.g., “Ventricular Arrhythmia”) or designate more specific subtypes via child labels. This structure enabled flexible model training strategies (e.g. grouping rare subtypes under a parent class to mitigate class imbalance). During annotation, users first selected a predominant rhythm for the full waveform and then mark specific regions (e.g., noise artifacts or other arrhythmias) where applicable. For instance, a 2-minute segment may be labelled globally as Sinus Rhythm, with one or more short X-second regions identified as another rhythm, which significantly reduces annotation time compared to fully segmented labelling.

To mitigate performance constraints inherent to Label Studio, which exhibits latency and interaction delays beyond approximately 300,000 data points, or 10 minutes of ECG sampled at 500 Hz, we adopted a segmentation strategy that partitioned longer waveform regions into contiguous 2-5 minute chunks. These segments were grouped as sequential tasks, with consistent annotations in the user interface denoting patient identity and temporal ordering. This approach preserved critical clinical context while maintaining interface responsiveness.

To maintain synchronization with our physiologic data platform, we implemented webhook functionality within Label Studio. These webhooks notify a custom Flask server each time a label is created, modified, or deleted, enabling real-time updates to an AtriumDB dataset and supporting rapid retraining cycles as new labelled data becomes available. This continuous feedback loop between expert labelling and model development is central to our clinician-in-the-loop workflow and ensures that improvements in the dataset immediately translate into model refinement.

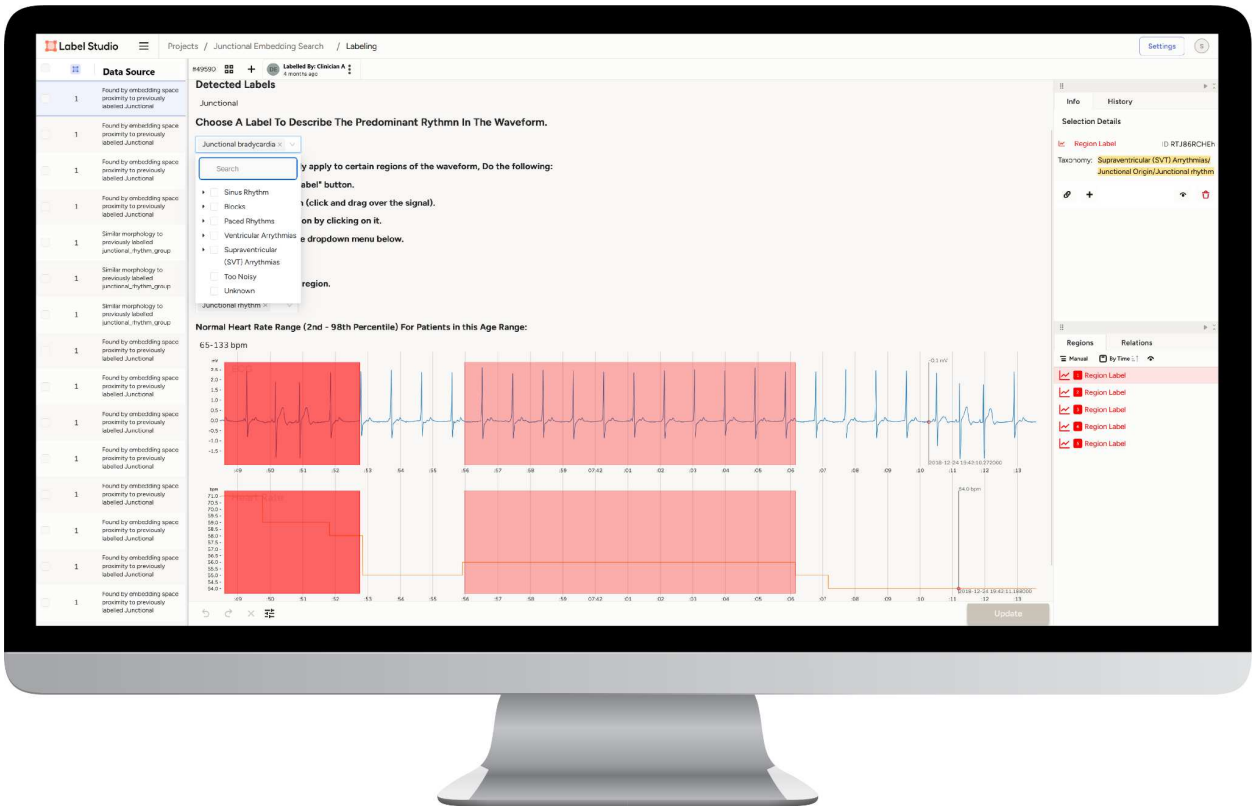


Figure 3. Custom Label Studio interface for ECG annotation. The dropdown box has the label schema encoded into it and allows labellers to quickly select the rhythm. At the bottom of the screen you can see the lead II ECG tracing with the heart rate plot below it. The red boxes are region labels and make a label only applicable to that space in time. On the far right the tasks that are apart of the project are listed.

Summary of Different Label Sources

To construct a high-quality training dataset for the arrhythmia classifier, we implemented a multi-pronged strategy to source labels from both retrospective and prospective data streams. Given the rarity and subtlety of JET, no single labelling approach was sufficient. Instead, we leveraged four complementary techniques to enrich the dataset with arrhythmias of interest. We started with MUSE and applied each of the four techniques mentioned below in order, moving onto the next once the current one hit diminishing returns or we could apply a better technique:

Step 1. MUSE: Structured diagnoses associated with 12-lead ECG studies. (human only)
Step 2. ICU Observations: Bedside annotations triggered in real time by clinical staff. (human only)
Step 3. Uncertainty Sampling: Model-guided selection of ambiguous or low-confidence windows for labeling by clinical experts. (model guided selection with human annotations)
Step 4. Embedding Search: Retrieval of morphologically similar unlabelled ECG segments via model-derived feature embeddings for labeling by clinical experts. (model guided selection with human annotations)

Each of the four strategies mentioned above yielded different levels of labelling efficiency, patient coverage, and arrhythmia density.

Sourcing Labels from MUSE

Our initial strategy for sourcing arrhythmia labels used the presence of a formal 12-lead ECG study as a proxy indicator for an arrhythmia event. Specifically, we extracted all 12-lead studies from the MUSE system at SickKids and aligned them with continuous ECG waveform data stored in an AtriumDB dataset by matching timestamps. The diagnostic interpretation associated with each 12-lead study, typically authored by an electrophysiologist or generated via rule-based algorithms, was attached to the corresponding ECG segment and presented to expert labellers in Label Studio for verification.

However, this approach introduced two critical challenges. First, we frequently observed discrepancies between the diagnostic interpretation derived from the 12-lead ECG and the rhythm classification provided by expert annotators based on the corresponding waveform segments. This was due to a clock mismatch between the Philips monitor and the MUSE system that we could not quantify nor correct for. Consequently, the ECG strips presented to annotators often did not correspond precisely to the diagnostic moment captured by the 12-lead system, thereby compromising the validity of the associated diagnostic labels. These misalignments necessitated manual re-labeling of all ECG segments, substantially increasing the annotation burden.

Second, even when correctly aligned, the segments sourced via 12-lead studies were predominantly composed of normal sinus rhythm (73% in our cohort). While sinus rhythm examples are essential for training a balanced classifier, the disproportionate prevalence of non-informative segments significantly limited the efficiency of expert labelling efforts. Given the constraints on clinician time, it became clear that this method was not viable as a primary mechanism for surfacing high-value arrhythmia examples.

Sourcing Labels from ICU Observations

To surface clinically relevant arrhythmia examples in real time, we implemented a clinician-in-the-loop annotation workflow that leverages the situational awareness of ICU providers. Initially deployed via Slack (Salesforce Inc.), a secure mobile messaging platform, the system allowed nurses and physicians to rapidly report observed arrhythmias through a structured form described in Figure 4. Upon submission, the input was automatically transmitted to a Flask-

based server, which queried the dataset and extracted two segments: a one hour window following the reported start time and a 20-minute window at the end of the encounter, which typically contained Sinus Rhythm.

This dual-segment retrieval strategy was designed to provide the model not only with positive examples of arrhythmias but also with patient-specific sinus rhythm morphology, facilitating finer-grained contrastive learning. In cases where the patient encounter had not yet ended at the time of submission, a background scheduler re-queried the dataset every five minutes until the terminal segment became available and could be uploaded for labelling. This flexible, low-friction reporting mechanism enabled bedside clinicians to contribute high-yield labelling candidates with minimal interruption to care, significantly increasing the number of rare rhythm examples, particularly junctional arrhythmias, compared to passive approaches. The integration of annotation triggers into tools already used during clinical care not only enhanced data relevance but also improved labeller efficiency by reducing the proportion of sinus rhythm segments.

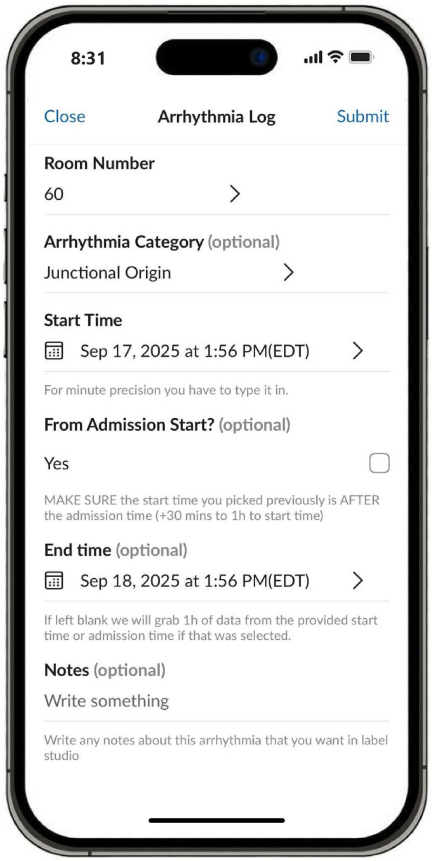


Figure 4. Slack-based clinician workflow for reporting arrhythmia events in real time. Designed for minimal workflow disruption, the form requires only two fields, patient room number and start time, while offering optional fields such as arrhythmia category (restricted to parent-level classes) and free-text clinical notes. These optional fields were later used to filter segments in Label Studio by rhythm class. The form also includes a "from admission start" checkbox, enabling clinicians to estimate timing without needing to retrieve the precise encounter start time, further reducing workflow disruption.

Following completion of ICU-based labeling, the resulting dataset contained approximately 80 minutes of annotated ECG data per patient but was limited by a relatively small cohort size. Figure 5 illustrates the clustering of all labelled segments, where embeddings from the same patient exhibit tight grouping, indicating minimal intra-patient variability within rhythm classes. While labeling one hour per patient provided valuable initial coverage, this approach is highly time-consuming for clinicians and yields diminishing returns because ECG signals are inherently repetitive

within a single patient. Nevertheless, it was essential for training an initial model capable of supporting model-guided sample selection. To improve generalization and reduce annotation burden moving forward, we aim to increase inter-patient diversity by labeling shorter segments (two minutes) from a larger number of unique patients. This strategy maximizes variability while minimizing clinician workload. To operationalize this approach, we adopted a model-guided labeling framework.

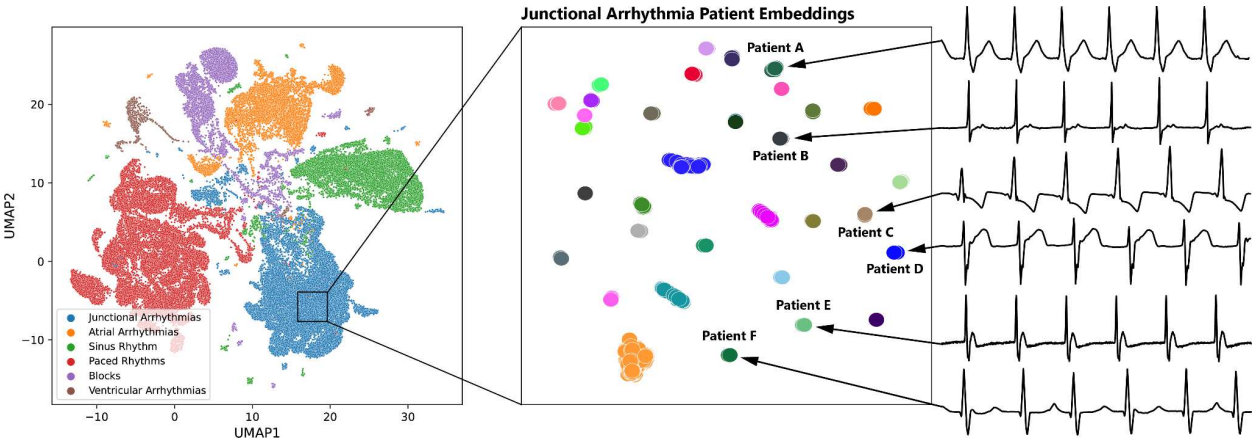


Figure 5. Uniform Manifold Approximation and Projection (UMAP)[13] projection of the model's embedding space. The left subplot displays the full set of labelled ECG segments colored by their parent rhythm class. On the right, six representative ECG segments from six unique patients from within the junctional cluster illustrate the morphological diversity of junctional arrhythmias.

Model Guided Labelling Workflow

The final two labeling strategies leverage a machine learning model trained on data obtained through the initial labeling methods to identify additional candidate segments within the dataset for our labellers to label. Although the model does not yet demonstrate sufficient performance for clinical deployment, it is capable of highlighting regions within the unlabelled data that are likely to contain arrhythmias. This enables a targeted labelling approach, focusing on rhythm classes the model finds challenging or are underrepresented in the current dataset. To contextualize these strategies, we first provide a summary of the model architecture and training process. We report no performance metrics for this model in the present work; it functions here only as a mechanism for prioritizing candidate segments for expert review, and its evaluation and further technical details are deferred to a subsequent publication.

Foundational Model

The foundation model was trained on continuous lead II ECG data collected from the SickKids ICU. It uses a Wave2Vec 2.0-based[14] architecture, which includes a 4-layer convolutional encoder followed by 12 transformer blocks configured similarly to BERT-Base[15]. Training was performed using Contrastive Masked Segment Coding (CMSC)[9], a self-supervised learning approach that combines local and global contrastive objectives. Locally, the model learns to reconstruct masked ECG segments from the surrounding context, while globally it learns patient-level rhythm patterns by comparing adjacent segments from the same patient against segments from others. This enables the model to learn temporally-rich and morphology-aware representations from unlabelled data.

Arrhythmia Classifier

To adapt the foundation model for supervised classification, self-supervised components such as masking and contrastive losses were removed. A lightweight classification head was added (Figure 6), consisting of temporal

average pooling and three fully connected layers, followed by a softmax output over four rhythm classes: Junctional Arrhythmia, Sinus Rhythm, Other Rhythm, and Noise. These classes were selected to reflect clinically relevant distinctions and to support the detection of junctional arrhythmias in real-world ICU settings.

Training was conducted using the expert-labelled lead II ECG segments obtained from the previously described methods. The model was optimized using weighted binary cross-entropy loss to address class imbalance. Physiologically realistic augmentations were employed such as powerline interference, electromyography (EMG) noise, and baseline wander; to improve robustness under ICU conditions. Standard normalization and interpolation techniques were applied to insure signal consistency and handle missing data.

The classifier's softmax outputs were used to identify low-confidence predictions for uncertainty sampling, enabling the selection of ambiguous segments near the decision boundary. Additionally, the 256-dimensional embeddings from the final hidden layer were used for embedding-based sampling, allowing the retrieval of morphologically similar but diverse examples from the unlabelled dataset. These capabilities enabled targeted mining of an AtriumDB dataset, improving the efficiency with which experts labelled the rare class. .

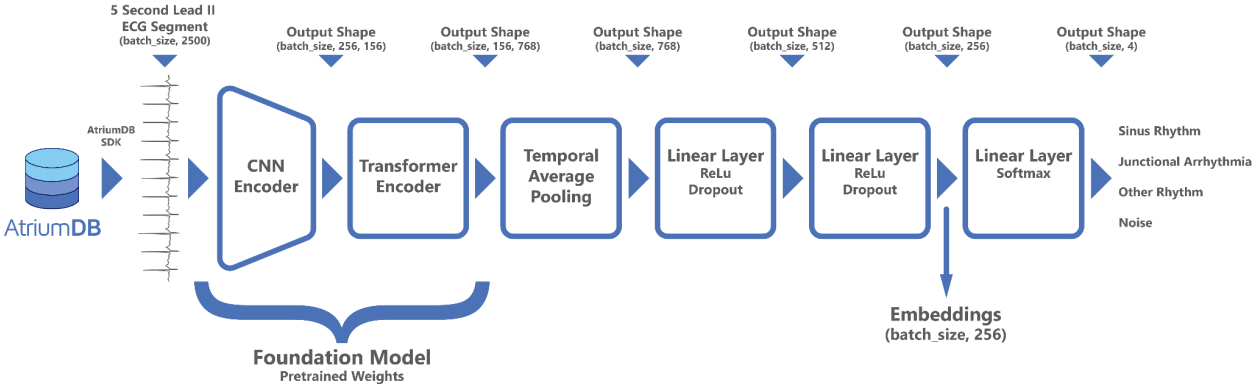


Figure 6. High-level architecture of the arrhythmia classifier.

Sourcing Labels from Uncertainty Sampling

Uncertainty sampling is a widely used technique in active learning, wherein a model is used to identify samples for which its predictions are least confident, those that lie near the decision boundary. These cases are often the most informative for training, as they expose the model's limitations and guide learning toward harder examples. In our context, the combination of labelled data from MUSE-derived 12-lead studies and clinician-reported ICU observations provided enough coverage to train a preliminary arrhythmia classifier. While limited in performance, this initial model was sufficient to enable active mining of the unlabelled ECG data dated back to 2016.

We applied the preliminary model to the full Lead II ECG dataset and selected ECG segments based on model uncertainty. For the initial round, we extracted 5-minute segments from 25 unique patients where the model's softmax score for the Junctional Arrhythmia class fell between 40% and 65%, a range indicative of ambiguous, borderline predictions. For later rounds we used 2-minute windows after the longer windows proved longer than needed. We also extracted 2.5 minutes from five minutes before the end of each patient's stay (not right at the end to avoid noise from the leads coming off) to try to get an example of sinus rhythm. This is keeping with our strategy of showing the model both an arrhythmia and sinus from the same patient so it can learn subtle morphological differences. Both ECG segments for each patient were then pushed to Label Studio for expert labelling. This process was repeated six times with between 20-50 tasks each time depending on available clinician time. Each round expanded the number of sampled patients, enriching the training dataset with ECG morphologies the model found most difficult to classify. As the model improved across successive iterations, the segments retrieved through uncertainty sampling shifted in character. Increasingly, the uncertain windows contained low signal-to-noise ratios or atypical morphologies that were difficult to classify, even for human experts. This inflection point marked the diminishing returns of uncertainty-based sampling.

We transitioned to embedding-based retrieval once approximately half of the returned uncertainty-sampled windows were being labelled as 'unknown.' Since segments labelled as ‘unknown’ are excluded from the training set, reaching this point meant that the majority of clinician labeling effort was no longer producing usable training data. The 50% level was adopted as an operational heuristic during the labeling process rather than as a pre-specified threshold; the precise value is not critical to the methodology, as a similar threshold in this regime would have prompted the same transition decision

Sourcing Labels from Embeddings

In contrast to uncertainty sampling, embedding-based sampling is a form of representation-aware active learning that leverages the model’s internal structure to surface underrepresented but high-confidence samples. Rather than querying the decision boundary, embedding search targets the interior of a known class, retrieving morphologically similar rhythms from new patients. We applied this strategy once the classifier had learned a stable representation of the key rhythm classes, using it to surface additional junctional morphologies from new patients. Here, our goal was to further enrich the junctional arrhythmia class by identifying previously unseen morphologic variants across the population, while reducing annotation burden and avoiding redundant examples. Figure 7 represents a high-level overview of the embedding search technique.

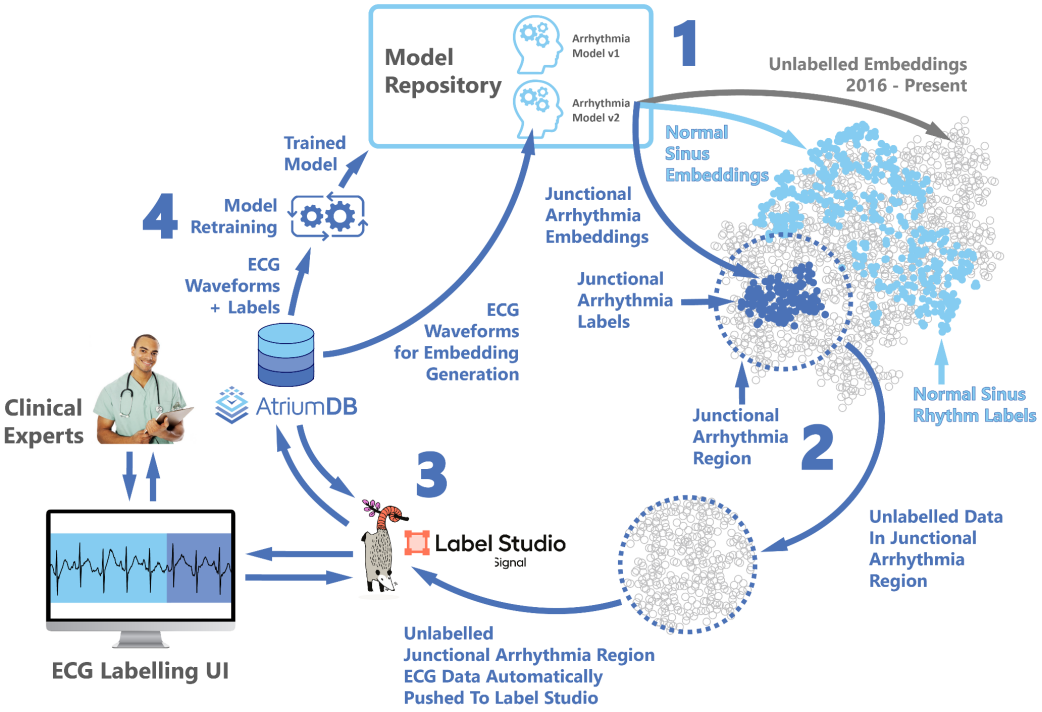


Figure 7. Embedding-based active learning pipeline for junctional arrhythmia enrichment. (1) A trained arrhythmia model generates 256-dimensional embeddings for all unlabelled ECG segments (2016-present), projected here via UMAP alongside labelled normal sinus rhythm (light blue) and junctional arrhythmia (dark blue) embeddings. (2) A hypersphere search region is defined around the global junctional arrhythmia centroid in embedding space, identifying unlabelled candidate segments (grey circles) that fall within the inclusion hypersphere. (3) These candidate segments are automatically exported to Label Studio, where clinical experts annotate them through a dedicated ECG labelling interface. (4) Newly labelled junctional arrhythmia examples are stored in AtriumDB and used to retrain the model, which is versioned in the model repository, closing the active learning loop.

To achieve this, we applied our arrhythmia classifier to the entire dataset and extracted 256-dimensional embeddings from the final hidden layer of the network for every 5-second ECG segment. Using the set of previously labelled junctional arrhythmia segments, we computed a global centroid in this high-dimensional embedding space, as well as per-patient centroids for each labelled junctional arrhythmia patient. We then defined a hypersphere centered at the global centroid, with a radius equal to the mean plus one standard deviation of the distances between each patient's centroid and the global center. This hypersphere captured the typical distribution of junctional arrhythmia morphology as represented by the model and defined our search region for new candidate segments. More information on the centroid calculation and inclusion criteria are available in Appendix 2.

Within this embedding-defined region, we searched for 2-minute segments of ECG that exhibited low "embedding variability", that is, whose 24 constituent 5-second windows were tightly clustered in embedding space. This constraint was found to increase the likelihood that a given 2-minute segment contained only one class, thereby reducing the cognitive load on annotators and increasing labelling throughput. For each selected patient, we also extracted an additional 1-minute segment from the end of their encounter to serve as a likely sinus baseline, preserving an intra-patient contrast between arrhythmia and sinus rhythm for each patient.

As shown in Figure 8, the UMAP projection of the model's embedding space illustrates the structure of the learned representation. This visualization confirms that junctional arrhythmias occupy a coherent and distinct region in embedding space, separable from other classes. The right plot focuses on that junctional region, showing a subset of previously labelled junctional arrhythmia examples alongside unlabelled segments sampled from within the search hypersphere. These are the candidate samples identified for expert annotation via Label Studio. Six representative ECG strips from six unique patients are also shown in Figure 5, highlighting the substantial inter-patient variability in junctional arrhythmia morphology.

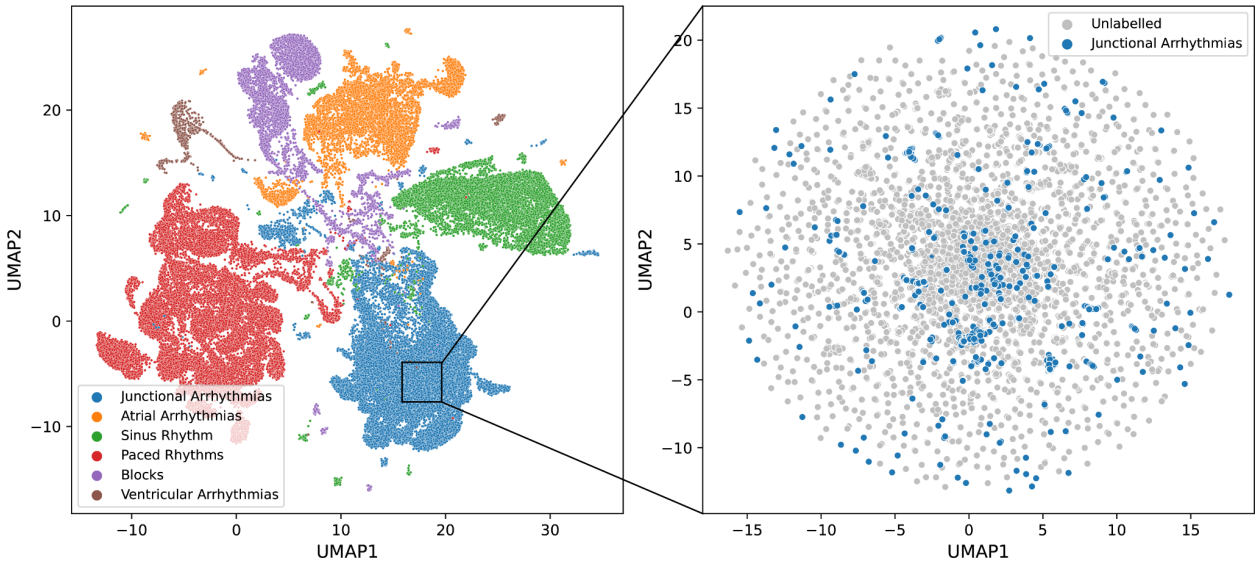


Figure 8. UMAP[13] projection of the model's embedding space. The left subplot displays the full set of labelled ECG segments colored by their parent rhythm class. The right subplot displays the labelled (blue dots) vs. unlabelled data (grey dots) from within the junctional cluster.

Results

Labelled Dataset

The final labelled dataset comprises 181.2 hours of expert-annotated lead II ECG from 1,447 unique pediatric patients across four rhythm classes (Junctional Arrhythmia, Sinus Rhythm, Other Rhythm, Noise). We evaluate the pipeline against a single question: how much more efficiently does it surface the rare class (junctional arrhythmia) than random review of the same dataset?

To establish the background rate, we labelled 200 one-minute segments drawn at random from different patients. Only 2 of 200 (1.0%; 95% CI 0.2-3.2%) contained junctional arrhythmia, confirming its extreme rarity under naive sampling. Taken as a whole, the pipeline produced a dataset that was 15.3% junctional by labelled hours and 24.7% junctional by patient, an enrichment of roughly 15-fold and 25-fold, respectively, over the 1.0% baseline (Fisher exact test on patient counts, $P<.001$). These two views capture different gains: labelled hours reflect the volume of rare-class signal collected, while unique patients reflect the morphological diversity of that signal, which is the more consequential property for building a dataset that generalizes. This conclusion is robust to uncertainty in the baseline: even against the upper bound of its confidence interval (3.2%), the pipeline remained enriched several-fold.

Because the four stages are sequentially dependent, each producing the labels that make the next possible, their individual yields are best read as a progression rather than a competition. Efficiency compounded as the pipeline matured: the share of labelled hours confirmed as junctional rose from 7.6% (MUSE) to 14.2% (ICU observations) to 24.3% (uncertainty sampling) to 35.7% (embedding search), with the patient-level share rising from 9.8% to 52.5% across the same sequence. The two retrospective and observation-based stages were the least efficient per hour of clinician time, but were a necessary foundation. They supplied the seed labels required to train the classifier on which the two higher-yield, model-guided stages depend. The value of the pipeline lies in this cascade, in which the early investment unlocks the later efficiency, rather than in any single stage in isolation.

For the embedding-guided retrievals specifically targeting the junctional region, 60.2% of sampled segments were confirmed as junctional once labelled, with the remainder sinus (18.7%), other rhythms (17.7%), and noise (3.4%). This 60.2% describes retrievals aimed at the junctional cluster; it exceeds the 35.7% hours-level figure in Table 1 because the latter also includes the per-patient sinus baseline segments we deliberately co-sample to preserve intra-patient contrast. The non-junctional retrievals are not wasted effort; they lie near the junctional region in embedding space and represent the morphologically adjacent cases most useful for downstream discrimination.

We emphasize that these results characterize the labeling pipeline, not the classifier. All metrics reflect how efficiently expert attention was directed toward the rare class; we make no claim regarding the diagnostic performance of the model used to guide sampling, which will be the subject of a subsequent publication.

Table 1. Label yield by sourcing strategy. For each method, the junctional columns report the number and percent of that method's own labelled hours and patients confirmed as junctional. For example, with embedding search, 214 patients out of the 408 total discovered through that method ended up being junctional (52.5%). Enrichment is each method's junctional share divided by the 1% random-sampling baseline, shown separately for hours and patients. The random baseline is shown for comparison and is excluded from the totals. The Total row reflects the full pipeline; its enrichment is lower than that of the later stages because it is a volume-weighted blend that includes the high-volume, lower-yield seed stages (MUSE and ICU observations) required to bootstrap the model-guided stages. Note hours and patients do not sum down the column; segments and patients surfaced by more than one method are counted once in the total.

	All Labels		Junctional Arrhythmia Labels			
Labeling Method	**Hours**	**Total Patients**	**Hours (% of total per method)**	**Patients (% of total per method)**	**Enrichment, hours (vs 1% baseline)**	**Enrichment, patients (vs 1%)**
Random Sampling Baseline	3.33	200	0.03 (1%)	2 (1%)	1x	1x
MUSE	75.72	991	5.75 (7.6%)	97 (9.8%)	7.6x	9.8x
ICU Observations	79.21	47	11.23 (14.2%)	18 (38.3%)	14.2x	38.3x
Uncertainty Sampling	12.91	96	3.14 (24.3%)	37 (38.5%)	24.3x	38.5x
Embedding Search	21.36	408	7.63 (35.7%)	214 (52.5%)	35.7x	52.5x
Total	181.21	1447	27.77	358	15.3x	24.7x

To ensure demographic and physiologic diversity, the dataset was stratified by both sex and age group, capturing a representative distribution across the pediatric population from neonates to late adolescents. As shown in Table 2, the dataset includes contributions from both male and female patients, while Table 3 further details the label distribution across age groups.

Table 2. Distribution of labelled ECG hours and patient counts across rhythm classes, stratified by sex. Patients may appear under more than one class; the total is deduplicated.

			Male		Female	
	Hours	**Patients**	**Hours**	**Patients**	**Hours**	**Patients**
Junctional Arrhythmia	27.77	358	14.86 (53.5%)	203	12.91 (46.5%)	155
Sinus Rhythm	88.64	1080	54.57 (61.6%)	594	34.07 (38.4%)	486
Other Rhythm	46.17	315	25.84 (56%)	176	20.33 (44%)	139
Noise	18.63	647	12.21 (65.5%)	360	6.42 (34.5%)	287
Total	181.21	1447	107.48 (59.3%)	797	73.73 (40.7%)	650

Table 3. Distribution of labelled ECG hours and patient counts across rhythm classes, stratified by pediatric age group. Note patients may appear under more than one age group in the case of multiple admissions or having multiple classes.

	Neonate (0-28 Days)		Infant (29 Days-1 Year)		Toddler (1-3 Years)		Child (4-12 Years)		Adolescent (13-18 Years)	
	Hours	**Patients**	**Hours**	**Patients**	**Hours**	**Patients**	**Hours**	**Patients**	**Hours**	**Patients**
Junctional Arrhythmia	3.12 (11.1%)	51	13.43 (47.9%)	144	1.87 (6.7%)	31	8.02 (28.6%)	92	1.51 (5.4%)	42

Sinus Rhythm	18.59 (20.8%)	205	31.00 (34.7%)	318	8.96 (10.0%)	113	19.19 (21.5%)	313	10.90 (12.2%)	170
Other Rhythm	15.05 (32.7%)	78	22.16 (48.2%)	116	2.22 (4.8%)	37	5.72 (12.4%)	72	0.74 (1.6%)	20
Noise	3.63 (19.3%)	168	4.29 (22.9%)	189	5.99 (31.9%)	64	3.48 (18.5%)	157	1.36 (7.2%)	92

Table 4. Comparison of labeling time requirements by label acquisition strategy. The results are presented using two key time constructs: *Annotation time* and *Task duration*. *Annotation time* refers to the annotator's active labeling time per task. Outliers in annotation time were removed because they typically occur when an annotator steps away and leaves a task open, which would otherwise inflate time estimates and misrepresent efficiency. *Task duration* represents the length of ECG content presented in each task, while *annotation time per minute of task* normalizes effort to the amount of ECG reviewed. Junctional label minutes per annotator hour refers to how many minutes of junctional labels you get for each hour of annotator time spent. Task duration varied within uncertainty sampling, which began with 5-minute windows and moved to 2-minute windows; annotation time per minute of task for that stage is computed using its mean duration of 234 seconds. ECG minutes reflect content presented for review, whereas Table 1 reports labelled span, which is smaller because not all of a window receives a label and unknown segments are excluded from the dataset.

Labeling method	**Tasks analyzed**	**Task duration (s)**	**Median annotation time (s)**	**Annotation time per minute of task (s/min)**	**Tasks per annotator hour**	**ECG minutes per annotator hour**	**Total annotator hours**	**Junctional label minutes per annotator hour**
Random Sampling Baseline	200	60	18.5	18.5	92	92	2.17	0.9
MUSE	1799	120	54	27.0	37	74	48.43	7.1
ICU observations	595	300	33	6.6	33	163	18.29	36.8
Uncertainty Sampling	210	120-300	65	16.7	32	124	6.59	28.6
Embedding Search	550	120	41	20.5	58	115	9.53	48.0

Discussion

Principal Findings

Developing clinically deployable machine learning models for pediatric arrhythmia detection requires overcoming the labeling bottleneck and addressing the difficulty of enriching rare classes within large-scale physiologic datasets. This study demonstrates that a multi-strategy labeling pipeline combining retrospective data mining, clinician-in-the-loop annotation, and active learning can accelerate dataset creation while preserving clinical fidelity.

We began with two initial strategies: extracting labels from MUSE 12-lead electrocardiogram studies and real-time ICU observations. While MUSE provided broad coverage it was inefficient; it yielded predominantly sinus rhythm data and required extensive manual verification. ICU observations, although limited in patient count, produced high-yield arrhythmia segments with minimal workflow disruption. With ICU observations we took one hour of arrhythmia data from each patient and 20 minutes of possible normal sinus rhythm. This was an arbitrary amount of time we selected in the beginning however we have now changed that to 20 minutes of arrhythmia and 10 minutes of sinus. labellers spent 58% of the total time labeling on MUSE tasks with only 7.6% of the labels containing junctional and only achieving an efficiency of 7.1 junctional label minutes per annotator hour making it our lowest yield method of finding arrhythmias. ICU observations almost doubled our yield of junctional labels taking us from 7.6% to 14.2%. If repeated, we would prioritize ICU observations from the outset, as they offered the most clinically relevant data with the least annotation overhead helping us get to model-guided annotation faster.

The MUSE and ICU observations provided sufficient labelled data to train a preliminary classifier that could generate embeddings for active learning. This model enabled uncertainty sampling and embedding-based retrieval, which dramatically improved label efficiency and inter-patient variability. Embedding-guided sampling, in particular, enriched rare arrhythmia classes such as junctional arrhythmias while reducing annotation burden. These findings highlight the importance of representation-aware approaches in pediatric machine learning development, where physiologic variability across age groups and clinical contexts limits the utility of larger adult datasets. By leveraging the scale of the physiological waveform dataset accumulated at SickKids and integrating expert feedback loops, we created one of the largest labelled pediatric ECG corpora to date, enabling robust model training under real-world ICU conditions.

Taken as a whole, the pipeline enriched the junctional class roughly 15-fold by labelled hours and 25-fold by unique patients over random sampling (Fisher exact test, $P<.001$). The patient-level gain is the more consequential for dataset quality. Embedding-guided retrieval alone contributed 214 previously unseen junctional patients, substantially widening the range of junctional morphologies represented. Because the stages are sequentially dependent, this enrichment is a property of the full cascade rather than of any one method; the rising per-stage yield reflects efficiency compounding as the pipeline matured, with the low-yield retrospective and observation stages functioning as the necessary seed for the model-guided ones. Each strategy offered distinct advantages and trade-offs. MUSE and ICU observations required the greatest total annotator time, at approximately 48.4 hours and 18.3 hours, respectively (Table 4). In contrast, uncertainty sampling and embedding-guided selection were far more efficient, producing the most diverse set of patients and enriching the minority class most efficiently requiring 6.6 hours with a yield of 24.3% junctional and 9.5 hours with a yield of 35.7%, respectively. These figures underscore that most labeling effort occurs before reaching model-guided techniques with the percentage of labels that end up being the underrepresented class increasing with each successive strategy. Overall, approximately 80 hours of clinician time was utilized in dataset creation, which is relatively modest given the scale and clinical impact of this use case.

However, clinician feedback emphasized that concentration declines after about an hour of continuous labeling, and we observed a corresponding drop in label quality. To mitigate fatigue, it was more effective to provide smaller batches (20 minutes to one hour of labeling) rather than assigning hundreds of tasks at once, improving engagement and consistency.

When normalized by ECG content, ICU observations and uncertainty sampling were the most efficient, reflecting their 5-minute task durations. In contrast, MUSE data and embedding-guided selection, which used 2-minute windows, were slower to complete on a per-minute basis. Median lead times and tasks per annotator hour were broadly similar across methods, but longer windows provided richer context, making annotations easier. Based on these findings, we recommend task durations in the 2-5 minute range to balance efficiency with annotation quality.

Limitations

A limitation of this work concerns the random-sampling baseline used to benchmark enrichment. Because junctional arrhythmia is rare, our 200-segment random sample contained only two junctional cases, which establishes that the class occurs on the order of 1% under naive sampling but does not pin down its exact background rate; the 95% confidence interval on this estimate spans roughly 0.2% to 3.2%. As a result, the enrichment factors reported here should be read as approximate rather than precise. We note, however, that our central conclusion does not depend on the exact base rate: even under the most conservative reading of the baseline, the upper bound of its confidence interval, the full pipeline and each model-guided stage remain substantially enriched relative to chance, and the difference remains statistically significant. Estimating a sub-1% base rate with high precision would require expert labeling of many thousands of randomly drawn segments, an effort that is impractical under real-world clinical time constraints and is precisely the bottleneck this pipeline is designed to address.

Conclusion

This study addresses a key barrier to machine learning adoption in pediatric critical care by introducing a scalable, clinically informed labeling framework. Our approach accelerates data curation and establishes a generalizable methodology for domains where expert annotation is time-intensive and rare events dominate. Future work will focus on expanding label diversity and validating model performance in silent trials prior to bedside deployment. Ultimately, the combination of clinician-in-the-loop and model-guided active learning represents a transformative step toward real-time, clinician-augmented machine learning systems capable of improving patient safety and outcomes.

Acknowledgments

Generative AI was used in the preparation of this manuscript. Claude (Anthropic) Opus 4.8 was used solely for language editing of author written text (proofreading, correcting grammar and spelling, and improving clarity and readability). It was not used to generate content, to conceive or design the study, to acquire, process, or analyze data, to interpret results, or to identify references. All AI-assisted edits were reviewed, edited and approved by the authors, who are fully accountable for the accuracy and integrity of the manuscript in its entirety.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR) through a Project Grant (Application Number: 480831), The University of Toronto Data Science Institute Catalyst Grant, and The University of Toronto EMH Seed Grant. The article processing fee was funded by the CHIR grant. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Data Availability

Due to patient privacy and institutional data governance policies, the ECG waveform data used in this study cannot be publicly released. However, access to de-identified subsets of the dataset may be made available to academic researchers upon request and subject to ethics approval and data use agreements with The Hospital for Sick Children (SickKids). The code used to train the arrhythmia classifier and its weights are not publicly available. Qualified academic collaborators may request access to specific components for non-commercial research under a data use agreement. Code for AtriumDB is open source and is available on PyPi as an installable package. Source code is available on GitHub at atriumdb.io. Label Studio is open source and available at labelstud.io.

Authors' Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by SV and WD. ECG labeling was done by AB, AA, MK, D Ehrmann, D Eytan and MM. The first draft of the manuscript was written by SV. The manuscript was commented on and edited by SDG, MK, AG, WD, AS, AA, D Ehrmann and RG. All authors read and approved the final manuscript.

Abbreviations

ECG: electrocardiogram
ICU: intensive care unit
JET: junctional ectopic tachycardia
ML: machine learning
UMAP: Uniform Manifold Approximation and Projection

References

1. Mayaya P-C, Tinică G, Chistol RO, Moraru L, Damian SI, Frăsinariu OE, Furnica C. Incidence and Risk Factors for Early Postoperative Arrhythmias in Congenital Heart Disease – Systematic Review. J Cardiovasc Emergencies Târgu Mureș, Poland: De Gruyter Brill Sp. z o.o., Paradigm Publishing Services; 2025;11(2):43–53. doi: 10.2478/jce-2025-0005

2. Pantelidis P, Bampa M, Oikonomou E, Papapetrou P. Machine learning models for automated interpretation of 12-lead electrocardiographic signals: a narrative review of techniques, challenges, achievements and clinical relevance. J Med Artif Intell AME Publishing Company; 2023 May 30;6(0). doi: 10.21037/jmai-22-94

3. Skippen PW, Sanatani S, Gow RM, Froese N. Diagnosis of postoperative arrhythmias following paediatric cardiac surgery. Anaesth Intensive Care 2009 Sept;37(5):705–719. PMID:19775033

4. Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng Med Biol Mag 2001 May;20(3):45–50. doi: 10.1109/51.932724

5. Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. Sci Data Nature Publishing Group; 2020 May 25;7(1):154. doi: 10.1038/s41597-020-0495-6

6. Reyna MA, Sadr N, Alday EAP, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A, Clifford GD. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. Brno, Czech Republic; 2021. p. 1–4. doi: 10.23919/CinC53138.2021.9662687

7. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LH, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data Nature Publishing Group; 2023 Jan 3;10(1):1. doi: 10.1038/s41597-022-01899-x

8. Oh J, Chung H, Kwon J, Hong D, Choi E. Lead-agnostic Self-supervised Learning for Local and Global Representations of Electrocardiogram. arXiv; 2022. doi: 10.48550/arXiv.2203.06889

9. Kiyasseh D, Zhu T, Clifton DA. CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients. arXiv; 2021. doi: 10.48550/arXiv.2005.13249

10. McKeen K, Masood S, Toma A, Rubin B, Wang B. ECG-FM: An Open Electrocardiogram Foundation Model. arXiv; 2025. doi: 10.48550/arXiv.2408.05178

11. Shani C, Zarecki J, Shahaf D. The Lean Data Scientist: Recent Advances Toward Overcoming the Data Bottleneck. Commun ACM 2023 Jan 20;66(2):92–102. doi: 10.1145/3551635

12. Goodwin AJ, Eytan D, Greer RW, Mazwi M, Thommandram A, Goodfellow SD, Assadi A, Jegatheeswaran A, Laussen PC. A practical approach to storage and retrieval of high-frequency physiological signals. Physiol Meas IOP Publishing; 2020 Apr;41(3):035008. doi: 10.1088/1361-6579/ab7cb5

13. McInnes L, Healy J, Saul N, Grossberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 2018 Sept;3:861. doi: 10.21105/joss.00861

14. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Adv Neural Inf Process Syst Curran Associates, Inc.; 2020. p. 12449–12460. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

15. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. Proc 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Vol 1 Long Short Pap Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. doi: 10.18653/v1/N19-1423

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MultimediaAppendix1InterRaterAgreementStudy.docx
- MultimediaAppendix2EmbeddingBasedSearchCentroidDefinitionsandInclusionCriteria.docx