Comparing Radar and Force-Sensor Method agreement for Non-Invasive Respiratory Monitoring of Freely-Moving Mice

Master's thesis in Biomedical Engineering

Axel Andersson
Academic supervisor: Frida Sandberg
Company: TrackPaw Scientific AB

Department of Biomedical Engineering
Faculty of Engineering, LTH
June 2026

Abstract

Continuous non-invasive monitoring of respiratory rate in laboratory mice can support research across drug efficacy, sleep, and stress and pain assessment. Established respiratory rate estimation methods exist but many conflict with the 3Rs principle through invasiveness, motion restriction, or animal stress. A cross-method comparison between two non-invasive systems on the same freely-moving animal offers a verification path that does not require an invasive reference. The goal was to conduct such a comparison, analyzing agreement between methods and when methods become unreliable dependent on behavioral state.

A signal-processing pipeline was developed for a 60 GHz pulse coherent radar (Acconeer A121) mounted above a home cage, extracting respiratory rate from chest-wall-displacement phase information. Twelve pipeline iterations were each accepted against a non-regression gate on a synthetic-signal benchmark.

Resulting respiration rate estimates from the final pipeline were compared against an existing force-sensor respiratory rate metric (TrackPaw) on simultaneous recordings spanning four cages and multiple 24-hour long sessions, with neither system treated as ground truth. Agreement was characterized by Bland-Altman analysis with a repeated-measures formulation. Behavioral-state robustness was characterized by per-method coverage and by agreement between the radar flagging motion and TrackPaw's flagging travel.

After 30 s pre-pairing aggregation, the joint bias between the methods was +0.03 Hz with 95 % limits of agreement of −0.14 and +0.20 Hz. Neither method reliably produced estimates during active behavior, when radar and TrackPaw both flag motion, and most coverage from both methods happened when no motion was flagged.

Continuous non-invasive respiratory rate monitoring in the home cage appears feasible at rest. Reliable estimation during active behavior remains an open problem for both methods.

Popular science summary

Watching mice breathe using floors and invisible energy

How can a mouse's breathing be measured without touching, holding, or even watching the animal? Two very different sensors did exactly that, in parallel, and agreed on the answer.

Breathing tells researchers a lot. A drug that quiets pain may slow breathing. A poor night's sleep can speed it up. In studies of asthma, anxiety, sleep, or heart disease, the breathing rate of a laboratory mouse can be one of the most informative numbers in the experiment. The trouble is collecting that number without disturbing the animal's regular behavior, impacting what the measurement really tells us.

Each established method comes with a price. Sealed chambers force the animal to sit still and can stress the mouse into changing the very breathing being measured. Implanted sensors require surgery and can leave lasting effects on the animal. Even watching with a stopwatch in hand burns human hours and adds an observer that mice can sense and react to.

This thesis tested two methods that quietly read breathing while the mouse roams freely in a home cage. The first is a small radar mounted above the cage that sends short pulses of high-frequency radio waves through the lid and listens for the faint echo from the chest wall rising and falling. The second is a platform of force sensors under the cage floor that picks up the tiny force shifts between an in-breath and an out-breath pushing and pulling on the floor. Neither sensor ever touches the animal, and the mouse is free to walk, climb, eat, and sleep.

There is a twist. Neither device can be called the truth. So the question changes from whether one device gets the right answer to whether two sensors, built on completely different physical principles and knowing nothing about each other, will agree. Such an agreement would be hard to explain unless both happen to be reading the same underlying signal.

Across simultaneous recordings from four cages and many hours of mouse rest, sleep, and ordinary activity, the two methods landed within roughly 10 breaths per minute of each other when measuring once every 30 seconds. That might sound like a lot until you realize mice typically breathe near 200 breaths per minute! That is a tight match for entirely independent technologies. When the mouse moved, however, both methods went blind in similar ways. The radar lost the breathing rhythm in the clutter of small body movements, and the force-sensor platform could not separate breathing when a mouse was running around.

Quietly counting breaths around the clock may help squeeze more usable data out of every animal study, lowering the number of mice needed to answer the same scientific question. That fits the principle of 3Rs, the call to replace, reduce, and refine the use of laboratory animals. The same kind of unobtrusive monitoring may, in time, support long-term welfare checks in many other animal facilities. The mouse, meanwhile, never knows the device is there.

Acknowledgments

This degree project has been conducted at the Department of Biomedical Engineering, Faculty of Engineering, Lund University.

First of all I would like to thank my supervisors Frida Sandberg and Erik Månsson, and my examiner Martin Stridh for letting me start this project and giving me the opportunity to learn more about radar signal processing.

Thank you Hooi Min and the whole team at Truly Labs for allowing me to run the radar study with you. It has been a pleasure working with you every step of the way. You always kept me informed of what was going on in the lab and put together the data I needed from the facility. Thanks for the coffee and office tour as well!

Thank you Frida for expressing your vast knowledge in a way that challenged me but never felt judgmental or harsh. I was often inspired to try and do better after our meetings.

Thank you Erik! Your enormous competence in the field of radar has most likely made me underestimate how difficult this project could have been. More than anything, thank you for always asking "what's next?" I think I thrive when good enough never feels like the goal and the question of how we could do better is the only thing we try to reach for. With that said, perhaps this is not the magnum opus. Still, maybe you think this is better than just good enough.

Thank you to friends and family and to all our mice friends who decided to show themselves! You guys truly (pun intended) put in the hours.

Author’s declaration of generative AI assistance

Generative AI assistance was used during the preparation of this report. The use was limited to correcting grammar and spelling, iterating on the visual design and rendering of the result figures, locating candidate literature for review, maintaining the bibliography file, and refactoring software. Generative AI was not used to write the thesis prose or to select signal-processing and statistical analysis methods. All scientific claims, design choices, and conclusions in this report remain the author’s responsibility.

Contents

List of Figures

List of Tables

Abbreviations

PCR	Pulsed Coherent Radar
RR	Respiratory Rate
STFT	Short-Time Fourier Transform
PSD	Power Spectral Density
FFT	Fast Fourier Transform
SNR	Signal-to-Noise Ratio
CFAR	Constant False Alarm Rate
HPBW	Half-Power Beam Width
HWAAS	Hardware-Accelerated Average Samples
LoA	Limits of Agreement (Bland-Altman)
BA	Bland-Altman
IQR	Interquartile Range
PPV	Positive Predictive Value (precision)
NPV	Negative Predictive Value
FPR	False Positive Rate
RFID	Radio-Frequency Identification
USB	Universal Serial Bus
RPi5	Raspberry Pi 5
IaC	Infrastructure as Code
LTH	Lunds Tekniska Högskola (Faculty of Engineering, Lund University)
3Rs	Replacement, Reduction, Refinement

1 Introduction

In this chapter, an introduction to the reasons for, and aims of, this thesis are described.

1.1 Motivation

In animal research, the rate at which an animal breathes, or respiratory rate (RR), is informative across many fields of research, including olfactory neuroscience, social behavior, learning and memory, and respiratory physiology [1]. In any scenario where measuring RR in mice is required, removing all confounders from the measurement is desirable. A method to monitor breathing continuously, non-invasively, and without motion restriction is therefore key.

Modern methods of measuring RR in mice all carry some limitation [1]. Plethysmography, measuring volume change, requires an attached mouthpiece or full body chamber impacting regular movement patterns. Harnessed accelerometers restrict movement or can be removed by the mice. Implanted sensors require surgery and can affect baseline biological traits in the animal. Observer methods (counting breaths manually) are labor-intensive and cannot be done in longer sessions. Even the presence of an observer could be a confounder for mouse behavior (e.g. via stress).

Advantages of a fully autonomous, non-interacting measurement device are continuous measurements at any time of day with no impact on the animal behavior. In addition, such a measurement device complements the principle of 3Rs [2]. The device is a refinement method with no inherent pain and discomfort for the animal. If the RR can be more continuously monitored, each test could offer more valuable knowledge and in a shorter period of time, possibly reducing the study size.

One method of monitoring respiration non-invasively (and non-interactively) is offered by TrackPaw and their force-sensor platform. The system is constructed as a typical lab cage for mice where the base of the cage contains multiple force sensors that are able to measure minimal changes in force, e.g. the force difference of inspiration and expiration in mice. To build confidence in a technology in line with the 3Rs, simultaneous RR measurement from another non-interactive system serves as a step toward verification without re-introducing methods that cause stress or add other confounders.

Radar is a promising choice for allowing RR monitoring without any interaction with the animal. By studying the displacement of the chest wall on a mouse, the radar could feasibly detect respiration as previously done in rodents [3]. The physical principle behind radar measurement is also fairly different from that of force sensors. Agreement between the methods of measurement would indicate a common phenomenon being measured where few other common phenomena intuitively exist.

1.2 Aims and objectives

The primary aim of this thesis is to develop a signal-processing pipeline that extracts RR from an Acconeer A121 Pulse coherent radar (PCR) on non-anesthetized freely-moving adult mice, and characterize agreement against TrackPaw's existing force-sensor RR metric using Bland-Altman analysis. Neither system is treated as ground truth.

The secondary aim is to characterize how reliability in RR extraction varies between behavioral states, namely motion against stationarity, derived from TrackPaw travel metric compared with radar motion readings.

There are two main questions this thesis will answer.

(RQ1) How well do the radar and force-sensor algorithms agree in detecting respiratory rate in freely-behaving laboratory mice, quantified by Bland-Altman agreement analysis?

(RQ2) Under which behavioral states does each method become unreliable?

RQ1 is not a pass or fail. No numerical agreement threshold will be considered irrefutable evidence of true respiratory rate being measured. RQ2 will divide behavioral state into motion or stationarity.

1.3 Thesis outline

The background (Section 2) develops the theory needed to interpret radar samples and the evaluation frameworks used in the comparison. The methods (Section 3) describe the recording setup, the signal-processing pipeline, the force-sensor platform interface, and the statistical analysis. The results (Section 4) address RQ1 and RQ2 in turn, preceded by internal validation of the radar pipeline. The discussion (Section 5) interprets the agreement and coverage results, considers confounds and limitations, and closes with methodological lessons. The conclusion (Section 6) summarizes the answers to RQ1 and RQ2.

2 Background

The working principles of the Acconeer A121 pulsed coherent radar are described first, including how chest-wall displacement maps to a per-distance phase signal carrying respiratory information, the configurable parameters, the timing model, and the antenna beam pattern. Signal-processing tools applied to the phase signal follow, namely the signal-to-noise definition, Butterworth bandpass filtering, the short-time Fourier transform, Welch's method, and noise-floor normalization. The TrackPaw force-sensor platform is described next at principle level, covering both the respiratory rate estimator and the travel estimator. The chapter closes with the evaluation frameworks used in the comparison, Bland-Altman agreement analysis with the repeated-measures formulation, and the confusion matrix.

2.1 Pulsed coherent radar (Acconeer A121)

By transmitting pulses of electromagnetic waves with precise timing, a PCR samples the backscatter (energy reflected back) at specific distances from the radar. At each distance d_n, the sampled value y_n can be expressed as in Equation 1 [4].

$$y_n = A_n \cdot e^{i4\pi \frac{d_n - d}{\lambda}} \tag{1}$$

Where A_n is an amplitude proportional to the amount of backscatter at d_n. The phase, the argument of the exponential, can be interpreted as an angular displacement between d_n and the actual distance d to the surface or object which caused the backscatter. Recall that $e^{i2\pi k}$ is an imaginary number traversing the unit circle N times as k increases from 0 to N, wrapping the angle around from 2π to 0 at every integer position. For clarity, the phase from Equation 1 can be rewritten as in Equation 2,

$$2\pi \frac{d_n - d}{\frac{\lambda}{2}} \tag{2}$$

where λ is the wavelength of the electromagnetic wave. At 60 GHz, the frequency of the Acconeer A121 PCR, $\lambda \approx 5$mm and the angular displacement thus wraps around at every ~ 2.5mm displacement between d_n and d. Rephrased, the phase cycle is approximately equal to a 2.5mm displacement between the measurement bin center d_n and the actual distance d at which the backscatter came from.

As a note on A_n, as previously stated it is proportional to the amount of backscatter at d_n, which is maximized when $d_n = d$.

2.1.1 Signal demodulation

Breathing in a mouse causes a small periodic displacement of the chest wall. By Equation 2, this displacement translates into a change in the phase of y_n while leaving A_n approximately unchanged. As long as the displacement is small compared with the wrap interval of 2.5 mm, the phase oscillates without wrapping at the RR. The RR is therefore present as a frequency in the per-frame time series of the phase of y_n at the distance bin closest to the chest wall [3].

The radar also captures large reflections from objects in the cage that do not move, such as the cage walls, the water bottle, and the lid. These produce contributions to y_n that have near-constant phase over the timescale of a breathing cycle, since the corresponding distances do not change. Subtracting the temporal mean of y_n per distance bin removes them.

2.1.2 Configuration parameters

The A121 offers a wide selection of configurable parameters that control where and how the radar transmit and receive pulses. First is the selection of distance bins at which to sample, i.e. what number and values of d_n. Multiple bins are able to run as part of the radar configuration, giving information about the amplitude and phase at each distance. By selecting a start point, step length and number of points, the radar will set up measurements at the corresponding distances. The start point and step length are expressed in units of the half wavelength distance of 2.5mm, meaning that a start point of 10 and step length of 2 corresponds to the first distance bin being placed at $10 \cdot 2.5 = 25$mm from the radar and all subsequent bins are at $2 \cdot 2.5 = 5$mm offsets from this start point. Rephrased, the A121 can configure a number of distances to sample at, and sample one value per distance as in Equation 1.

Another configurable parameter is pulse-length, called profiles in the case of the A121. Five profiles are available, each with a fixed pulse length, and the choice trades pulse energy against minimum usable range and resolution [5], [6]. With a longer pulse more energy is transmitted in each pulse, which can increase the amount of backscatter and thus the SNR. The minimum usable range is affected because the A121 has the transmit and receive antennas close together on the same chip. While a pulse is being transmitted some of the pulse energy can travel directly from one antenna to the other. This short-path signal, called direct leakage, can dominate reflections from objects close to the radar. A longer pulse means the leakage is present for longer, and the minimum range at which a reading can be clear of leakage increases.

A third configurable parameter sets how many times each measurement at each distance bin is repeated and averaged on-chip. This is hardware accelerated averaging of samples, or HWAAS [5], [7]. Averaging reduces random noise, and doubling HWAAS adds about 3 dB of SNR. The trade-off is time, since the duration to produce samples grows linearly with HWAAS.

2.1.3 Timing

The duration to produce samples is part of a wider timing hierarchy. Each measurement at one distance bin takes a duration τ_{point} that includes the HWAAS samples and some overhead. A pass through all configured distance bins is a sweep, with duration

τ_{sweep} equal to the sum of τ_{point} across the bins. A frame is, configurably, one or more sweeps grouped together, with duration τ_{frame}. The frame rate is the inverse of the frame duration,

$$f_{\text{max}} = \frac{1}{\tau_{\text{frame}}}, \tag{3}$$

and is the highest rate at which complete frames of data can leave the radar [8].

The frame rate can be set explicitly or left unset. With an explicit rate, the A121 paces frames against the on-chip 24 MHz crystal oscillator [5], producing a stable interval between frames. Without an explicit rate, the radar produces frames as fast as the connected host can request them, which ties the actual frame timing to the host's latency and scheduling rather than the radar's own clock [8].

Between frames the A121 enters a configurable idle state. READY is the idle state ready to take the next frame immediately, which can sustain a higher frame rate by minimizing sweeping period [8].

2.1.4 Antenna beam pattern

An antenna does not have a strict field of view, but it has a region of higher sensitivity. The A121′s antennas are most sensitive along a forward direction, and the sensitivity falls off as the angle from that direction increases. The standard way to characterize this falloff is the angular width over which the power stays within half of the peak, called half-power beam width or HPBW [5]. The HPBW is therefore not a hard cutoff but a reference contour, and reflections from outside it are attenuated rather than excluded, so objects beyond the HPBW limits can still be detected with less signal strength.

The A121′s HPBW is not symmetric. The beam is wider in one plane, called the H-plane, than in the perpendicular E-plane, with approximate widths of 65° and 53° respectively [5]. The angular sensitivity is therefore oblong rather than circular. When the targets to be detected are not arranged symmetrically around the radar, this gives a clear choice for how to orient the antenna. Aligning the wider H-plane along the longer arrangement of targets covers more of them with stronger signal.

2.2 Signal processing

2.2.1 Signal-to-noise ratio (SNR)

In any measurement, the power (or strength) of the signal compared to the power of the noise is important to consider. This is the signal-to-noise ratio, or SNR. In this thesis the signal is the small periodic part of the radar reflection caused by chest-wall motion during breathing, and the noise is everything else, such as random fluctuations in the radar electronics, reflections from the cage walls or bedding that do not move with the breath, and occasional larger movement from the mouse itself.

The signal power is the power spectral density (PSD) [9] at the breathing frequencies in a recording, and the noise power is the PSD at the same frequencies in a recording made of the same cage but with no mouse. An SNR of 1 means the breathing peak is no taller than the empty-cage background and an SNR of 10 means it is ten times taller.

2.2.2 Butterworth bandpass

A Butterworth bandpass is a filter whose passband is as flat as possible, so frequencies inside the band pass through approximately unchanged in magnitude. A side-effect of a Butterworth filter is that it introduces a frequency-dependent time delay, which can shift fast events in the signal forward or backward in time relative to where they occurred.

This time-delay distortion is removed by applying the filter twice, once forward in time and then once backward in time on the result [10]. The second pass cancels the delay introduced by the first, leaving a filtered signal that is time-aligned with the input. This zero-phase property preserves the timing of events in the signal.

2.2.3 Short-time Fourier transform (STFT)

The short-time Fourier transform (STFT) recovers time-varying frequency content by chopping the recording into overlapping segments and taking a Fourier transform on each segment separately. The result is a two-dimensional grid, with one axis frequency and the other time. The value at each grid point is the spectral content of the segment centered at that time. The discrete STFT, in the form used by scipy [11], is given in Equation 4.

$$S[q,p] = \sum_{k=0}^{N-1} x[k] \cdot \overline{w[k-ph]} \cdot e^{-i2\pi q \frac{k}{N}} \tag{4}$$

Where $x[k]$ is the input signal, $w[m]$ (overline denoting the conjugate) is a window function of length M samples that smoothly tapers each segment to zero at the window edges, h is the hop size in samples that sets how far the window slides between consecutive segments, and N is the FFT length in samples. The frequency index q corresponds to a step of $\frac{1}{NT}$ Hz, where T is the sampling interval, and the time index p corresponds to a step of hT seconds. The overline denotes complex conjugation. Rephrased, $S[q,p]$ is the Fourier transform at frequency index q of the windowed segment centered at time index p.

Three parameters together control the behavior of the STFT. The window length M sets the frequency resolution, where a longer window can distinguish more closely-spaced frequencies but blurs out fast changes in the signal. The hop size h sets how often a new spectrum is produced. A smaller hop gives finer time resolution at the cost of more overlap between segments. The FFT length N, when chosen larger than M, pads the windowed segment with zeros, placing the spectrum on a denser frequency grid without changing the underlying frequency resolution.

2.2.4 Welch's method

Welch's method [9] is a related spectral estimator that builds on the same windowed Fourier transforms as the STFT. Within an analysis window the signal is divided into K overlapping sub-segments, a PSD is computed from each sub-segment, and the K PSDs are averaged together into a single PSD. The averaging reduces the variance of the resulting PSD by approximately a factor of K compared with the PSD of any one sub-segment.

The STFT is the special case of K = 1, where each PSD is kept individually without averaging.

2.2.5 Noise-floor normalization

The PSD of a recording can vary with frequency and with distance bin even without a mouse present, due to fixed contributions from the radar electronics and from static reflectors in the cage. These fixed contributions can be measured by recording the same cage with no mouse and computing the PSD of that empty-cage recording. Normalizing the recording's PSD against the empty-cage PSD removes the fixed contributions and expresses the result in SNR units. A value of 1 means the recording has the same power as the empty cage at the same frequency and distance bin, and a value greater than 1 means more.

Without this rescaling, raw PSDs can slope upward around fixed frequencies that also appear in the empty-cage recording and are therefore decisively not a present respiratory signal. This can bias later peak-detection toward those frequencies. After rescaling, any remaining frequency peaks most likely indicate real signals.

This normalization is conceptually similar to a detection method in radar where the threshold is set relative to a local noise estimate rather than to a fixed value, called constant false alarm rate, or CFAR [4].

2.3 Force-sensor platform

The TrackPaw platform replaces the cage floor of a typical mouse cage with a 3 by 5 grid of 15 cells, each standing on three force-sensors that record the load on the cell over time. Two outputs from these readings are used in this thesis for comparison against the radar. The first is RR estimates produced at intervals near 250 ms during normal operation, with larger gaps when no valid estimate is available. The second is travel, the total distance moved by a mouse over each 60 s window, reported per identified animal.

The platform also reads RFID chips implanted in the mice to identify which animal is moving.

Preprocessing of the incoming force-sensor readings produces a 1035 ms time delay, applied as an offset to the RR series before pairing against the radar.

2.3.1 RR estimator

The RR estimator runs independently on each cell. The per-cell force-sensor signal is bandpass filtered between 1 Hz and 10 Hz. The output signal is tracked as a respiratory amplitude from the force difference traced by inspiration and expiration of a mouse standing on the cell.

A positive zero-crossing of the filtered signal is detected. The signal must first dip below a negative threshold and then rise through a positive threshold of equal magnitude for a positive crossing to count. Each detected crossing produces a period (the time since the previous positive crossing) and a peak amplitude (the largest magnitude observed during that period).

The most recent 50 periods are retained in a circular buffer. Once the buffer is full, the mean and standard deviation of the 50 period durations are computed, and the normalized standard deviation is the standard deviation divided by the mean. A rate is reported when the normalized standard deviation is below 0.25 and the implied rate lies between 60 and 600 breaths per minute, equivalently between 1 Hz and 10 Hz.

The per-cell output carries the rate, the mean and standard deviation of the period amplitudes, and a detected flag. The cage-level RR paired against the radar is the median of the per-cell rates across cells reporting a valid detection.

2.3.2 Travel estimator

The travel estimator works from a position estimate of the mouse on the cage floor, updated at each force-sensor sample. Two anchor points are kept, the most recent committed position and the one before it. A new position contributes to the accumulated distance only when separated by more than 25 mm from both anchors, requiring two consecutive moves of that magnitude before any distance is credited.

Accumulated distance is committed every 10 seconds while the mouse is being tracked, and once more when tracking ends. The committed distance is attributed to the RFID identity associated with the mouse, or to a separate unattributed bucket when identity is not yet resolved. The output for each cage is a sum of millimeters per minute per RFID identity, at a native 60 s bucket resolution.

All sessions in this study use single-housed mice, so the RFID-attributed and unattributed buckets refer to the same animal. The per-minute travel value paired against the radar is the sum across both buckets.

2.4 Evaluation of systems

To answer the research questions, methods for agreement and behavioral state analysis are required.

2.4.1 Bland-Altman agreement analysis

Bland-Altman analysis characterizes the agreement between two measurement techniques [12]. For each pair of simultaneous measurements $x_{1,i}$ and $x_{2,i}$, the difference $d_i = x_{1,i} - x_{2,i}$ and the mean $m_i = \frac{x_{1,i}+x_{2,i}}{2}$ are computed, and a Bland-Altman plot displays d_i against m_i. The mean of the differences gives the bias, and the 95 % limits of agreement (LoA) are computed as in Equation 5.

$$\text{LoA} = \text{bias} \pm 1.96 \cdot \text{SD}(d) \tag{5}$$

Where $\text{SD}(d)$ is the standard deviation of the differences.

The procedure above assumes the differences are independent samples of a single population. When the same subject is measured repeatedly across multiple sessions, this assumption is broken in two ways. Within a session, time-adjacent paired measurements are correlated when the RR estimates are comprised of near time averaging of raw measurements. Across sessions, the bias of the differences may change with day-to-day differences in animal posture or cage layout, and pooling all pairs into a single sample mixes those between-session changes into the within-session variation.

The repeated-measures procedure of Bland and Altman 1999 addresses the latter issues by treating each session as a separate measurement unit [13]. With S sessions in the study and n_s paired measurements in session s, the per-session mean $\overline{d}_s$ and standard deviation s_s of the differences are computed within each session. The variance of the differences is then decomposed into two components. The within-session component σ^2_{dw} is a weighted average of the per-session variances s_s^2 in which sessions with more paired measurements contribute more weight, capturing the typical spread of differences inside a single session. The between-session interaction component σ^2_{dI} is estimated from the spread of the per-session means $\overline{d}_s$ around the joint mean, capturing how much the bias differs between sessions. The joint bias is the mean of all paired differences across sessions, computed equivalently from the per-session means as in Equation 6.

$$\overline{d} = \frac{\sum_{s=1}^{S} n_s \overline{d}_s}{N}, \quad N = \sum_{s=1}^{S} n_s \tag{6}$$

The joint limits of agreement combine both variance components, computed as in Equation 7. The interaction component is set to zero when between-session variation does not exceed within-session variation.

$$\text{LoA}_{\text{joint}} = \overline{d} \pm 1.96\sqrt{\sigma^2_{\text{dI}} + \sigma^2_{\text{dw}}} \tag{7}$$

The first issue of dependent time-adjacent paired measurements can be combated by aggregation, averaging adjacent samples to a single value in larger bouts of time before pairing.

2.4.2 Confusion matrix

When two signals provide binary labels of the same underlying state, Bland-Altman analysis does not directly apply because the difference between two binary values is not a meaningful agreement quantity. A confusion matrix is used instead, with one side designated the reference label and the other the predicted label, and the four counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) tabulated.

From these counts, the performance metrics used in this thesis follow as in Equation 8.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{8}$$

NPV is the probability that the reference is negative given that the prediction is negative, the read-out for whether the predictor's negative state coincides with the reference's negative state. PPV is the positive equivalent.

3 Methods

The comparison reported in this thesis was performed by simultaneously recording the radar and the TrackPaw force-sensor platform from a study on single-housed adult mice held one per cage at Truly Labs, a contract research facility in Lund. Each session (one cage holding one mouse for one period of time) ran approximately 24 hours, with up to four cages recorded in parallel.

The physical cage and radar mount are described first, followed by the radar configuration with parameter rationale. The recording infrastructure is described next, covering per-cage worker processes, hourly-segmented storage, liveness monitoring, and integrity validation on ingest. A corner-reflector test that validated the configuration follows. The study protocol is then described, covering animal handling, housing conditions, and the trial-versus-main-study split. The signal-processing pipeline is described end-to-end at the final version, with cumulative changes across versions summarized in a table. A synthetic radar signal model used to benchmark the pipeline against a ground truth is described after, together with the non-regression acceptance criterion applied to every version and parameter change. The TrackPaw platform interface makes mention of a bug found in the respiratory algorithm, with pre and post fix results reported for Bland-Altman comparison. The statistical analysis closes the chapter, covering the Bland-Altman pairing protocol, the motion-overlap confusion matrix between systems, and per-method coverage.

3.1 Cage and hardware

The cages used in the study were the TrackPaw custom cage with the force-sensor platform and polycarbonate walls, a dimensional copy of a Tecniplast 1291H Type III H polycarbonate cage. A Tecniplast 1291 Filter Top lid covered the cage.

The radar was mounted on top of the lid via a 3D-printed baseplate that raised the radar above the lid surface, a height tuned experimentally for maximum amplitude of the reflection from the cage floor. A 3D-printed shield enclosed the radar assembly to protect the PCB from incidental contact and to cover the bright LEDs on the radar that could otherwise disturb the mice. The assembly was attached to the lid with adhesive. The 3D models for the baseplate and shield are listed in Appendix C.

The total distance from the radar to the cage floor was approximately 219 mm, computed from the baseplate height (4 mm), the lid height above the cage-body lip (40 mm), and the internal cage height (175 mm). A mouse could be at this distance lying on the floor, closer to the radar when standing on bedding, or at a longer slant range near the cage edges where the radar path passed through more of the cage interior.

The radar was mounted with the antenna pointing straight down at the cage floor and rotated so that the wider H-plane (65°) was aligned with the longer axis of the cage interior [5]. At the radar-to-floor distance of 219 mm this orientation placed the HPBW footprint of the antenna over most of the cage floor (Section 2.1.4).

The full physical cage setup is shown in Figure 1.

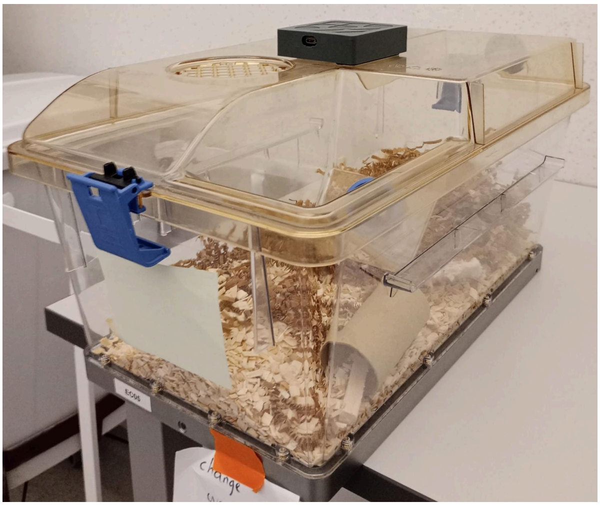

Figure 1: Physical setup of the cage with the radar mounted on the filter top lid.

The radar hardware in each cage consisted of an evaluation board (XE121) carrying the A121 PCR sensor and the integrated antenna, and a connector board (XC120) carrying a microcontroller that bridged the sensor to a USB HS 2.0 connection to the host.

All radar units were connected over USB to a single centralized Raspberry Pi 5 (RPi5), which could host up to four units simultaneously.

The four cages are labeled A, B, C, and D throughout this thesis. The same letter identifies the cage and the radar unit mounted on it. A topological view of the deployed hardware can be seen in Figure 2.

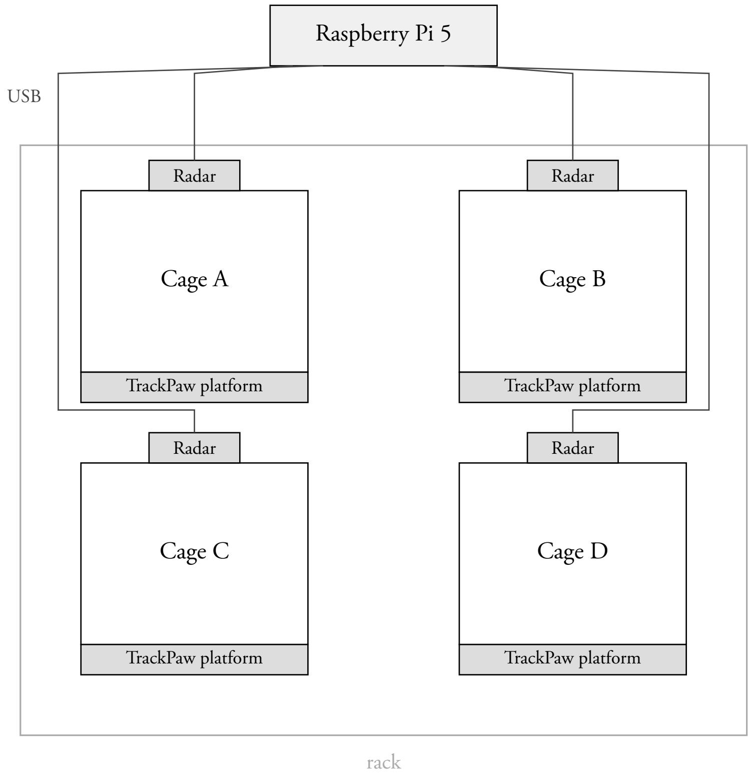

Figure 2: Hardware topology of the radar + TrackPaw deployment.

3.2 Radar configuration

The frame rate of the radar (Section 2.1.3) was set to 250 Hz, well above the Nyquist rate for the breathing band ($\leq$ 12 Hz [1]).

The start point and step length (Section 2.1.2) were chosen as integer multipliers of the radar's 2.5 mm minimum step. A start point of 60 placed the near bin at 0.15 m, and a step length of 8 gave 20 mm between consecutive bins. Ten distance bins were configured, placing the far bin at 0.33 m. The 0.15 to 0.33 m range covers the radar-to-floor path of 0.219 m with margin for slant range to the bottom corners of the cage interior.

Profile 2 (Section 2.1.2) was chosen as the highest-SNR profile whose minimum usable range remains below the near bin at 0.15 m.

HWAAS (Section 2.1.2) was set to 284, just below the maximum compatible with the 250 Hz frame rate at Profile 2 per the timing model in Equation 3. A higher HWAAS averages out more random noise per sample, so the largest value within the available timing budget was preferred, with a small margin to avoid jitter in the frame interval when nearing the maximum.

Both idle states were set to READY (Section 2.1.3) to sustain the 250 Hz frame rate. The remaining configuration parameters were set per recommendation of Acconeer for slow movements [14]. One sweep was configured per frame, with an automatic sweep rate, and both continuous sweep mode and double buffering disabled.

The full configuration parameter list is in Appendix A.

3.3 Recording and ingest

The recording system ran on the RPi5. The full host configuration was version-controlled, with the entire stack from bare hardware to running state re-deployable from the configuration repository.

One worker process per cage connected over USB to the cage's XC120, configured the sensor with the parameters in Appendix A, and streamed radar frames to disk as hourly-rotated segments. Hourly segmentation bounded the worst-case data loss to the current partial segment.

A worker was flagged as failed on abnormal exit, including crashes, uncaught exceptions, and USB I/O errors, or on a frame-delivery stall beyond a timeout. Failed workers respawned automatically. Sustained failure, defined as repeated crashes within a short window, escalated to a USB power-cycle of the affected device. The escalation recovered transient hardware faults that respawn alone could not.

A per-device liveness signal was exposed by the recording host and read by a remote dashboard. The dashboard provided real-time confirmation throughout each session, ensuring that silent failures could not run unnoticed for long.

Completed segments were periodically copied from the recording host to an analysis workstation. Each transferred file was validated by opening it through the same code path used by the analysis pipeline. Files that failed to open or were structurally incomplete were rejected as malformed, logged as corrupted, and excluded from further sync and analysis. Corrupted files therefore never entered the pipeline.

Hourly segmentation, prompt retrieval, and integrity validation together enabled a checkpoint at the end of the trial study (2026-03-25, one cage, 24 h). The data retrieved during the trial was analyzed offline to confirm that the radar configuration produced usable signal before committing to the main study. The full session log is in Appendix B.

3.4 Configuration validation

A trihedral corner reflector served as a known target for validating the radar's response to small displacements. The reflector was built from three copper-tape surfaces meeting at right angles. This geometry reflects radar waves back along the direction they came from, giving a strong and orientation-insensitive return. The 3D model is listed in Appendix C.

In the first test, the reflector was held outside the cage and tilted by a small amount. The phase of the radar sample at the range bin matching the reflector's distance was observed live for changes corresponding to the tilt. A sub-wavelength tilt produced a clear shift in the phase, confirming the displacement-to-phase relationship (Equation 2, Section 2.1.1). The test was qualitative, with the tilt amount not measured and the live signal inspected on screen rather than recorded.

In the second test, the reflector was held at the corner of the cage outside the cage wall, where the path from the radar passed through the wall and the reflector sat beyond the HPBW footprint (Section 2.1.4). A small tilt at this position still produced a detectable phase response, confirming that the radar's effective coverage extended to the corners of the cage.

3.5 Study protocol

The study was conducted at Truly Labs (Medicon Village, Lund) under a pre-approved ethical permit (number: 05517/2025) and Swedish animal welfare legislation. All animal handling was performed by trained Truly Labs personnel.

Animals were laboratory mice with body weight greater than 20 g. The lower weight bound was set to keep chest-wall displacement and reflection area large enough for reliable radar detection. Strain, sex, and age were not constrained by the protocol and not analyzed as covariates. Per-session details are listed in Appendix B. Each cage housed exactly one mouse per session, since the radar could not distinguish multiple animals in a shared cage. Each animal was acclimatized in the study cage for at least five days before recording, per Truly Labs protocol.

Cages were kept at constant temperature with a 12 h light-dark cycle. Animals had free access to food, tap water, and cage enrichment. Food was placed directly in the bedding rather than in a metal hopper, since a hopper above the cage floor would have produced a strong static reflection in the radar beam. Bedding placement removed the reflector entirely. The water bottle remained at the standard lid position, appearing as a persistent static reflector that the DC-removal step suppressed (Section 2.1.1). Bedding was standard lab bedding, loose enough for the animal to dig and burrow.

Data collection consisted of a trial study and a main study. The trial study ran on 2026-03-25 with one cage for 24 h, confirming that the radar configuration produced usable signal under study conditions before committing to the main study. All main study sessions are listed in Appendix B.

3.6 Signal processing pipeline

The pipeline operated on recorded segments, extracting the samples per bin and time and produced a single RR time series per segment. Development proceeded as a sequence of versions, each the successor to the previous. Per-version performance metrics for all twelve are reported in Results. The final version, v12, is the radar method used in all other analyses. Every numerical parameter cited below is listed in Appendix A.

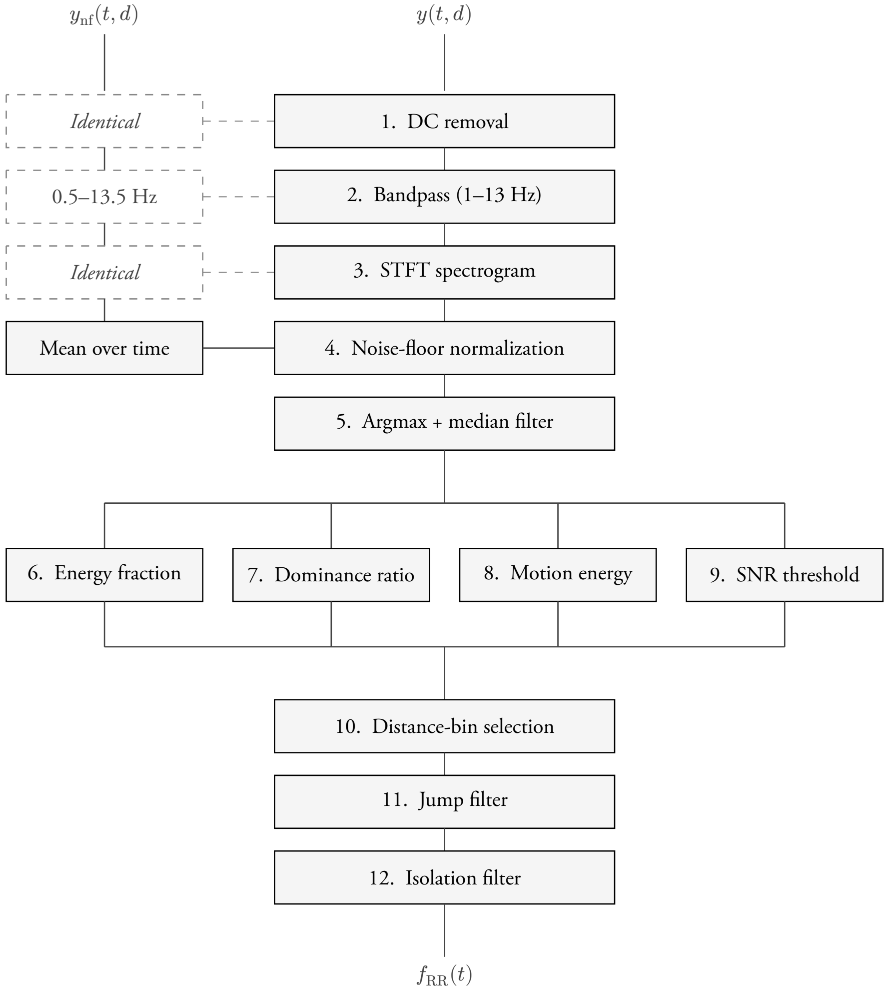

Figure 3: Block diagram of the v12 radar signal-processing pipeline. Empty-cage recording $y_{\mathrm{nf}}(t, d)$ passes the same preprocessing pipeline as the analyzed recording $y(t, d)$.

Time-domain conditioning. The first two steps conditioned the per-bin time series before any frequency analysis.

1. **DC removal** (Section 2.1.1). The temporal mean of each distance bin was subtracted to suppress all static reflections, e.g. the cage walls, the water bottle, and the filter top.
2. **Butterworth bandpass** (Section 2.2.2). A 5th-order Butterworth bandpass was applied forward and backward to the real and imaginary parts of each distance bin (zero-phase filtering [10]). The filter passband (1–13 Hz) covered mouse RR from rest (2–4 Hz) [1] up through sniffing. The filter was set with 0.5 Hz margins (0.5–13.5 Hz) when filtering noise floor. This was done so the analysis-band edges were divided by a less attenuated noise floor-band later, avoiding near zero divisions from uneven attenuation between the two. A bandpass rather than a highpass was chosen so that

spurious high-frequency content was actively suppressed. The 5th order was settled empirically.

Spectral analysis and normalization. Each conditioned per-bin signal was then mapped to frequency over time in an STFT and rescaled relative to a noise-floor reference.

3. **STFT spectrogram** (Section 2.2.3). An STFT [11] was applied per bin with a Hann window [15] of 2048 samples (8.2 s at 250 Hz, frequency resolution $\Delta f \approx 0.12$ Hz), a hop length of one-eighth of the window, and a 4× zero-padded FFT. Only the positive half of the spectrum was retained.
4. **Noise-floor normalization** (Section 2.2.5). Each spectrogram was divided by a per-(distance, frequency) reference PSD. For real recordings, the reference was computed from an empty-cage recording with the same radar configuration and no animal in the cage, collected once per cage at deployment, with the pipeline auto-selecting the reference by source cage at analysis time. References were not refreshed between sessions, so cage-furniture shifts during a study (bedding redistribution, water-bottle reseating, filter-top reseating, cage replacement between studies) could change the static-clutter pattern and the per-(distance, frequency) reference PSD. The division placed all peak detection in SNR space, normalizing out any frequency-dependent structure in the raw noise floor so it could not bias the dominant-frequency selection. The widened bandpass in step 2 kept the band- edge roll-off out of this division.

Per-bin estimation and quality masks. The next steps assigned a candidate dominant frequency per distance bin and per time frame, then applied independent masks, joined later, to mark estimates that were unreliable.

5. **Argmax and median filter.** Per bin and time frame, the dominant frequency, f_1, was selected as the argmax over the analysis band of the normalized spectrogram. A small median filter along the time axis smoothed single-frame excursions [16].
6. **Energy-fraction mask.** Per bin and time frame, energy was summed in main-lobe wide ($4\Delta f$) bins around the dominant frequency f_1 and the harmonics $2f_1$ and $3f_1$. Estimates were masked when this sum fell below 30% of the total in-band energy.
7. **Sole-frequency dominance mask.** Local maxima of the normalized spectrum were extracted per bin and time frame, with a minimum separation of one main-lobe width [17]. Estimates were masked when the second-highest peak exceeded 30% of the highest. Well-isolated fundamentals therefore passed. Spectra with two comparable peaks did not.
8. **Two-component motion-energy mask.** A motion-energy score was computed per time frame as the total log-energy of the normalized spectrogram summed over distance and frequency. Time frames whose score exceeded the 85th percentile across the segment were flagged as global motion. In addition, the peak-to-peak range of the score in a ± 5-frame neighborhood was computed, and frames whose local range

exceeded the 50th percentile of all local ranges were flagged as motion spikes. The combined mask was dilated by one frame on each side to absorb boundary frames.

9. **SNR mask** (Section 2.2.1). Estimates whose normalized spectrum at the dominant frequency did not clear $2.0 \times$ the noise floor were masked.

Cross-time finalization. The remaining steps collapsed the (bin, time) representation to a single RR time series and filtered the resulting series for outliers.

10. **Per-time-frame bin selection.** Per time frame, of the distance bins that passed all earlier masks and the median filter, the bin with the highest normalized PSD value at the dominant frequency was selected. This replaced an earlier band-mean-PSD selector that was biased by broadband noise elsewhere in the band.
11. **Jump filter.** Time frames whose selected RR deviated more than 0.2 Hz from the median of a symmetric ± 4-frame neighborhood were masked. The symmetric window prevented near-threshold outliers from corrupting the reference for subsequent frames.
12. **Isolation filter.** Time frames with fewer than two non-masked neighbors within a ± 4-frame window were masked. This removed lone surviving estimates.

The cumulative changes across pipeline versions are summarized in Table 1.

Version	Change
v1	Welch PSD per bin over the full segment, argmax over the analysis band, highest-peak bin selected. Constant RR per segment.
v2	Sliding-window Welch periodogram (one per hop) replaces the segment-wide Welch. One RR per time frame.
v3	STFT spectrogram replaces the sliding Welch. Noise-floor normalization added.
v4	Median filter on the dominant-frequency series along time.
v5	Sole-frequency dominance mask added.
v6	SNR mask, jump filter, and moving-average smoothing added.
v7	Energy-fraction mask replacing the sole-frequency dominance mask.
v8	Two-component motion-energy mask added.
v9	Sole-frequency dominance mask reintroduced. Median-filter window widened. Isolation filter replaces the moving-average smoothing.
v10	Filter passband widened beyond the analysis band so the noise floor at the band edges was unattenuated by the bandpass.
v11	4× zero-padded FFT introduced for finer frequency resolution.
v12	Per-time-frame bin selection switched from band-mean PSD to peak PSD at the dominant frequency.

Table 1: Cumulative changes across pipeline versions v1 to v12. Each row describes the change introduced by that version relative to the previous.

3.7 Benchmarking the pipeline on synthetic signals

To evaluate the processing pipeline, a synthetic radar model was created so each version could be benchmarked against a ground truth.

3.7.1 Synthetic signal model

The synthetic benchmark generated a fixed bank of 50 seeded synthetic signals (250 000 samples each at $f_s = 250$ Hz, 16.7 min per signal) approximating real A121 output under the configured radar parameters. The ten distance bins, indexed $n = 1, \ldots, 10$, had centers d_n uniformly spaced from $d_1 = 150$ mm to $d_{10} = 330$ mm to match the deployment range.

Each per-bin sample was constructed from the polar form Equation 1. The amplitude of the animal contribution at distance bin d and time t was modeled as a Gaussian function of the offset between the bin center and the gross animal position $d_{\text{gross}}(t)$,

$$A(t,d) = \exp\left(-\frac{\left(d_{\text{gross}}(t) - d\right)^2}{w^2}\right), \quad w^2 = 500\text{mm}^2, \tag{9}$$

giving maximum amplitude when the animal sat at the bin center and a smooth decay as the gross position moved away.

The animal range varied with time as the animal moved, decomposing into a slowly varying gross-position component and a small superimposed breathing oscillation,

$$d_{\text{refl}}(t) = d_{\text{gross}}(t) + d_{\text{breath}}(t). \tag{10}$$

Each component was modeled separately.

The gross position changed by discrete decision intervals of 0.5 s. At each interval, a decision was made for the animal to either stand still or move toward or away from the radar at a constant velocity. Movement occurred with probability 0.1, with the direction drawn uniformly and the speed drawn uniformly from 5 to 50 mm/s. The chosen state was held for the full interval before the next draw, with the position clamped so that the breathing oscillation never carried the animal beyond the 150 to 330 mm range.

The breathing displacement was a small sinusoidal oscillation, with fixed amplitude $A_{\text{breath}} = 0.3$ mm, overlaid on the gross position. The instantaneous frequency was held piecewise-constant within 30 s bouts. Each bout's frequency was drawn from a truncated normal centered at 4 Hz with standard deviation 1 Hz, bounded to the [2, 8] Hz range. The instantaneous phase was the time-integral of the frequency, keeping the sinusoid continuous across bout boundaries.

To mimic the body parts that move with breathing alongside the chest, three additional reflectors were added. Each reflector co-moved with the animal at a fixed range offset from the gross position, drawn from a Gaussian with standard deviation 15 mm. The reflector breathed at the same instantaneous breathing frequency as the main animal but with an independently drawn breathing amplitude (uniform between 30% and 80% of A_{breath}) and breathing-phase offset (uniform on $[0, 2\pi)$). The peak reflection amplitude was drawn uniformly between 0.2 and 0.8 (relative to the main animal's peak).

A static contribution modeled persistent reflectors such as the cage walls. The contribution had a constant amplitude C much larger than any other contribution, with a phase drawn independently per bin,

$$y^{\text{static}}(d) = C\exp(i\psi(d)), \quad C = 40, \quad \psi(d) \sim \text{Uniform}[0, 2\pi). \tag{11}$$

The animal contribution carried independent Gaussian noise of standard deviation 0.4 on amplitude and phase before polar combination,

$$y^{\mathrm{anim}}(t,d) = (A(t,d) + \eta_A(t,d)) \exp\left(i \left[4\pi \frac{d_{\mathrm{refl}}(t) - d}{\lambda} + \eta_\varphi(t,d) \right] \right), \quad (12)$$

with $\eta_A, \eta_\varphi \sim \mathcal{N}(0, 0.4^2)$ independent across (t,d). The reflector contributions $y^{\mathrm{refl},r}(t,d)$ were constructed by the same polar form at offset gross position $d_{\mathrm{gross}}(t) + \Delta_r$, $\Delta_r \sim \mathcal{N}(0, 15^2)$, with the per-reflector breathing parameters described above and no amplitude or phase noise. Summing the animal, reflector, and static contributions gave the full per-bin synthetic signal in the notation of the pipeline diagram,

$$y(t,d) = y^{\mathrm{anim}}(t,d) + \sum_{r=1}^{3} y^{\mathrm{refl},r}(t,d) + y^{\mathrm{static}}(d). \quad (13)$$

The model parameters were plausible-range defaults drawn from rodent-breathing literature and visual inspection of recorded radar samples. They were not formally fit. The residual risk that the synthetic regime differs from the real-recording regime is addressed in the discussion.

An empty-cage variant of the synthetic model provided the noise-floor reference for the synthetic case. The variant was generated by removing the animal signal and the additional reflectors from the per-bin signal, leaving only the static contribution and the additive noise. It was formatted to match the empty-cage recording used as the real-data reference, so the signal processing pipeline ran on both synthetic and real noise recordings without modification.

3.7.2 Benchmarking

All version and parameter changes of the pipeline were accepted against the same criterion. Aggregate metrics on the synthetic benchmark had to not regress relative to the predecessor, specifically the RMS (in areas where an estimate is produced) error, the bias, and the false positive rate (FPR) at $\Delta = 1$ Hz had to not increase by more than the seed-to-seed standard deviation observed across the 50 synthetic signals. Hypotheses for changes drew from any source, including analytic reasoning, literature, and cross-inspection of real-recording outputs against the TrackPaw reference. Each candidate change was validated only against the synthetic benchmark, so synthetic non-regression was the sole acceptance criterion. Later iterations, once aggregate metrics stayed consistent with successive versions, were motivated by artifacts surfacing in real-recording inspection that the synthetic generator did not reproduce. v10 to v12 each addressed one such artifact.

For v10, the empty-cage reference was passing through the same bandpass as the analyzed recording, so band-edge roll-off attenuated the reference at the analysis-band edges. Dividing the recording spectrum by an attenuated reference inflated the normalized SNR at those edges and biased peak detection outward. Bandpassing the empty-cage reference with a passband 0.5 Hz wider than the analysis band held the reference flat across the band of interest.

For v11, the unpadded FFT bin spacing of approximately 0.12 Hz set a floor on the dominant-frequency estimate's resolution. Reporting RR in 0.12 Hz steps discarded usable information about peak location within a bin. A 4× zero-padded FFT refined the bin grid and let the dominant peak be located more precisely, which also removed the visible step structure that would otherwise appear in the Bland-Altman comparison against the TrackPaw reference.

For v12, the per-frame bin selector previously ranked bins by band-mean PSD across the analysis band. On real recordings, that selector could track movement-induced broadband elevation more strongly than respiratory dominance, because broadband content lifted the mean. Switching to the peak PSD at the dominant frequency selected the bin where the breathing peak stood out most. The synthetic spectrum was dominated by the breathing peak with little broadband content, so band-mean and peak-PSD selection agreed there and the change was invisible on the benchmark.

3.8 Force-sensor platform

The TrackPaw platform delivered RR and travel time series, used as the reference for the radar comparison. Each was labeled with the source cage, allowing pairing with the corresponding radar unit by cage label. The TrackPaw algorithm and configuration were held constant (one pre and post-fix version) for the full duration of the study, fixing the processing baseline across sessions.

Bland-Altman analysis (Section 2.4.1) of early results showed a systematic bias between the paired measurements. A bug was found in the platform's RR estimator (Section 2.3.1), where the sum of the 50 buffered periods was divided by 49 instead of 50, underestimating the reported rate by approximately 2.04%. The pre-fix and post-fix versions of the estimator were both run against the same recordings.

Radar and TrackPaw recorded against independent timestamps. The 1035 ms RR offset, equal to the platform's preprocessing delay, was subtracted from each TrackPaw RR timestamp before pairing.

3.9 Statistical analysis

Agreement was assessed with a Bland-Altman analysis (Section 2.4.1). Behavioral state robustness was assessed with a confusion matrix (Section 2.4.2) of motion overlap between systems. Coverage, how often each system produced an RR estimate, was also computed.

3.9.1 Bland-Altman agreement analysis

For agreement analysis, radar and TrackPaw RR samples were paired by nearest-neighbor matching within ±3 s, with unmatched samples dropped from either side. The paired differences were used to compute Bland-Altman joint statistics across sessions per Bland-Altman 1999 [13], with the bias as in Equation 6 and the limits of agreement as in

Equation 7. The full Bland-Altman protocol was applied three times against the same recordings: against the pre-fix TrackPaw RR series, against the post-fix series, and against the post-fix series after a pre-pairing aggregation step. The aggregation binned radar and TrackPaw RR samples respectively in 30 second intervals where all contained samples were averaged. If less than half of the maximum possible samples were present in the bin before averaging it was discarded. Aggregation was done to mitigate time-adjacent dependency of data points in both systems. Each plot was overlaid with boxplots in 0.25 Hz bins along the mean axis, drawn for bins containing at least 100 pairs. Each box summarized the differences within the bin by the median, the interquartile range (IQR), and 1.5× IQR whiskers, with outliers omitted. Equal IQRs and whisker spreads across the mean range were the expected outcome only after aggregation, since time-adjacent dependency had been mitigated.

3.9.2 Confusion matrix motion and coverage analysis

For behavioral state robustness analysis, radar and TrackPaw systems required labels for motion. Each 60 s *window* (Section 2.3.2), the native resolution, of TrackPaw travel was reduced to a binary travel label, true (active) when travel was greater than zero and false (stationary) when travel was zero. The > 0 threshold aligned with the platform's 25 mm two-step deadzone, which quantized sub-threshold motion to exactly zero. The radar pipeline motion-energy mask (motion mask) provided the binary label for radar motion detection. Each radar *frame*, at the native radar RR resolution, was marked active when masked and stationary otherwise.

A per-window test characterized motion agreement on windows where the within-window motion mask frames were unambiguous (almost entirely active or stationary). This aggregate radar label was set stationary when the fraction of masked frames was below 10 % and active when the fraction was 90 % or higher. Windows where the fraction of masked frames lay between the thresholds were dropped. The proportion of paired windows surviving the filter was the kept-fraction. On the kept windows a confusion matrix between the aggregate radar label and the travel label was summarized by NPV, the probability that TrackPaw was stationary when the radar was dominantly stationary, and PPV, the probability that TrackPaw was active when the radar was dominantly active, per Equation 8.

Per-frame the motion mask was paired with the travel label of the enclosing window, giving two conditional rates, $P(\text{motion mask} = \text{active} \mid \text{travel} = \text{active})$ and $P(\text{motion mask} = \text{active} \mid \text{travel} = \text{stationary})$. The ratio of the two, the per-frame lift, summarized how much more often the mask fired inside active windows than inside stationary windows.

Both tests were computed per cage and pooled across cages. The per-window test was pooled by concatenating the kept windows from every cage and recomputing the confusion matrix on the union. The per-frame test was pooled by summing radar and TrackPaw labels across cages and recomputing the rates from the sums.

Radar and TrackPaw coverage, the share of all time where each system produced a RR estimate, was computed. A non-motion fraction was computed as the fraction of time when the motion mask or TrackPaw travel label was stationary, respectively. The coverage within non-motion was computed as the fraction of non-motion frames or windows for which the corresponding method delivered a valid RR estimate. Coverage within motion was conversely computed as estimate share within active frames or windows respectively.

4 Results

Internal validation of the radar pipeline opens the chapter, covering the empty-cage noise floor used as the per-cage SNR denominator, the final version pipeline on a representative synthetic signal, the per-version aggregate metrics across the synthetic bank, and the final version pipeline on a representative real recording. The radar-vs-TrackPaw RR Bland-Altman agreement addressing RQ1 follows, reported across three subsections, against the pre-fix RR series, the post-fix RR series, and the post-fix RR series after a 30 s pre-pairing aggregation. A behavioral state robustness section addressing RQ2 closes the chapter, validating the radar's motion mask against TrackPaw travel and reporting per-method coverage across behavioral states.

4.1 Empty-cage noise floor

The empty-cage noise floor for all cages is presented in Figure 4. The PSD is highest within the 1 to 13 Hz analysis band and drops sharply above it, with similar shape across cages and modest variation in absolute level. This is the per-(distance, frequency) reference that the noise-floor normalization step of the pipeline divides by.

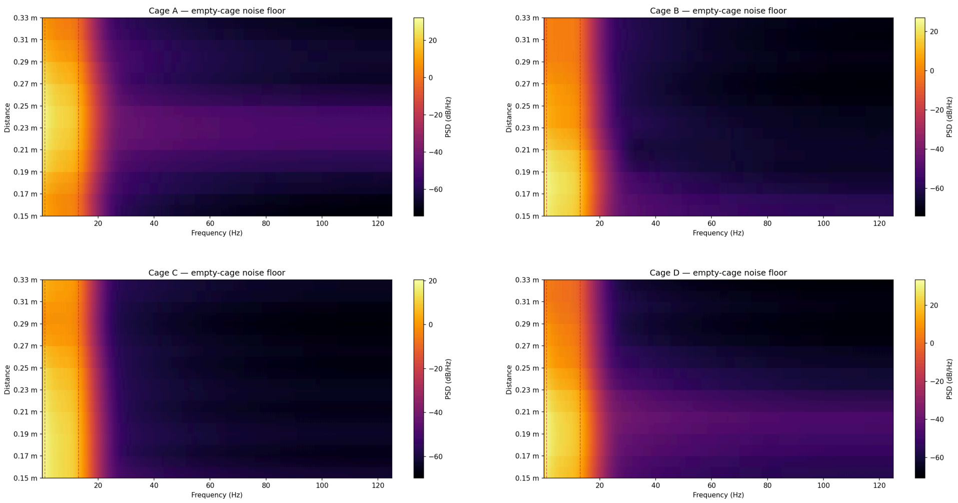

Figure 4: Empty-cage noise-floor PSD per range bin and frequency for each main-study cage. Dashed verticals delimit the analysis band.

4.2 Final pipeline on a synthetic model signal

The final analysis pipeline output on a representative synthetic signal is shown in Figure 5. The top panel overlays the extracted RR on the spectrogram against the ground-truth breathing frequency, with motion-masked regions shaded. The middle panel shows per-frame log-energy with the 85th-percentile global threshold, and the bottom panel shows the local log-energy range over a ±5-frame neighborhood with the 50th-percentile spike threshold. The extracted RR tracks the piecewise-constant ground truth closely across the signal, with motion masking active where either threshold is exceeded.

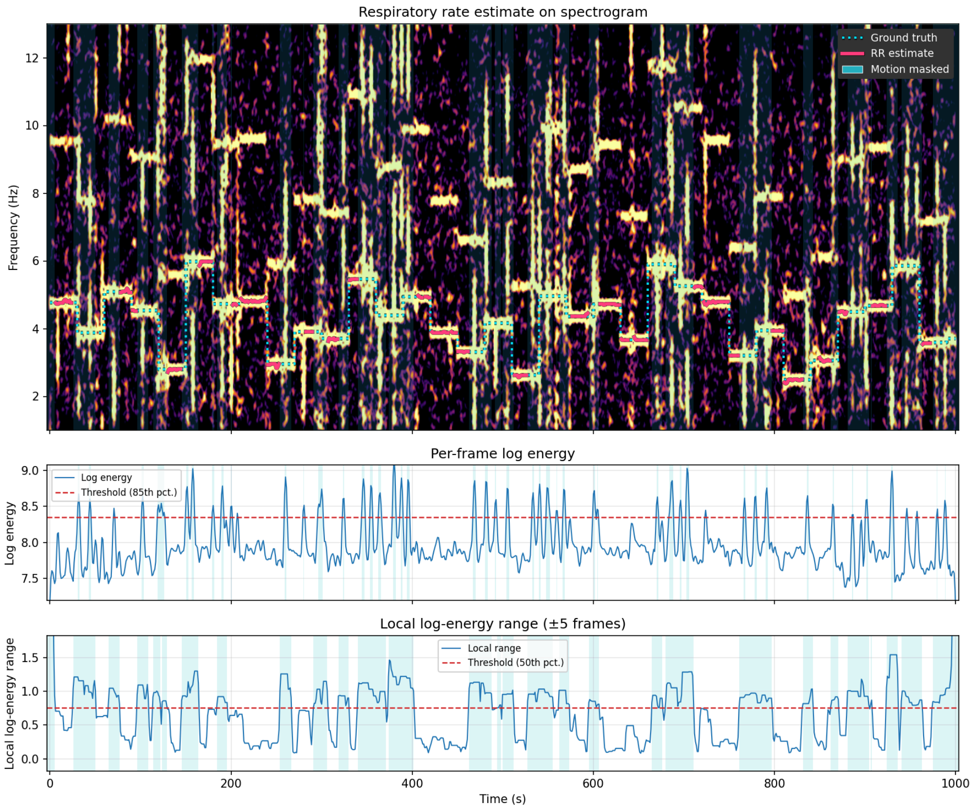

Figure 5: Final analysis pipeline on a representative synthetic radar signal: extracted RR overlaid on the spectrogram and on ground-truth (top). Motion-energy diagnostic (bottom).

4.3 Pipeline evaluation across versions

The per-version aggregate metrics across the 50 synthetic signals are shown in Table 2. The RMS error drops from 1.14 Hz at v1 to 0.05 Hz at final version, the bias from +0.19 Hz to essentially zero, and the false-positive rate at $\Delta = 1$ Hz from 39 % to under 0.1 %. Coverage, the fraction of frames for which the pipeline produces an RR estimate, stays at 100 % through v4, drops to around 90 % from v5 to v7 as quality masks reject unreliable estimates, and falls to about 40 % from v8 onward when the motion-energy mask flags frames with high or rapidly varying energy. Motion coverage, the same fraction restricted to non-motion frames (those not flagged by the motion-energy mask), remains around 86 % once the mask is active. The largest improvements are concentrated in v1 to v9, with later versions producing no change at the table's precision.

Version	Cov.	Mot. cov.	RMS (Hz)	Bias (Hz)	FPR @ Δ=1 Hz
v1	100%	—	1.069 ± 0.203	+0.237 ± 0.45	34.8 ± 11.9%
v2	100%	—	0.564 ± 0.477	+0.058 ± 0.086	3.5 ± 4%
v3	100%	—	0.753 ± 0.609	+0.101 ± 0.126	4.5 ± 4.7%
v4	100%	—	0.705 ± 0.584	+0.091 ± 0.12	4.2 ± 4.5%
v5	88.3%	—	0.224 ± 0.141	−0.001 ± 0.014	0.4 ± 0.3%
v6	84.4%	—	0.147 ± 0.03	+0 ± 0.004	0.3 ± 0.3%
v7	88.1%	—	0.153 ± 0.099	+0.004 ± 0.012	0.2 ± 0.2%
v8	41.1%	86.7%	0.109 ± 0.025	+0 ± 0.005	0.1 ± 0.2%
v9	40.8%	86.2%	0.045 ± 0.004	+0 ± 0.003	0 ± 0%
v10	40.8%	86.2%	0.045 ± 0.004	+0 ± 0.003	0 ± 0%
v11	40.8%	86.2%	0.045 ± 0.004	+0 ± 0.003	0 ± 0%
v12	40.9%	86.2%	0.045 ± 0.004	+0 ± 0.003	0 ± 0%

Table 2: Aggregate metrics per pipeline version, mean ± seed-to-seed standard deviation across 50 seeded synthetic signals. Coverage (Cov.) and motion coverage (Mot. cov.) are reported as means only. Motion coverage is undefined for versions without the motion-energy mask.

4.4 Final pipeline on a representative real recording

The final analysis pipeline output on a representative real recording is shown in Figure 6 (cage B, session B1, hour 5). The panels follow the same template as Figure 5, with the extracted RR overlaid on the spectrogram (top), the per-frame log-energy with the 85th-percentile global threshold (middle), and the local log-energy range over a ±5-frame neighborhood with the 50th-percentile spike threshold (bottom). The breathing-frequency band sits near 3 to 4 Hz across the calm portion of the recording, with the extracted RR tracking it. The second half of the hour shows broadband energy and is mostly motion-masked. Visual inspection also shows segments with clear breathing energy in the spectrogram that the RR estimate does not track, indicating the final pipeline does not pick up all available signal. The few RR estimates that pass the mask in the motion-heavy second half show elevated values, consistent with the breathing-rate increase that accompanies physical exertion such as movement.

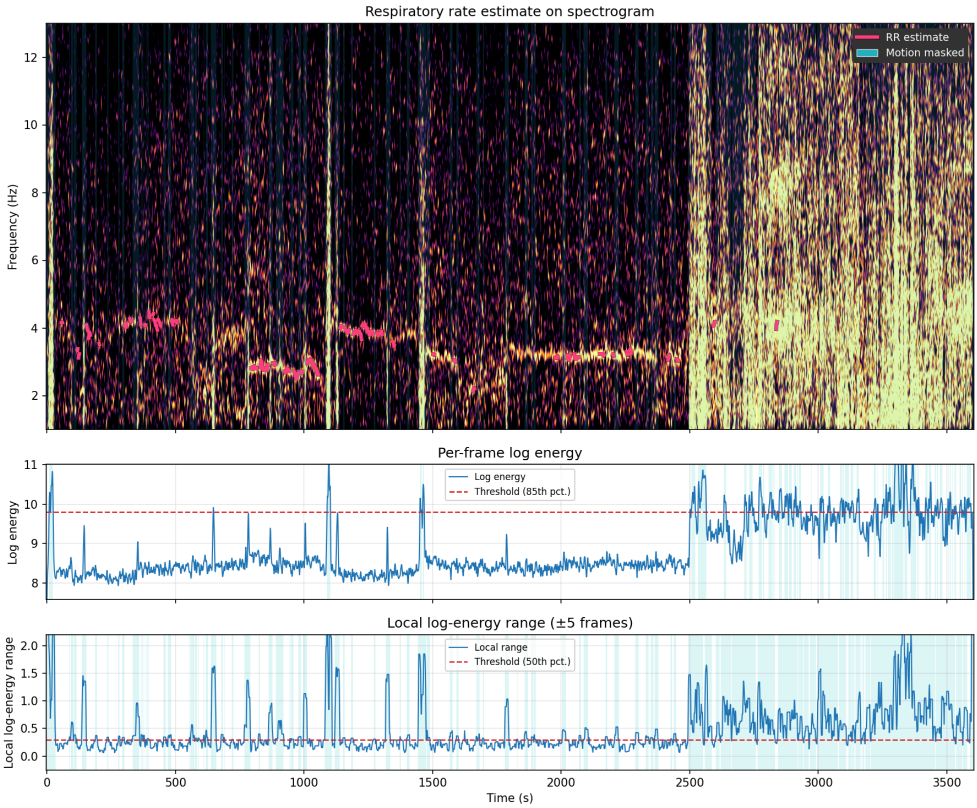

Figure 6: Final analysis pipeline on a representative real radar recording (session B1, hour 5): spectrogram with extracted RR (top). Motion-energy mask (bottom).

4.5 Agreement analysis (RQ1)

Bland-Altman of radar against TrackPaw RR is reported in three subsections on the same paired recordings: the pre-fix RR series, the post-fix RR series, and the post-fix RR series after the 30 s pre-pairing aggregation. Figure 7 shows the joint Bland-Altman view for each, and Table 3 summarizes the joint bias, joint LoA (per Bland-Altman 1999 [13]), n_pairs, and n_sessions for all three. Per-session detail for each subsection is in Appendix B. Boxplots overlaid summarize the distribution of differences within each 0.25 Hz mean-axis bin (median, IQR, 1.5× IQR whiskers), drawn for bins containing at least 100 pairs.

4.5.1 Pre-fix Bland-Altman

Across all sessions, the joint bias is +0.17 Hz with limits of agreement of −0.34 and +0.68 Hz. The systematically positive bias and slanted cloud motivated the investigation that surfaced the off-by-one bug in the platform's RR estimator. The box medians trend upward with the mean frequency, matching the slanted cloud. The IQR and whisker spread widen at the lower and upper edges of the mean range, where bin counts are smaller.

4.5.2 Post-fix Bland-Altman

With the off-by-one corrected, the joint bias drops to +0.04 Hz, with limits of agreement of −0.31 and +0.40 Hz. The bias has decreased substantially compared to pre-fix, and the limits of agreement have also tightened. The off-by-one accounts for the larger share of the systematic disagreement surfaced. The box medians flatten near the joint bias across the central bins, removing the upward trend seen pre-fix. The IQR and whisker spread still widen at the lower and upper edges of the mean range.

4.5.3 Aggregate Bland-Altman

With the 30 s pre-pairing aggregation step, the joint bias remains essentially unchanged at +0.03 Hz, but the limits of agreement tighten substantially to −0.14 and +0.20 Hz. Box IQR and whisker spread are roughly equal across mean-frequency bins, the expected outcome of aggregation.

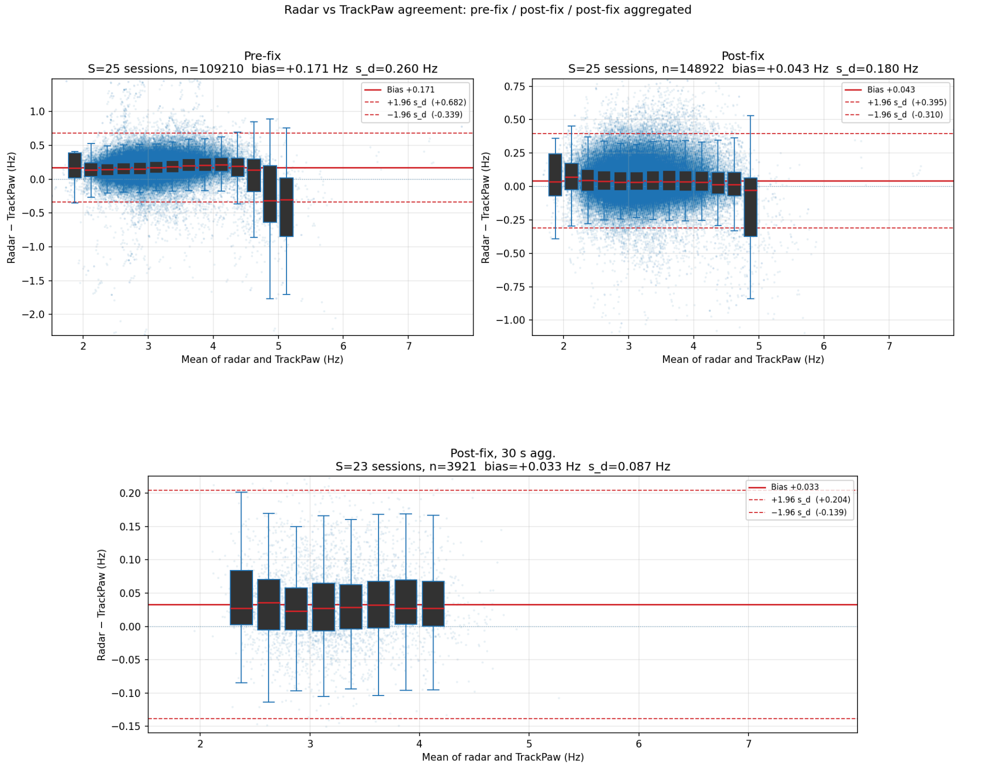

Figure 7: Joint Bland-Altman 1999 across pre-fix, post-fix, and post-fix with 30 s pre-pairing aggregation. Bias and 1.96 SD LoA bands marked per panel.

Subsection	**Joint $\overline{d}$ (Hz)**	**Joint LoA (Hz)**	n_{pairs}	n_{sessions}
Pre-fix	+0.171	-0.339 / +0.682	109210	25
Post-fix	+0.043	-0.310 / +0.395	148922	25
Post-fix, 30 s agg.	+0.033	-0.139 / +0.204	3921	23

Table 3: Joint Bland-Altman 1999 comparison across pre-fix, post-fix, and post-fix with 30 s aggregation.

4.6 Behavioral state robustness (RQ2)

RQ2 asks under which behavioral states each method becomes unreliable. Two views are reported. The radar's motion mask is first compared against TrackPaw travel as a cross-platform validation that the mask captures motion broadly defined. Per-method coverage across behavioral states is then reported, quantifying how often each method delivers a valid RR estimate as a function of motion.

4.6.1 Radar motion mask vs. TrackPaw travel

Agreement between the radar motion mask and the TrackPaw travel label is shown in Figure 8 as a confusion matrix on the kept windows pooled across all four cages, and broken down per cage in Table 4. The table reports the comparison at per-frame and per-window level as described in Section 3.9.

Across the four cages the per-frame mask fires often inside active windows and less often inside stationary windows, giving a per-frame lift above one. The per-window NPV is near one while the per-window PPV ranges near $60 - 70\%$. The kept-fraction is slightly below half of all windows.

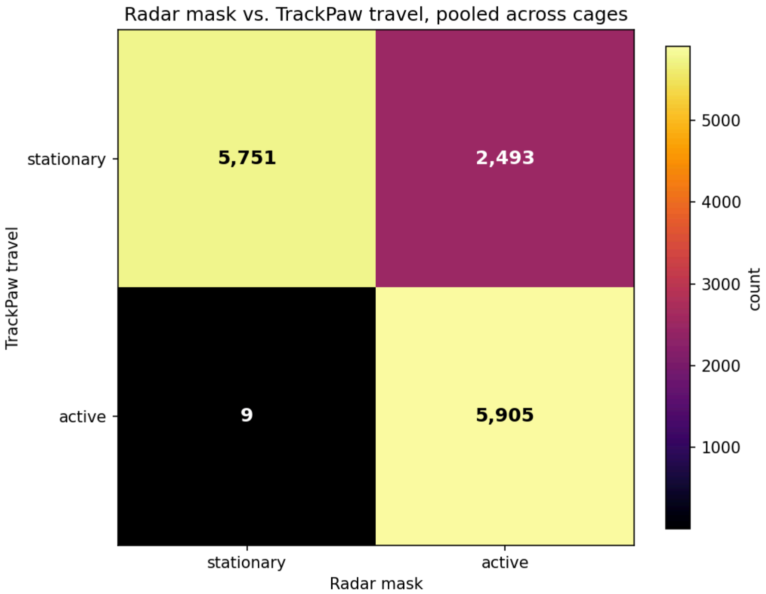

Figure 8: Confusion matrix between the radar motion mask aggregate label and the TrackPaw travel label, pooled across all cages over the 14158 windows where the radar aggregate produced a value. Cells show absolute window counts.

Cage	n_{kept}	$P(m\|a)$	$P(m\|s)$	Lift	NPV	PPV	Kept
A	3080	87.9%	38.8%	2.269	99.6%	59.1%	42%
B	4590	91.1%	37%	2.461	99.9%	76.6%	42.5%
C	4609	91.6%	38%	2.413	99.9%	67.5%	48.6%
D	1879	93.5%	35.8%	2.612	100%	79.9%	47.3%
joint	14158	90.9%	37.6%	2.42	99.8%	70.3%	44.8%

Table 4: Radar motion mask agreement with TrackPaw travel label, per cage and pooled (joint row). Per-frame, $P(m|a)$ and $P(m|s)$ are the conditional probabilities of the mask being active (m) given the enclosing travel label was active (a) or stationary (s). Lift is the ratio of the probabilities. Per-window, n_{kept} kept windows, NPV, PPV, kept-fraction, are computed on the windows where the radar aggregate label produced a value

4.6.2 Per-method coverage

Coverage results are shown in Figure 9 (a representative time-matched view) and Table 5 (per-method coverage averaged across all sessions). For session B1, the top panel shows all RR estimates from each method during the first 4 hours. The bottom panel shows time-matched paired RR estimates only, pre aggregation. Shading shows where the TrackPaw travel label or radar motion mask are active. RR estimates from either method are sparse within travel-shaded windows. One low-frequency radar cluster is visible where TrackPaw disagrees. In the later strongly shaded segment, TrackPaw produces some estimates and the radar a few, with one short cluster of paired estimates forming.

Table 5 reports per-method coverage and the non-motion fraction, mean ± SD across sessions. Radar produces no coverage within motion frames by construction, while TrackPaw produces a small amount. Most of the coverage from both systems falls in non-motion.

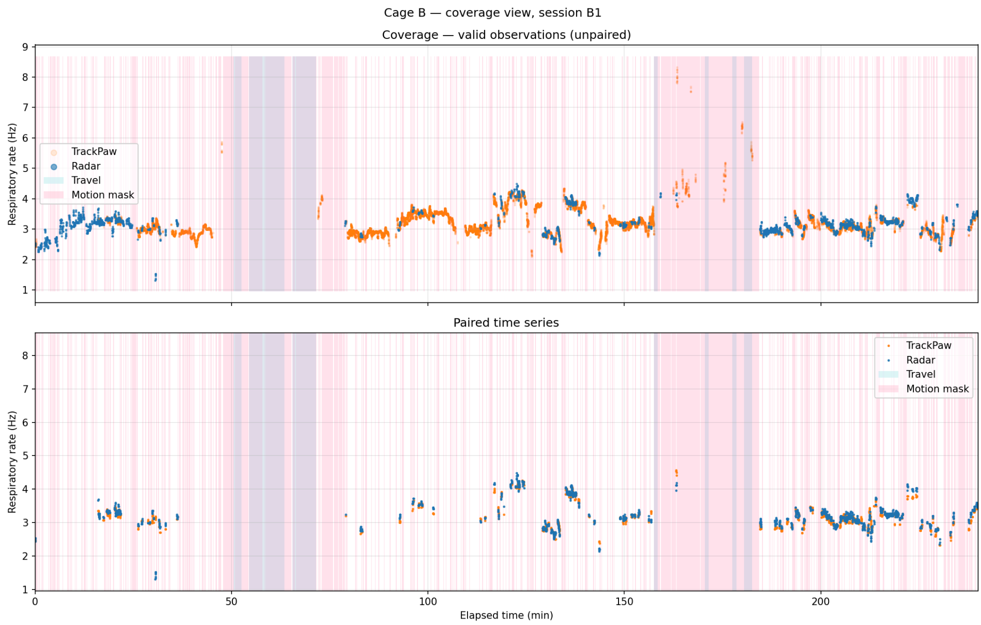

Figure 9: Coverage view on a representative session (B1, first 4 hours). Top: all RR estimates from each method. Bottom: time-matched RR pairs. Travel shading: 60 s bins with TrackPaw travel > 0 mm. Motion mask shading: per-frame radar energy motion mask as applied by the pipeline.

Method	Coverage	Non-motion fraction	Cov. in non-motion	Cov. in motion
radar	12.8 ± 6.5%	48.8 ± 3.2%	26.1 ± 12.8%	0 ± 0%
trackpaw	55.7 ± 14.4%	78.6 ± 8.7%	68.4 ± 13.1%	7.2 ± 3.3%

Table 5: Per-method coverage, mean ± SD across sessions. Each method uses a method-specific definition of non-motion. The coverage (cov.) in or outside motion values are not directly comparable across methods.

5 Discussion

The relationship between the synthetic and real signal regimes is examined first, followed by the framing of agreement without a designated ground truth. The agreement results addressing RQ1 and the coverage results addressing RQ2 are interpreted in turn, after which methodological lessons with directions for future work are presented, followed by the study's limitations. Societal and ethical aspects of the agreement-without-ground-truth framing, together with the application of the methods to clinical studies, close the chapter.

5.1 Synthetic evaluation compared to real data

As the results show, the pipeline achieves a near perfect match on the synthetic signal model ground truth (given an estimate is made). As Figure 5 shows, there are areas obscured by a scattered energy band, clear harmonics and some amount of background noise as well. Comparing to a real representative recording, Figure 6, scattered energy bands are more evenly scattered in real recordings and harmonics are not present. The high energy span around the synthetic ground truth is comparable in width to the energy span where the estimated real signal exists. The discrepancy of harmonics is not worrisome for the produced pipeline. Harmonics can exist and are accounted for as part of the pipeline step 6 but signals with no apparent harmonics are not algorithmically discarded. Although not visually indistinguishable, the spectrograms show comparative signal visuals. As for the scattered and noisy sections, both share their existence but not all together characteristic. Since the synthetic evaluation mainly served as a tool to track a ground truth and realistically mask areas of high uncertainty, the model was useful. However, the discrepancy in visual comparison to real data show that the model does not confidently describe how the real radar samples in areas of weak signal, even though it can mimic energy scattered and noise to some degree.

Had the model been generally more accurate in areas of discrepancy to that of real data, using the model for motion mask validation would be possible. The idea to compare TrackPaw travel and radar to imply that the motion mask is in fact motion is essentially a work around as to not rely on comparing travel from the synthetic model with the mask. One way to answer RQ2 if the model was more accurate would be to compare the model data with the model travel signal. Because of the visual model to real data discrepancy in areas of weak signal, the confidence in the model as ground truth for cross-behavioral robustness was low. Thus, the work around option is argued to be a closer to reality comparison for behavioral identification and comparison.

The choice of modeling per-bin amplitude in the synthetic benchmark as a Gaussian is probably not entirely accurate relative to the actual radar receiver characteristics. The model, like the radar, has a maximal amplitude when the reflection distance matches the bin distance. From there, the amplitude was required to fall off symmetrically with positive or negative displacement, continuously, and with all distance bins picking up a near but non-zero contribution at any reflection distance. The simplest function meeting these requirements was a Gaussian. The synthetic spectrogram in Figure 5 compared to the real recording in Figure 6 still shows reasonable and comparative results, so the Gaussian guess was treated as an adequate description of amplitude fall-off. Not constructing the synthetic benchmark from truer-to-life recordings of the scenario was nevertheless a limitation.

Synthetic respiration models for radar exist in the human-radar literature. Singh et al. simulated chest-wall motion from chemical-mechanical respiratory control mechanics, with high correlation against respiratory-belt recordings in human participants [18]. El Abbaoui et al. synthesized radar baseband signals from a sum-of-sinusoids chest-wall motion model in humans, used to train a deep-learning rate estimator [19]. The model used in this thesis instead targets mice at the radar-signal level, generating synthetic per-bin radar samples directly from a Gaussian amplitude profile around a single chest-wall reflector. No analogous simulator for the mouse-cage scenario was found in the literature reviewed.

5.2 Defining ground truth

In this thesis, two systems intend to measure respiratory rate and do so in unconventional ways. The radar system infers that breathing is a phase shift in electromagnetic backscatter as the chest-wall moves, TrackPaw infers that breathing is an oscillating force measurement in an expected range. The size of the movement or force measured is expectedly tiny. The chest-wall movement of a mouse and the corresponding force shift are assuredly small compared to that of a human for example. It is reasonable to intuitively guess that the measured quantity is so small that it simply should not consistently be measurable by the two systems. Perhaps the choice of describing the systems as unconventional stem from their unintuitive quality.

The use of a ground truth does not exist in the framing of the two systems in this thesis. Neither is a given gold standard and neither is labeled truth. When a system is not a standard, the only honest way to label the system the ground truth is if that system can, in large majority, be thought of as truth without proof. Such a truth relies heavily on group intuition to avoid opposition. A system that feels unequivocally functional to measure a phenomenon can be labeled a ground truth even if it is not a standard. In the case of measuring respiratory rate in mice, a human counting periodic chest-wall movements in a replayable camera recording on mice seems very plausible and could be labeled truth. It is as intuitive to believe as a human being able to count the number of swings a playground swing completes. Neither is a gold standard but both are believably true.

In this thesis, both systems used are non-standard and, to some degree, non-intuitive because of how small the measured phenomenon is. To avoid the opposition inherent to labeling one system truth, the systems are only compared in agreement. Something that the thesis uses much as a ground truth is the travel metric from TrackPaw. This is rather consistent with the intuition based labeling of a system as ground truth. Expecting an array of force-sensors to find a center of mass as it moves around is comparable to believing the human ability to detect the position of a mouse as it moves around on their back.

This intuition-based labelling of a ground truth is accepted elsewhere in practice. The Jackson Laboratory builds ground-truth datasets for the Envision home-cage monitoring platform by having human observers annotate the recordings, treating the manual annotation as labelled truth against which the machine-learning outputs are compared [20]. The platform is non-invasive and operates on freely-moving mice. The framing matches the argument above, a trained observer assessing a recording is intuitively believable as truth, even though the manual annotation is not a gold standard.

5.3 Agreement (RQ1)

Bland-Altman agreement with aggregates of 30 s remove time-adjacent dependency of samples from both systems, something that is important for Bland-Altman tests to avoid adjacent pair dependency. Moreover, pairing without aggregation lacks substantive claim. Pairing in nearest neighbor, when a common time aggregate does not exist to build pairs, is simple and should bin near simultaneously measurements together. However, the nearest neighbor requires a tight knowledge of the true sampling to noted timestamp error, and error between both systems. If there is no time difference between recorded time and true sample time in both systems, and both systems share a clock, nearest neighbor could match perfectly in time with a known max time difference. In essence, the correct pair assignment would be known. This is not the case and validating the exact timelag between recorded time and true sample time, as well as clock difference between the two systems, is highly speculative. For this reason, nearest neighbor has a larger than wanted max time difference allowed for shaping pairs allowing pairs that are clearly skewed from each other in time.

The choice of 30 s aggregation mitigates stated issues. TrackPaw averages 50 periods of RR when producing values. Assuming a minimal breathing rate of 2 Hz, the 50 period buffer is fully replaced after 25 seconds. For radar, the window size when shaping PSD is 8.2 s long with samples produced up to once per second, meaning neighboring samples are also dependent. Aggregating on 30 s removes dependency of RR estimates from both. Averaging all points per 30 s bin is a natural choice for a representative value per bin. Dropping bins with few (less than half of maximum) data points isolates comparison to aggregate values where the representative average is based on larger sample sizes. Moreover, the aggregate averages out outliers or small group readings that appear in areas of low SNR. Comparing to noted clusters in Section 4.6.2, spurious low frequency measurements could average with nearby data, and areas of high travel retain small groupings that are probably discarded as there are not enough estimate in the near region.

The aggregate Bland-Altman results (Figure 7, bottom plot, and Table 3 last row) show a high degree of agreement. There is a bias (ca. 0.03 Hz) between the RR estimates where the radars consistently show a higher RR than the paired TrackPaw estimate. The bias is equivalent with the quantization error of the radar analysis pipeline (0.12 Hz resolution with 4× zeropadding) but for the quantization error to persist as a systematic one step above after aggregation is unexpected. The quantization error should just as likely be one step down and an average over multiple samples should cancel any bias. Comparing to synthetic benchmarking in Table 2, no quantization step bias exist compared to ground truth either. Instead, the bias is left unexplained. One possibility is that another bug exists in the TrackPaw system but not one that was discovered.

Disregarding the source of the bias, a note must be made of the agreement the systems show. Assuming a breathing rate of 3 Hz, given a randomly chosen aggregate RR estimate from either system of exactly 3 Hz, the other one will predict an estimate between 2.86 – 3.20 Hz in the 95th percentile. As stated in the introduction, no agreement threshold will be argued as irrefutable evidence that the measured signal is respiration. What *is* argued is that both systems measure the same phenomenon. Perhaps the observed signal is another phenomenon entirely. Therefore, any and all readers are urged to hypothesize another method-agreeing signal ranging in a 2 – 4 Hz interval.

5.4 Robustness across behavioral states (RQ2)

How to aggregate the radar motion mask to 60 s windows was not obvious. With the goal of comparing travel as closely to the motion mask as possible, aggregating the radar label was done by dropping windows where the mask was ambiguous, only comparing windows where the mask was dominantly active or stationary. This loses an explanation and evaluation of what occurs in areas where the mask is more ambiguous. To supplement the per-window comparison, the conditional probabilities were calculated per-frame, losing no mask frames. Since the share of windows that were kept after radar label aggregation was substantial, both comparisons are used when relating the motion mask to travel in this section.

The simple answer to robustness is that neither system manages to estimate RR in areas of motion. TrackPaw, as seen in Table 5, has some coverage during travel. Since the travel metric natively can have shares of time of zero movement even when the 60 s window show travel > 0, the coverage is low enough that it could even reside entirely in moments during travel when the mouse is not moving. The dominant majority of all RR estimates from TrackPaw are nevertheless in windows of zero travel.

For the radar, defining and validating what motion is was difficult. The force-sensor platform can reasonably detect travel as a center of mass moving. The radar cannot even rely on the singular dimension of radial distance since a mouse could be moving at a constant radial distance to the radar. The so called motion mask assumes that a moving mouse equals more backscatter spread across the entire passband. The results in Figure 8 and Table 4 support this claim. The strong NPV agrees that when the mask does not flag a frame as active TrackPaw will, in nearly every case, agree that there has been no travel. However, this is on the aggregate motion mask label which does not evaluate the likelihood that any masked frame overlaps a travel window. Instead, the conditional probabilities do show that the mask is active more often given travel than no travel without aggregation. The statistics together show that a stationary radar flag will occur during no travel and an active flag will most likely be during travel, but the mask will undoubtedly be active in times of zero travel as well.

This is expected in two ways. Firstly, the radar mask is percentile-based and statistically removes areas of highest global and local energy. No matter how strong the signal is and how often it is clearly distinguishable by inspection in a spectrogram, there will be areas masked as motion. This is noticeable in locally masked slivers in Figure 6 where the signal is visually present but masked because of the tight threshold.

Secondly, the travel metric is a deadzone measurement. Any smaller movements or body posture adjustments done by the mouse would not count as travel but could affect the amount of backscatter and frequency content in the passband. Consider a mouse spinning in place. The chest-wall reflections would align with the radar receiver and unalign over time. There would be no travel recorded but radar measurements would be affected and probably disrupted. Rephrased, travel are larger net movements of mice, what radar detects and labels as motion correlates highly with travel but there are other motion patterns that also disrupt RR estimates from the system.

Table 5 also shows that the radar in total has very low total and in motion coverage compared to TrackPaw. While both struggle with travel, the radar struggles more overall with estimating RR.

Motion masking in prior radar respiration studies relies on the radar's own activity signal, either via internal range-Doppler power thresholds [21] or by recording while the subject is anesthetized and trivially still [3]. Neither verifies the masking decision against an external label of motion. Pairing the radar motion mask against TrackPaw travel offers a new way of verifying the radar's motion label. The agreement noted in Figure 8 and Table 4 supports the mask as an indicator of net motion, with the caveat that smaller postural movements not registered as travel can still disrupt the radar signal.

5.5 Methodological lessons and future work

Several improvements to the radar acquisition and pipeline became apparent during the work and could be addressed in a follow-up study. Visually from Figure 6, the signal appears to be traceable in more spots than the pipeline produces. There are performance improvements that could be done with the recorder data, most importantly in avoiding false readings. For even better results, many radar parameters could be tweaked. As a recommendation, using higher profiles and filter direct leakage could up the SNR, a tuned HWAAS could reduce noise, another framerate, perhaps a lower one to allow for higher HWAAS, could also be tuned. Other parameters not mentioned in the thesis could possibly aid as well.

In the pipeline, a more carefully chosen welch sliding window approach might help remove areas of fluctuating signal strength, leveraging an average in the window. The spectrogram is rather noisy which welch's method might help mitigate. Simply removing the negative frequency spectrum was also an oversight. Folding the negative frequencies onto the positive ones (averaging per frequency) could contain more information on signal than simply ignoring it.

Given enough real data recorded, radar research and modeling knowledge, perhaps a more sophisticated synthetic model can be built, including static reflectors at specific points with realistic phase jitter and amplitudes. If real data was collected with a ground truth comparison (e.g. computer vision or manual annotation methods), creating a synthetic model from destructured real recordings could be possible.

Beyond pipeline and modeling improvements, the noise-floor handling itself was a methodological limit worth flagging. Using fixed noise floor division has qualitative limits. The same noise floor recording was applied across all sessions, although bedding and enrichment, altered between sessions, would change the static reflector profile of the cage. The empty-cage recordings shown in Figure 4 already differ between cages. A study retaining the recorder noise floor division should record one noise floor per session to mitigate drift. CFAR methods (Section 2.2.5), although not thoroughly explored because of an early synthetic benchmark regression, might remove the need for a separate noise floor recording entirely by adjusting dynamically.

5.6 Limitations

For selection of mice the only requirement was mice of weight 20 grams or above. No constraint on strain, sex or age was made since all were assumed to have comparable radar cross section and size in chest-wall displacement. This was not backed by literature and simply assumed. The weight of mice was deemed as correlated with the size of mice and therefore chest-wall area. This, much in the same way, was an assumption. Maximizing chest-wall area was expected to increase the radar reflection sought. This motivated the minimum weight criterion. Not researching the biological factors impact on radar, or comparing during the study, was an oversight that might have yielded a stronger signal. Comparable rodent radar and camera studies restrict inclusion to a single strain and sex. Polonelli et al. recorded female Swiss Webster mice only [21], and Breuer et al. recorded three male Sprague Dawley rats in a 360-375 g, 9-11 week range [22]. The lack of a literature-backed inclusion protocol may have widened cross-session variability.

Sample size is always a limitation. The repeated measures Bland-Altman agreement plot estimates a variance of biases from each sessions, of which there are only a few. The low population does increase variation and uncertainty as to the limits of agreement between methods.

Every pipeline revision was accepted or rejected on a non-regression gate against the synthetic-signal benchmark, so the synthetic model also served as the optimization target for the pipeline. Reported synthetic metrics therefore describe a best-case ceiling rather than independent validation, and strong synthetic performance cannot on its own establish real-world robustness. The same caution applies to noise-floor normalization on the synthetic data, where the reference was the same generator with the animal contribution removed. That reference matched the synthetic noise exactly and gave a near-ideal normalization, so the synthetic results may overstate how well the normalization copes with the imperfectly characterized noise of real recordings.

The configuration validation in Section 3.4 confirmed that the radar responded to small movements of a known target, but the test was qualitative. The displacement applied to the corner reflector was not measured, so detection at the roughly 0.3 mm scale of mouse chest-wall motion used in the synthetic model was never verified directly, and the procedure as performed is not reproducible. A controlled test driving a reflector through a known displacement on the order of the modeled chest-wall motion would strengthen the validation and make it reproducible.

5.7 Societal and ethical aspects

The idea of comparing two methods for agreement without a ground truth could, at its peak, be deemed completely worthless. If both methods show high levels of agreement, it is defensible stating that both are measuring the same phenomenon. It is another thing to know that the phenomenon in question is the intended phenomenon. If the two methods instead do not show agreement at all, the result is essentially uninterpretable regarding what is being measured or which one, if any of the two, is correct.

The main reasons this thesis chose an agreement analysis without ground truth was twofold. The primary reason was technical knowledge and interest for the two methods used. The secondary reason was based on a want to fully remove any confounders in measurement related to invasive or stressing factors for the animal.

This thesis set out to investigate agreement between non-invasive, even non-interacting, methods to build confidence in the technology for use in future animal research. The goal to promote a method in line with the principle of 3Rs felt dishonest if the comparing method did not manage to reduce pain and stress for the animals or refine valuable data per study. For this reason, and to remove confounder in measurements, it was important to the thesis that all methods used were fully non-interacting with the animal and that the study protocol allowed for full acclimatization and stress free management of animals.

The force-sensor platform reads RFID chips implanted in the mice, and implantation is an invasive procedure. These chips were part of routine husbandry at the facility and predated the study rather than being placed for it. Because every cage held a single mouse, the identity attribution the chips provide was not used in any comparison, so the implants did not enter the measurements and do not bear on the non-interacting character of the methods evaluated here. The implants nonetheless remain an invasive husbandry reality that the non-interacting framing of the compared methods does not remove.

5.8 Application to clinical studies

As both systems showed a high level of agreement, it is not a disregardable option to continue investigating the use of the systems in clinical studies. The relevant question is application and where either system is useful.

Plethysmography measures more than rate. It records tidal volume and the airflow at each inhalation and exhalation [1]. Neither system compared in this thesis captures these. The radar pipeline reports rate alone, with the phase-derived displacement reduced to a spectrum peak rather than a per-breath waveform. The force-sensor platform may carry richer waveform content in the raw signal, although only the averaged rate estimate was used in this thesis. For studies that require volume or per-breath morphology, plethysmography or an implanted sensor remains the appropriate tool.

The value of the two systems is found at a different scale. Long-period averages and trends in respiratory rate can be tracked continuously and without interaction, where plethysmography would require restraint or a sealed chamber. The advantage is not precision at one moment in time but coverage over hours or days, applied to questions where average rate, trend, or behavior-conditioned shifts in rate are the relevant readouts. Studies of stress response, drug effect, sleep, or circadian variation can be served by a continuous non-invasive rate metric. Studies of single-breath morphology cannot.

6 Conclusion

For RQ1, radar-derived and TrackPaw-derived RR estimates showed close agreement on the same animals recorded simultaneously. With pairing on 30 s aggregates, the joint bias was +0.03 Hz with limits of agreement of −0.14 and +0.20 Hz (Figure 7, Table 3). Without aggregation, the post-fix joint bias was +0.04 Hz with limits of agreement of −0.31 and +0.40 Hz. The agreement supports the interpretation that the two systems, built on independent physical principles, track the same phenomenon in the breathing-frequency band. A small positive bias persisted after aggregation that the analysis could not attribute to a known source.

For RQ2, neither system reliably produced RR estimates inside motion windows (Table 5). The radar pipeline by construction returned no estimates inside motion-flagged frames, and the force-sensor platform produced only a small fraction of estimates inside travel windows. The bulk of RR coverage from both methods fell in non-motion. The radar motion mask agreed strongly with TrackPaw travel as a motion indicator (Figure 8), with a high negative predictive value supporting use of the mask as the radar pipeline's coverage gate. The moderate positive predictive value was consistent with the percentile-based mask design and with motion patterns that disturb the radar without registering as travel.

Under the conditions tested, continuous non-invasive RR monitoring of laboratory mice in the home cage appears feasible during rest, with the radar pipeline and the force-sensor platform showing tight mutual agreement on aggregated estimates. Reliable RR estimation during active behavior remains an open problem for both methods.

Several extensions follow from the limitations identified. On the radar acquisition side, configuration could be tuned further, with higher profile selection and direct-leakage filtering targeting better SNR, and a tuned HWAAS at a lower frame rate offering further noise reduction. On the pipeline side, a more carefully chosen Welch sliding-window approach could mitigate the noisy spectrogram, and folding negative frequencies onto the positive half could retain information currently discarded. The synthetic-signal benchmark could be extended with additional static reflectors carrying realistic phase jitter and amplitudes, ideally derived from real recordings paired with an independent ground-truth annotation such as computer vision or manual labeling. Acquisition itself would benefit from a per-session empty-cage noise-floor recording, or alternatively from dynamic adjustment via CFAR methods that could remove the dependence on a single fixed reference.

References

[1] J. Grimaud and V. N. Murthy, “How to Monitor Breathing in Laboratory Rodents: A Review of the Current Methods,” *Journal of Neurophysiology*, vol. 120, no. 2, pp. 624–632, 2018, doi: 10.1152/jn.00708.2017.

[2] W. M. S. Russell and R. L. Burch, *The Principles of Humane Experimental Technique*. London, U.K.: Methuen, 1959.

[3] G. Sun *et al.*, “Contactless Monitoring of Respiratory Rate Variability in Rats Under Anesthesia With a Compact 24-GHz Microwave Radar Sensor,” *Frontiers in Veterinary Science*, vol. 12, p. 1518140, 2025, doi: 10.3389/fvets.2025.1518140.

[4] M. A. Richards, *Fundamentals of Radar Signal Processing*, 2nd ed. New York, NY, USA: McGraw-Hill Education, 2014.

[5] Acconeer AB, “A121 Pulsed Coherent Radar (PCR) Datasheet.” 2025.

[6] Acconeer AB, “A111 Profiles and Direct Leakage.” Accessed: Apr. 25, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/a111/radar_data_and_control/profiles.html

[7] Acconeer AB, “Signal Strength (HWAAS, Profile, Subsweeps).” Accessed: Apr. 25, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/radar_data_and_control/a121/signal_strength.html

[8] Acconeer AB, “A121 Timing Model.” Accessed: Apr. 25, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/radar_data_and_control/a121/timing.html

[9] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2005.

[10] SciPy developers, “scipy.signal.filtfilt — Zero-Phase Forward-Backward Filtering.” Accessed: Apr. 25, 2026. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.filtfilt.html

[11] SciPy developers, “scipy.signal.ShortTimeFFT — Short-Time Fourier Transform.” Accessed: Apr. 25, 2026. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.ShortTimeFFT.html

[12] J. M. Bland and D. G. Altman, “Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement,” *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986, doi: 10.1016/S0140-6736(86)90837-8.

[13] J. M. Bland and D. G. Altman, “Measuring Agreement in Method Comparison Studies,” *Statistical Methods in Medical Research*, vol. 8, no. 2, pp. 135–160, 1999, doi: 10.1177/096228029900800204.

[14] Acconeer AB, "Breathing Reference Application (A121)." Accessed: May 07, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/ref_apps/a121/breathing.html

[15] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978, doi: 10.1109/PROC.1978.10837.

[16] SciPy developers, "scipy.ndimage.median_filter." Accessed: Apr. 25, 2026. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.median_filter.html

[17] SciPy developers, "scipy.signal.find_peaks and peak_prominences." Accessed: Apr. 25, 2026. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html

[18] A. Singh, S. U. Rehman, S. Yongchareon, and P. H. J. Chong, "Modelling of Chest Wall Motion for Cardiorespiratory Activity for Radar-Based NCVS Systems," *Sensors*, vol. 20, no. 18, p. 5094, 2020, doi: 10.3390/s20185094.

[19] A. El Abbaoui, D. Sodoyer, and F. Elbahhar, "Contactless Heart and Respiration Rates Estimation and Classification of Driver Physiological States Using CW Radar and Temporal Neural Networks," *Sensors*, vol. 23, no. 23, p. 9457, 2023, doi: 10.3390/s23239457.

[20] The Jackson Laboratory, "Analytical Validation of Digital Mouse Detection, Identification, and Activity in the Envision Platform," White paper, 2025. Accessed: May 13, 2026. [Online]. Available: https://resources.jax.org/white-papers/analytical-validation-of-digital-in-vivo-mouse-detection-identification-and-movement-in-the-envision-platform

[21] T. Polonelli, M. Glahn, S. Kron, S. Selbert, M. Garzola, and M. Magno, "A Non-Invasive Path to Animal Welfare: Contactless Vital Signs and Activity Monitoring of In-Vivo Rodents Using a mm-Wave FMCW Radar." Accessed: May 13, 2026. [Online]. Available: https://arxiv.org/abs/2512.05595

[22] L. Breuer *et al.*, "Camera-Based Respiration Monitoring of Unconstrained Rodents," *Animals*, vol. 13, no. 12, p. 1901, 2023, doi: 10.3390/ani13121901.

Appendices

Appendix A Sensor and pipeline parameters

Parameter	**Value**	**Unit**	**Rationale**
frame_rate	250	Hz	Section 3.2
HWAAS	284	—	Section 3.2
profile	2	—	Section 3.2 / Section 2.1
start_point	60	—	Section 3.2
step_length	8	—	Section 3.2
num_points	10	—	Section 3.2
inter_frame_idle_state	READY	—	Section 3.2
inter_sweep_idle_state	READY	—	Section 3.2
sweeps_per_frame	1	—	Section 3.2
sweep_rate	None (auto)	—	Section 3.2
continuous_sweep_mode	False	—	Section 3.2
double_buffering	False	—	Section 3.2

Table 6: Radar parameter listing with rationale.

Parameter	**Value**	**Unit**	**Rationale**
analysis_band_min	1.0	Hz	Section 3.6 step 2
analysis_band_max	13.0	Hz	Section 3.6 step 2
filter_pass_min	0.5	Hz	Section 3.6 step 2
filter_pass_max	13.5	Hz	Section 3.6 step 2
filter_order	5	—	Section 3.6 step 2
stft_win_len	2048	samples	Section 3.6 step 3
stft_hop_div	8	—	Section 3.6 step 3
stft_mfft_factor	4	—	Section 3.6 step 3
energy_fraction_threshold	0.30	—	Section 3.6 step 6
sole_freq_dominance_ratio	0.30	—	Section 3.6 step 7
motion_global_pct	85	%	Section 3.6 step 8
motion_spike_pct	50	%	Section 3.6 step 8
motion_spike_window	10	frames	Section 3.6 step 8
snr_threshold	2.0	—	Section 3.6 step 9
jump_filter_thresh	0.2	Hz	Section 3.6 step 11
jump_filter_window	8	frames	Section 3.6 step 11
isolation_min_neighbors	2	—	Section 3.6 step 12
isolation_window	8	frames	Section 3.6 step 12

Table 7: Final radar analysis pipeline parameters listing.

Appendix B Study sessions

Session		Start	Dur (h)	Animal	Strain	Sex	Age (wk)
Cage	#						
A	1	2026-04-07 11:50	20.2	8219	C57bl/6	M	16
A	2	2026-04-09 09:19	22.5	8219	C57bl/6	M	16
A	3	2026-04-13 09:26	22.8	8219	C57bl/6	M	16
A	4	2026-04-15 09:22	23.1	8219	C57bl/6	M	16
A	5	2026-04-22 10:16	23.2	5298	C57bl/6	F	15
A	6	2026-04-24 08:20	0.0	5432	C57bl/6	F	15
A	7	2026-04-28 07:06	2.8	4859	C57bl/6	F	15
B	1	2026-04-07 12:46	19.3	6733	C57bl/6	M	16
B	2	2026-04-09 10:44	21.4	6733	C57bl/6	M	16
B	3	2026-04-13 09:41	20.8	6733	C57bl/6	M	16
B	4	2026-04-15 10:55	21.4	6733	C57bl/6	M	16
B	5	2026-04-22 10:19	23.7	6186	C57bl/6	F	15
B	6	2026-04-23 10:45	22.1	6097	C57bl/6	F	15
B	7	2026-04-27 12:10	21.0	5909	C57bl/6	F	15
B	8	2026-04-28 12:34	17.7	4322	C57bl/6	F	15
C	1	2026-04-07 12:37	19.6	6342	C57bl/6	M	16
C	2	2026-04-09 10:45	21.1	6342	C57bl/6	M	16
C	3	2026-04-13 09:18	23.0	6342	C57bl/6	M	16
C	4	2026-04-15 09:34	22.9	6342	C57bl/6	M	16
C	5	2026-04-22 10:57	23.1	4322	C57bl/6	F	15
C	6	2026-04-27 11:49	22.1	5298	C57bl/6	F	15
C	7	2026-04-28 11:16	18.7	5432	C57bl/6	F	15
D	1	2026-04-22 10:47	23.4	5909	C57bl/6	F	15
D	2	2026-04-27 12:26	21.4	6186	C57bl/6	F	15
D	3	2026-04-28 10:40	17.4	6097	C57bl/6	F	15

Table 8: Strain, sex and age in weeks (wk) per animal and session, with specified durations (dur). Animal ID is RFID extracted, equal ID is same animal. Age is noted as recorded at first session occurrence of an animal and not changed with new sessions of the animal.

Appendix C 3D-printed components

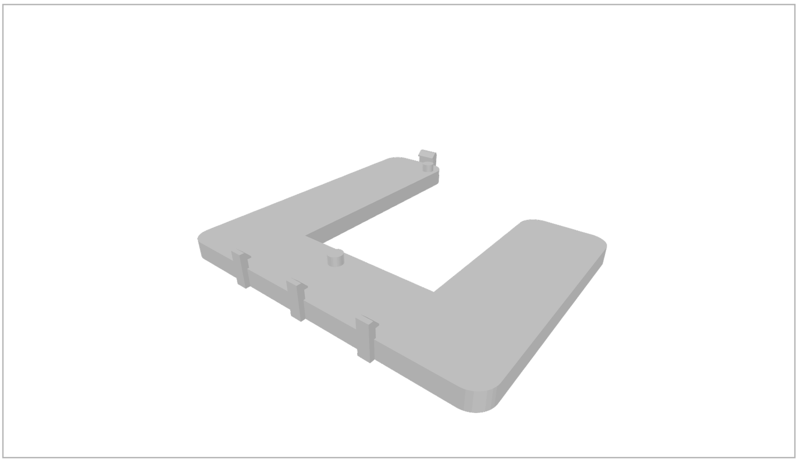

Figure 10: XE121 Mount baseplate. Dimensions: 68 × 67 mm.

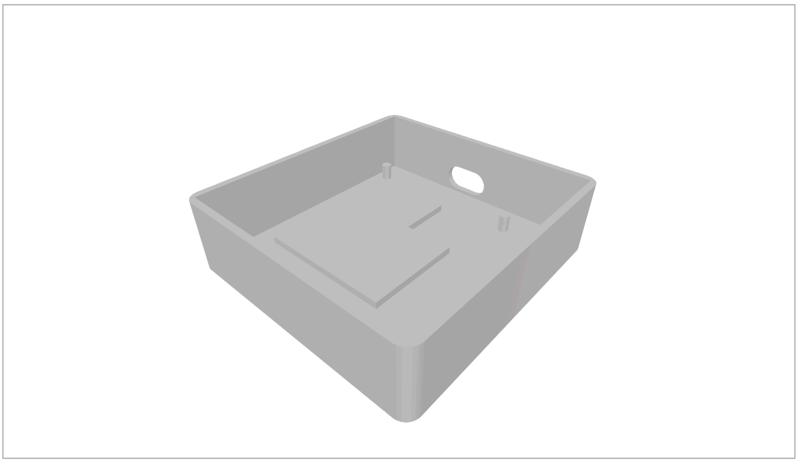

Figure 11: XC120 Shield. Dimensions: 78 × 74 × 21 mm.

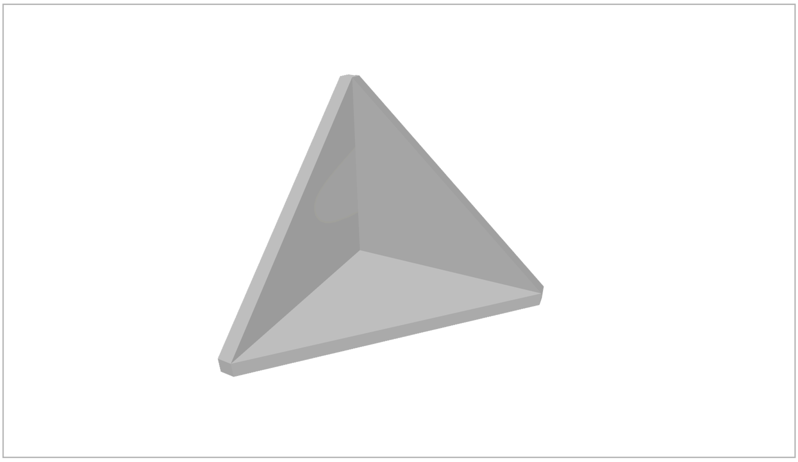

Figure 12: Trihedral corner reflector — three orthogonal 40 × 40 × 40 mm faces.