



Privacy-Preserving Presentation Attack Detection Using a Low-Resolution Thermal Sensor Array

by

Khawja Redwanul Islam

Supervisor: Dr Abdallah Naser

Project submitted as part of the requirements for the award of
MSc Data and Artificial Intelligence

August 2026

Declaration

DECLARATION

I, _____ Khawja Redwanul Islam _____ declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed: _____

Date: _____

FORM OF CONSENT

I, _____ Khawja Redwanul Islam _____ hereby consent that my Project, submitted in candidature for the MSc Data and Artificial Intelligence degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed: _____

Date: _____

Abstract

Facial biometric systems are vulnerable to presentation attacks, in which an artefact such as a printed photograph or a screen replay is presented in place of a genuine, living user. Presentation attack detection (PAD) is the countermeasure that decides whether a live person is present before recognition proceeds. Most published PAD systems rely on high-resolution visible-light or thermal imagery, which is effective but is itself privacy-invasive, since the sensor captures enough detail to identify the person it is meant only to check for liveness.

This project investigates whether PAD is achievable using a thermal sensor array of such low resolution that it cannot capture an identifiable face by construction. Using a Melexis EVB90640-41 evaluation board fitted with an MLX90641 sensor, logging 16×12 (192-pixel) thermal frames, a dataset of 1,850 frames was collected across 15 recording sessions from two participants, spanning four attack families: printed photographs, screen replays, warm-water heat decoys, and a combination attack that heats a photograph to defeat a temperature-only defence. Two physically motivated cues were extracted from each frame – peak temperature and the spatial topology of the warm region – and a one-class detector was fitted on genuine frames only, so that no attack data is required during training. Fusing the two cues achieves an Average Classification Error Rate (ACER) of 0.092 (APCER 0.039, BPCER 0.146) under session-grouped cross-validation, outperforming fifteen classical classifiers, five one-class anomaly detectors, and eleven deep architectures including a frozen 85-million-parameter Vision Transformer, while requiring under one kilobyte of stored model parameters.

The method also generalises to attack families withheld entirely from training (mean ACER 0.091), which classical supervised models trained on the other three families cannot match (mean ACER 0.34–0.53). However, the detector does not yet generalise across people: trained on one participant and tested on the other, the best cross-participant ACER is 0.167, and most configurations reject nearly every genuine frame from the unseen participant. This generalisation gap is reported as the project’s central finding and its principal limitation, and is only measurable because a second participant was recorded. The dissertation concludes that liveness detection at privacy-preserving thermal resolution is feasible within a known user population, but that person-independent generalisation remains an open problem requiring a larger participant pool.

Keywords: presentation attack detection, liveness detection, thermal imaging, low-resolution sensing, privacy-preserving biometrics, one-class classification, anomaly detection.

Acknowledgements

I would like to thank my supervisor, Dr Abdallah Naser, for his guidance throughout this dissertation. He made the project possible in the first place by lending me the Melexis evaluation board around which all of this work is built, and his supervision consistently pushed me toward testing my assumptions rather than asserting them. Several of the findings I am most pleased with came from being asked to check something I had taken for granted. I am grateful to him also for arranging laboratory access and for taking the setup photographs reproduced in Chapter 3.

I would also like to thank my sister and my brother for their help with data collection. Recording thermal data requires someone willing to sit still in front of a sensor for minutes at a time, more than once, and often at inconvenient hours. Their willingness to give up that time made the second half of this project possible, and the resulting data changed what I was able to claim.

Finally, I would like to thank my friends and family for their patience and encouragement over the course of this project.

Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
1 Introduction	1
1.1 Background	1
1.2 Problem statement	1
1.3 Research question and objectives	2
1.4 A note on scope and the sensor actually supplied	3
1.5 Contributions	3
1.6 Dissertation structure	3
2 Literature Review	4
2.1 Biometric presentation attacks and why they matter	4
2.2 Evaluating PAD: the ISO/IEC 30107-3 framework	4
2.3 PAD beyond the visible spectrum	5
2.4 Low-resolution thermal sensing and privacy	7
2.5 Generalisation: unseen attacks and unseen users	7
2.6 The physiology of facial temperature	8
2.7 Summary and the research gap	9
2.7.1 Systematic novelty search	10
3 Methodology	12
3.1 Sensor and software setup	12
3.2 Capture protocol	14
3.3 Ethical considerations	14
3.4 The dataset	15
3.5 Data quality	17
3.6 A methodological error found and corrected	18
3.7 Analysis pipeline	18

4	Design and Implementation	19
4.1	Design principle: work upward in complexity	19
4.2	Feature engineering	19
4.2.1	Temperature features	19
4.2.2	Topology features	20
4.3	Baseline thresholds	20
4.4	Classical supervised classifiers	20
4.5	One-class anomaly detection	20
4.6	The fused two-cue detector	21
4.7	Face-region cropping	21
4.8	Neural networks and transfer learning	22
4.9	Deployment feasibility	22
5	Results and Discussion	23
5.1	Temperature: necessary but not sufficient	23
5.1.1	Distance	23
5.1.2	Why temperature alone fails	24
5.1.3	A difference between participants	25
5.2	Topology: where the signal actually is	25
5.3	Are the topology features redundant?	26
5.4	Baselines and models	28
5.4.1	Single thresholds	28
5.4.2	Classical supervised models	29
5.4.3	One-class detection	29
5.5	Combining two cues: the principal result	30
5.6	Performance against attacks never seen in training	31
5.7	Cross-participant generalisation: the key limitation	32
5.8	Face-region cropping: an attempted fix that did not work	33
5.9	Neural networks and transfer learning	34
5.10	Overall ranking	35
5.11	Could this be built as working software? Deployment feasibility	36
5.12	Discussion	37
6	Conclusions and Recommendations	38
6.1	Revisiting the research question	38
6.2	Objectives revisited	38
6.3	Limitations	39
6.4	Recommendations for future work	40
6.5	Closing statement	41

7 Reflection	42
7.1 On a result that was too good	42
7.2 On discarding a recording while the participant was still in the room . . .	43
7.3 On not taking the better number	43
7.4 On expecting the big networks to win	44
7.5 On finding closely related work late	44
7.6 On the participants I did not have	45
Appendices	52
A Dataset Composition and Reproducibility	53
A.1 Per-session breakdown	53
A.2 Repository layout	54
A.3 Reproducing any result	54
B Ethics Approval and Participant Materials	56
B.1 Approval	56
B.2 Participant materials	56
B.3 Data protection measures	57
C Analysis Code Index	58
C.1 Key parameters	59
C.2 The three rules the suite enforces	60
D Complete Result Grids	61
D.1 Classical supervised models	61
D.2 One-class detectors	65
D.3 Cross-participant generalisation	66
D.4 Unseen attack families	67
D.5 Crop-window sensitivity sweep	69
D.6 Pretrained backbones	70

List of Figures

3.1	Laboratory capture setup	13
3.2	Frames per recording session	16
5.1	Peak face temperature against distance	24
5.2	Peak temperature by session with the live band shaded	25
5.3	Mean thermal frame per scenario	25
5.4	Topology features in two principal components	26
5.5	Correlation between the twelve topology features	27
5.6	Best result for each classical model	29
5.7	One-class anomaly detectors	30
5.8	Attacks rejected by the fused detector, per family	31
5.9	Within-participant against cross-participant performance	32
5.10	ACER across sixteen crop-window sizes	33
5.11	Custom networks and frozen ImageNet backbones	35
5.12	Error rates against aggregation window length	36

List of Tables

3.1	Composition of the completed dataset.	15
5.1	Conditioning of the topology feature set (bona fide frames).	28
5.2	Temperature, topology and fused two-cue detector, under session-grouped cross-validation.	30
5.3	Best deep-learning results (selected architectures).	34
5.4	Overall ranking of the strongest result per approach.	35
A.1	Every recording session in the dataset.	53
A.2	Project repository layout.	54
C.1	The analysis suite.	59
D.1	Complete classical supervised model grid.	61
D.2	Complete one-class detector grid.	65
D.3	Complete cross-participant grid.	66
D.4	Complete unseen-attack grid.	67
D.5	Complete crop-window sweep.	69
D.6	Complete pretrained-backbone grid.	70

Chapter 1

Introduction

1.1 Background

Biometric authentication systems, and facial recognition in particular, are increasingly used to secure personal devices, restricted spaces, and financial transactions. A well-documented vulnerability of these systems is the *presentation attack*: an attempt to deceive the sensor by presenting an artefact – a printed photograph, a face replayed on a screen, or a mask – in place of a genuine, living user [1]. Presentation attack detection (PAD), also called liveness detection, is the countermeasure that decides whether the biometric sample originates from a real person before recognition is permitted to proceed. It is important to state this distinction plainly, because it is easily conflated with the recognition task itself: PAD does not ask *who* is present, only *whether* a genuine living person is present at all.

The great majority of published PAD systems operate on visible-light or near-infrared imagery, often augmented with depth or high-resolution thermal channels [2]. These approaches are effective – George et al. [2] report an ACER of 0.3% on the WMCA multi-channel database – but they share a structural weakness: the same high-resolution imagery that supports reliable liveness detection is also rich enough to identify the person, which is precisely the privacy exposure a liveness check should not need to create. Thermal imaging offers a physically motivated alternative cue. A living human face maintains a structured pattern of surface heat, radiating from the vasculature beneath the skin, that inert artefacts such as paper or an LCD screen do not reproduce [3]. This project asks what happens to that argument at the opposite extreme of resolution: whether a thermal sensor coarse enough to make identification impossible in principle can still carry enough signal to make liveness detection possible in practice.

1.2 Problem statement

Existing thermal PAD research demonstrates that thermal imaging is a strong liveness cue, but does so using full-resolution thermal cameras that cost hundreds to thousands of pounds

and that themselves capture recognisable thermal faces [3, 2]. Separately, a distinct line of work – including that of this project’s supervisor – has established that very low-resolution thermal arrays (16×12 to 32×24 pixels) are sufficient for privacy-preserving human sensing tasks such as distance estimation, occupancy counting, and activity recognition [4, 5, 6, 7]. These two threads have rarely been combined, and where they have been, the thermal array has been paired with a conventional RGB camera, which reintroduces the privacy liability the low resolution was chosen to avoid (Section 2.7.1). The question of whether a thermal array coarse enough to be privacy-preserving by construction is, on its own, fine enough to support presentation attack detection therefore remains open with no established public dataset to answer it.

1.3 Research question and objectives

The research question that motivated this project, as agreed with the supervisor Dr Abdallah Naser, was:

Can a low-resolution thermal sensor array reliably distinguish a live human face from presentation attacks – printed photographs, digital screen replays, and heat-based decoys – in a privacy-preserving manner, and which lightweight approach performs best under this resolution constraint?

Five objectives followed from this question:

1. Assemble a thermal capture rig around the supervisor’s Melexis EVB90640-41 evaluation board and build a capture and logging pipeline.
2. Establish baseline behaviour – the working distance, the stable temperature range of a live face, and the sensor’s practical characteristics.
3. Collect a purpose-built dataset of bona fide and attack presentations, since no public dataset pairs this sensor with a representative set of attacks.
4. Compare lightweight, CPU-friendly detection approaches suited to a very small per-frame feature space.
5. Evaluate detection performance using the standard PAD metrics defined in ISO/IEC 30107-3 [8]: the Attack Presentation Classification Error Rate (APCER), the Bona Fide Presentation Classification Error Rate (BPCER), and their mean, the Average Classification Error Rate (ACER).

1.4 A note on scope and the sensor actually supplied

The original proposal specified the Melexis MLX90640, which produces a 32×24 (768-pixel) frame. The evaluation board actually supplied logs 16×12 (192 pixels) under every configuration tested, which is consistent with the board being fitted with the related MLX90641 die rather than the MLX90640; both chips share a dual-footprint evaluation board and an ambiguous silkscreen. This discrepancy is documented in full in Section 3.1, together with the diagnostic steps taken to rule out a software or configuration cause, and concludes that the board is populated with the MLX90641. The pipeline built for this project detects and handles either frame size automatically, so the discrepancy affects the interpretation of one experiment (face-region cropping, Section 5.8) rather than the feasibility of the project as a whole.

1.5 Contributions

This project makes four contributions. First, it provides an empirical answer, previously absent from the literature, to whether presentation attack detection is achievable at a resolution low enough to be privacy-preserving by construction. Second, it contributes a purpose-built dataset of 1,850 thermal frames across four attack families and two participants, collected under a documented and iteratively corrected capture protocol. Third, it introduces a two-cue fusion detector – peak temperature combined with a topological description of the warm region – that outperforms every classical and deep model evaluated against it while requiring under a kilobyte of stored parameters. Fourth, and identified only once a second participant existed to test it, it reports a measured cross-participant generalisation gap as a central finding rather than a hidden weakness, which is offered as the project’s most scientifically important contribution.

1.6 Dissertation structure

Chapter 2 reviews the literature on presentation attack detection, thermal and multi-channel PAD, low-resolution privacy-preserving thermal sensing, one-class generalisation to unseen attacks, and the physiological basis of facial temperature, and states the gap this project addresses. Chapter 3 describes the sensor setup, the capture protocol and the lessons learned while developing it, the resulting dataset, and a methodological error in validation that was found and corrected. Chapter 4 describes the feature engineering, the model families compared, and the design of the fused detector. Chapter 5 presents the results in full, including the cross-participant generalisation experiment and the face-cropping experiment that did not succeed. Chapter 6 draws conclusions against the research question and recommends future work, and Chapter 7 reflects on the research process itself.

Chapter 2

Literature Review

2.1 Biometric presentation attacks and why they matter

Facial biometric systems are attractive because they are convenient and contactless, but this convenience creates an attack surface that fingerprint or password authentication does not share: a face can be captured from a distance, reproduced cheaply, and presented to a sensor without the legitimate user's cooperation. A *presentation attack* is any such attempt to subvert a biometric system by presenting an artefact to the capture device rather than the genuine biometric characteristic [1]. Common attack instruments range from a simple printed photograph, through a face replayed on a phone or laptop screen, to physical masks of varying sophistication, up to custom-fitted silicone masks capable of defeating even deep face-recognition models at rates an order of magnitude higher than their nominal false-match rate [9].

Presentation attack detection (PAD) is the countermeasure that sits ahead of recognition and answers a narrower question than recognition does: not *who* is this, but *is a live person present at all*. This distinction matters for the present project because it bounds what the sensor needs to resolve. A system that only needs to detect liveness, rather than identity, does not need identity-revealing resolution – a claim this dissertation tests directly. Ramachandra and Busch [10] survey PAD methods for consumer RGB cameras and categorise them by the cue exploited: texture and image-quality statistics, motion and challenge-response, and remote photoplethysmography, among others. A recurring theme across this literature is that RGB-only defences are locked in an arms race with increasingly realistic replay attacks, which motivates the search for cues, such as thermal emission, that a screen or a printed sheet cannot reproduce regardless of image quality [11].

2.2 Evaluating PAD: the ISO/IEC 30107-3 framework

Comparability across PAD studies depends on a shared evaluation vocabulary, which is provided by ISO/IEC 30107-3, *Biometric presentation attack detection – Part 3: Testing*

and reporting [8]. The standard defines the Attack Presentation Classification Error Rate (APCER), the proportion of attack presentations incorrectly classified as bona fide, and the Bona Fide Presentation Classification Error Rate (BPCER), the proportion of genuine presentations incorrectly rejected. Their mean, the Average Classification Error Rate (ACER), is widely reported as a single summary figure, although the standard itself is explicit that APCER should where possible be considered per attack type rather than pooled, since a detector can average well while failing completely against one attack family – a distinction this dissertation relies on directly in Chapter 5, where per-family APCER reveals that temperature and shape cues fail on almost entirely different attacks. This project reports APCER, BPCER and ACER throughout, consistent with the standard, and reports per-family breakdowns wherever pooling would obscure a meaningful failure mode.

2.3 PAD beyond the visible spectrum

The limitation of visible-spectrum PAD follows from what a camera measures. A conventional camera records light reflected from a scene, and a printed photograph or a replayed video is itself a recording of reflected light. An attack instrument of sufficient quality is therefore not an approximation of the original signal but a faithful reproduction of it, and defences based on texture, image quality or reflectance are reduced to detecting artefacts of the reproduction process rather than any property the artefact fundamentally lacks. As reproduction quality improves those artefacts diminish, which is the arms race Ramachandra and Busch [10] describe.

Thermal imaging changes what is being measured. A thermal sensor records long-wave infrared radiation emitted by an object as a consequence of its own temperature, and in the case of a face that temperature is maintained by metabolism and perfusion. A printed photograph does not emit this radiation because there is nothing generating it; the property is not difficult for the artefact to reproduce so much as absent from it entirely. This is a categorical difference rather than a difference of degree, and it is why thermal sensing appears repeatedly in the PAD literature — Bhattacharjee et al. [9] propose a low-cost thermal camera specifically as a defence against silicone masks that defeat visible-spectrum recognition. Thermal facial imagery is additionally largely invariant to ambient illumination and carries vascular structure that a printed or displayed artefact does not reproduce [12].

The naive form of this argument, that spoofs are cold, is nonetheless insufficient, and the literature is clear that a motivated attacker will heat the artefact. Kowalski et al. [3] include artificially warmed masks among their attack instruments for exactly this reason, and evaluate presentation attack detection in thermal infrared against printed photographs, displayed photographs, 3D-printed masks, and latex and silicone masks. Their heated-mask

condition is the full-resolution analogue of the combination attack constructed for this project (Section 3.4), in which a printed photograph is warmed from behind specifically to defeat the temperature cue. The evidence from the present project agrees with theirs: a printed photograph warmed to 35.5°C sits inside the live temperature band, and 19% of all attack frames in this dataset fall within it (Section 5.1). Absolute temperature is a necessary cue but not a sufficient one, and the fact that a full-resolution study and a 192-pixel one independently converge on “warm the artefact” as the most dangerous attack is a useful point of agreement.

What survives heating is spatial structure. A live face is not uniformly warm: it is warmed unevenly by underlying vasculature and by the varying thickness of tissue, producing a large, structured region with internal variation. An artefact warmed from behind is heated by a source with its own geometry, and reproduces that geometry rather than a face’s. Distinguishing the two therefore becomes a question of the shape and internal structure of the warm region rather than of its temperature, which is the cue this project builds on.

Structure is not unfalsifiable either. A sufficiently determined attacker could construct an artefact with both the right temperature and a plausible heat distribution, and doing so is a matter of effort rather than of principle. What is harder to reproduce is behaviour over time: a living face fluctuates and dissipates heat continuously as blood flow and ambient conditions change, whereas a passively heated artefact follows the simpler thermal profile of whatever is warming it. That temporal dimension is not exploited in this project, which classifies each frame independently, and it is identified in Chapter 6 as a promising direction for further work.

George et al. [2] introduce the Wide Multi-Channel presentation Attack (WMCA) database, combining colour, depth, near-infrared and thermal channels, and report an ACER of 0.3% using a multi-channel convolutional network. This is the strongest published thermal-adjacent PAD result and the benchmark this project must be read against, but the comparison is asymmetric in an important way: WMCA uses a full-resolution thermal camera in combination with three further imaging channels, whereas this project uses a single 192-pixel thermal channel and nothing else. The gap between 0.3% and this project’s 9.2% ACER is therefore not simply a measure of algorithmic quality; it is substantially a measure of what 192 pixels of one modality can achieve against roughly two orders of magnitude more information across four modalities. Heusch et al. [13] report a related result using shortwave-infrared imaging instead of thermal, which together with George and Marcel [14]’s systematic comparison across channels helps justify, from the literature, why thermal specifically was chosen for this project: it is the channel most directly tied to a physiological signal – heat – that an inert artefact cannot reproduce without active heating, as opposed to depth or NIR cues that a sufficiently good 3D-printed or curved artefact can approximate.

2.4 Low-resolution thermal sensing and privacy

A separate strand of research establishes that thermal arrays far coarser than a conventional thermal camera are nonetheless useful for human-centred sensing, precisely because coarseness removes the privacy liability of the sensor. This strand includes several works by this project’s supervisor, Dr Abdallah Naser, whose prior research investigates low-resolution thermal sensor arrays for privacy-preserving human monitoring. Naser et al. [4] demonstrate accurate discrete distance estimation from a low-resolution thermal array; this project’s own distance study (Section 5.1.1), which finds peak facial temperature stable to within 0.43°C across the 30–50 cm working range, is a direct descendant of that finding and of the calibration procedure described in Naser et al. [5], whose emissivity and ambient-temperature settings this project matched for comparability. Naser et al. [6] extend this line of work to multi-sensor fusion, and Naser et al. [7] address sensor placement and human segmentation on low-resolution arrays – directly relevant to the torso-confound problem this project encounters in its own cropping experiment (Section 5.8), since segmentation at 16×12 resolution is exactly the problem that experiment attempts, and fails, to solve cheaply.

The privacy argument itself is made explicitly by Newaz and Hanada [15], who argue that a sufficiently low-resolution infrared array preserves the ability to detect and characterise human presence while removing the capacity for identification – the same claim this dissertation makes about its 192-pixel frames, and one it can now support with a concrete number: at 16×12 , no facial feature large enough for a human or a recognition model to identify an individual survives, a claim substantiated visually in Chapter 3. More recent work, Zhang and Wu [16], pushes low-resolution thermal sensing further by extracting coarse depth and ranging information from a thermal array alone, describing an “excellent resolution-privacy balance”; this is a useful reference point for future work, since a second low-resolution sensor could plausibly recover some of the shape information this project’s single-sensor cropping experiment could not (Section 5.8). At the lower extreme, Yin et al. [17] demonstrate human activity recognition from an 8×8 Grid-EYE array, and Vandersteegen et al. [18] demonstrate person detection running on a low-cost microcontroller, supporting this project’s own deployment feasibility claim (Section 5.11) that a millisecond-scale, sub-kilobyte detector is realistic to run on commodity embedded hardware. None of this body of work, however, addresses presentation attack detection specifically – a gap made explicit in Section 2.7.

2.5 Generalisation: unseen attacks and unseen users

A deployed PAD system will eventually face an attack instrument its designer did not anticipate. Arashloo et al. [19] make the founding argument for treating PAD as an anomaly-detection problem rather than a two-class classification problem: a model trained

only on genuine presentations, with no attack examples at all, is by construction unable to overfit to the specific attacks available during development, and therefore has a principled claim to generalise to novel attacks. Nikisins et al. [20] evaluate this claim directly, under a protocol – one-class models scored against attack families withheld from training – that this project’s own unseen-attack experiment (Section 5.6) mirrors closely, and with a similar conclusion: one-class models degrade far more gracefully against unseen attacks than supervised classifiers do. This project’s result is a stronger version of the same finding: the fused one-class detector achieves a mean ACER of 0.091 against attack families withheld entirely from training, while supervised classifiers trained on the remaining three families reach a mean ACER between 0.34 and 0.53 against the withheld family, with some individual configurations performing worse than chance.

A second, orthogonal generalisation question concerns the user rather than the attack: does a detector trained on one person’s thermal signature transfer to another person? Fatemifar et al. [21] address a closely related question with *client-specific* anomaly detection, arguing that per-person models outperform a single pooled model precisely because individual physiological baselines differ – a finding this project’s own cross-participant experiment (Section 5.7) independently confirms and quantifies: participants differ in mean peak facial temperature by 0.64°C (Cohen’s $d = 2.04$), which is sufficient on its own to explain why a detector fitted to one participant’s thermal centre rejects a large fraction of a second, unseen participant’s genuine presentations. Where Fatemifar et al. [21] treat this as grounds for a client-specific design, this project reports it as a limitation to be resolved with a larger and more diverse participant pool, since a strictly client-specific model requires enrolment data from every future user, which is a materially different and more burdensome deployment model.

2.6 The physiology of facial temperature

The live temperature band this project measures — $35.35 \pm 0.45^{\circ}\text{C}$ at the frame peak — requires grounding in the medical literature rather than resting on this project’s own data alone.

The closest independent comparison is Vardasca et al. [22], who recorded the inner canthi of fourteen healthy participants with a calibrated FLIR E60sc thermal camera under controlled conditions, following the Glamorgan protocol, and report a mean of $35.4 \pm 0.5^{\circ}\text{C}$. That figure agrees with the present project’s live band to within 0.05°C in the mean and 0.05°C in the standard deviation, despite being obtained with a full-resolution medical thermal camera rather than a 16×12 array, at a working distance of 100 cm rather than 30–50 cm, and at a specific anatomical region of interest rather than as a frame maximum. The convergence is useful precisely because the two measurements were made in such different ways: it indicates that the band this project fits is a property of human

facial temperature rather than an artefact of a low-resolution sensor or of a particular capture protocol.

The same study also provides an independent check on the distance behaviour reported in Section 5.1.1. Vardasca et al. [22] find that inner-canthen temperature varies by only 0.1°C across distances from 80 to 120 cm, and that measurement error grows outside that range. The present project measures a slope of -0.0213°C per centimetre over 30–50 cm, equivalent to a total change of 0.43°C across the working range. Both studies therefore find facial temperature to be stable to within a few tenths of a degree over their respective operating distances, which is the property a deployed liveness detector depends on: the system must not need to know how far away the user is standing.

Not all published figures agree so closely, and the difference is instructive. Micheletti et al. [23] report regional facial baselines across twenty healthy subjects that sit lower still: $34.3 \pm 0.7^{\circ}\text{C}$ at the inner canthus, and 33.4 ± 0.8 and $33.3 \pm 0.8^{\circ}\text{C}$ at the forehead and cornea respectively – the inner-canthen figure roughly one degree below this project’s, and the forehead and cornea roughly two. The most likely explanation is measurement site and protocol rather than disagreement about physiology: fever-screening guidance identifies the inner canthus as the most thermally stable site precisely because it is a small region, and a mean taken over a tightly drawn region of interest will sit below a frame maximum taken over the whole face [24]. That distinction matters for how the present results should be read. This project reports the hottest pixel in the frame, which is a deliberately simple and reproducible statistic at 16×12 resolution, but it is not the same quantity as a region-of-interest mean, and the two should not be compared without noting the difference.

2.7 Summary and the research gap

Three observations from the literature reviewed above combine to define the gap this project addresses. First, thermal imaging is an established and effective PAD cue, but only at full sensor resolution: both Kowalski et al. [3] and George et al. [2] demonstrate strong thermal PAD using cameras that cost from hundreds to thousands of pounds and that themselves capture identifiable thermal faces. Second, low-resolution thermal arrays are an established platform for privacy-preserving human sensing – distance estimation, occupancy, segmentation and activity recognition – as the supervisor’s own body of work and related studies demonstrate [4, 5, 6, 7, 15, 16], but this body of work is directed at human-sensing tasks such as distance estimation, occupancy and activity recognition rather than at presentation attack detection, and the one study that does apply a low-resolution thermal array to presentation attack detection pairs it with an RGB camera (Section 2.7.1). Third, whether liveness detection is possible at a resolution low enough to be privacy-

preserving by construction, and cheap enough to deploy widely, is consequently an open question with no established public dataset to answer it.

This dissertation answers that question, and the answer is qualified rather than unconditional: within a known participant, a low-resolution thermal array supports liveness detection competitive with – and in this project’s results, better than – deep models applied to the same data (ACER 0.092 against 0.121–0.251 for eleven pretrained and custom networks, Chapter 5). Across an unseen participant, it does not yet generalise reliably (ACER 0.167 at best). That second half of the answer is, as Section 2.5 anticipates, the more scientifically interesting of the two.

2.7.1 Systematic novelty search

Because the gap claimed above is partly a claim about absence, a systematic search was carried out to test it. The query ("thermal array" OR "low-resolution thermal") AND ("presentation attack" OR "liveness" OR "anti-spoofing") was run across Scopus, Google Scholar and the USPTO full-text database in August 2026, together with hardware-specific variants naming the MLX90640 and MLX90641 devices.

The search did not return an empty result, and the gap must therefore be stated more narrowly than the three observations above imply. Two items are directly relevant.

Imam [25] reports face presentation attack detection using a low-cost thermal sensor array, collecting approximately 46,000 thermal images and classifying them with AlexNet and ResNet-50, reporting accuracies of 0.9798 and 0.9691 respectively. The overlap with the present project is substantial and must be acknowledged plainly: that work was carried out at Nottingham Trent University, uses the same Melexis EVB90640-41 evaluation board, and includes among its attack classes a category described as *picture attached with hot objects*, which is functionally the same construction as the combination attack used here. A United States patent covering thermal-based presentation attack detection for biometric systems (US 12,277,803, priority 2021) further confirms that the general idea is neither unexplored nor unclaimed.

Four differences nonetheless distinguish the present work, and the contribution should be framed in terms of these rather than in terms of priority. First, and most importantly, Imam [25] fuses the thermal array with a conventional RGB camera. That design choice is reasonable for maximising accuracy, but it forfeits the privacy argument entirely: a system containing an RGB camera captures an identifiable face regardless of what the thermal channel resolves. The present project uses the thermal channel alone, which is what makes the privacy-preserving claim defensible. Second, that work reports accuracy, precision, recall and F1, whereas ISO/IEC 30107-3 [8] specifies APCER, BPCER and ACER for presentation attack detection; the two sets of figures are therefore not directly comparable, and accuracy in particular can conceal a detector that fails completely against one attack family. Third, the present project uses a one-class formulation requiring no attack data

during training, and demonstrates generalisation to attack families withheld entirely from training, whereas a supervised CNN trained on a fixed set of attack classes has no such guarantee. Fourth, the present project reports cross-participant generalisation explicitly, and finds that it fails; no equivalent analysis appears in the prior work.

One methodological caution is worth recording, and is offered as an observation rather than a criticism, since the published description is not detailed enough to settle the question. Where a large image count is drawn from a comparatively small number of continuous recordings, a randomly partitioned train/test split will place near-duplicate frames on both sides and inflate the reported score. The present project encountered exactly this failure and documents it in Section 3.6, where a random frame-level split produced an ACER of 0.000 that collapsed to chance under session-grouped validation. Any comparison between the accuracies reported in prior work and the ACER values reported here should be read with that difference in validation protocol in mind.

Chapter 3

Methodology

3.1 Sensor and software setup

The sensing platform is a Melexis EVB90640-41 evaluation board [26], configured through the MlxCIRT 90640 software. The logging separator was set to a semicolon so that exported CSV files match the format of the supervisor's own sample recordings, and log output was directed to a named file per scenario so that recordings save to the correct folder without manual renaming. Settings were checked against the supervisor's sample file and matched exactly: emissivity 1.000, frame rate 2 fps, ADC resolution 2, an interleaved reading pattern, and the de-interlace, IIR and TGC filters enabled. Identical settings were used for every recording in the dataset, including sessions recorded weeks apart, so that sessions are directly comparable. Figure 3.1 shows the assembled setup in the laboratory.



Figure 3.1: The capture setup, photographed in the laboratory. (a) the working distance from the subject’s face to the sensor is set with a tape measure, here at the 50 cm end of the 30–50 cm range; (b) a live bona-fide capture, with the thermal field displayed by the MLxCIRT software showing the characteristic warm facial region; (c) a printed-photograph attack presented to the sensor, with the target propped and the operator clear of the field of view; (d) the EVB90640-41 evaluation board connected to the capture laptop. The sensor is hand-held for these photographs only; all dataset recordings were made with the sensor and the target both stationary.

Establishing which sensor is fitted

The evaluation board is supplied as a dual unit, and its silkscreen lists both the MLX90640 (32×24) and the MLX90641 (16×12) [26]; the printed label therefore does not identify which die is actually populated. Because every frame-level feature in this work depends on the frame geometry, this was established rather than assumed.

Recordings made for this project log 192 values per frame, while the sample file supplied by the supervisor contains 768. Three possible explanations were tested. Interpolation was set to $\times 1$, $\times 2$ and $\times 4$ in turn, and confirmed to affect only the on-screen display rather than the logged data. The configuration registers of both files were compared, which showed that the supervisor’s recording used the Chess reading pattern while this project’s used Interleaved; switching to Chess and re-recording left the logged count unchanged at 192. No combination of software settings produced 768 values.

The conclusion is that the board used for this project is populated with the MLX90641, whose array is 16×12 and therefore yields exactly the 192 values observed [27]. All

work reported in this dissertation uses that geometry. The distinction matters beyond bookkeeping: at 16×12 a face at the working distance spans roughly forty pixels, which is what makes the privacy argument in Section 2.4 defensible, and it is also why the face-cropping experiment in Section 5.8 has so few pixels to work with. The analysis pipeline detects frame geometry automatically and handles either size, so the same code would run unmodified on a 32×24 board.

3.2 Capture protocol

The capture procedure was: clear the log, open the live view to begin recording, present the target for approximately 60 seconds at a distance of 30–50 cm, then minimise the window to stop. At 2 fps this yields approximately 120 frames per scenario, each saved to its own CSV file.

Recordings discarded, and why

The first attempt at the printed-photograph attack was unusable: the peak temperature was 34.8°C against 35.4°C for a live face – a difference of only 0.6°C – with 23% of the frame reading hot. A room-temperature sheet of paper should read close to ambient; reviewing the frames revealed warm hand- and arm-shaped regions moving through the scene, indicating the sensor was mostly seeing the operator rather than the paper. A second attempt, holding the photograph differently, improved the main area but left a bright band along the bottom edge of every frame, 1.8°C hotter than the rest, where the equivalent rows in a live recording were 0.1°C cooler. Left uncorrected, this artefact risked teaching a classifier to associate a hot bottom strip with an attack, rather than anything related to thermal physics. On a third attempt the photograph was propped so that nothing touched it, and the operator stepped entirely out of the sensor’s field of view before recording; this produced 0% hot pixels and a peak only 2.4°C above ambient, and this discipline – prop the target, stay out of frame, confirm the live view is dark before starting – was adopted for every subsequent recording. This failure was not visible from summary statistics alone and was caught only by inspecting individual frames and comparing regions of the image, which motivated the frame-level integrity checks described in Section 3.5.

3.3 Ethical considerations

Ethical approval for this project was obtained prior to data collection; the approval and the data-protection measures applied are documented in Appendix B. All participants were given the Participant Information Sheet and provided informed consent using the approved Participant Consent Form (Appendix B) before any recording took place. In line with the approved application, the dataset contains no names, photographs, video or audio; frames

are labelled by participant number only, and the sensor’s own resolution means no frame is capable of revealing a participant’s identity, which was itself part of the ethical justification for the project’s privacy-preserving framing.

Participant 1 is the researcher: all bona fide recordings from the first phase are of the author, stated here rather than left to be inferred. Two consequences follow. The researcher was not blind to the class of any recording during capture, which is why every recording decision – most importantly the rejection of a mis-framed session in Section 3.4 – is reported with the evidence it rested on. More significantly, a detector fitted largely on one person may be learning that person rather than liveness; that is measured directly in Section 5.7, which establishes the project’s central limitation.

3.4 The dataset

The completed dataset contains 1,850 frames across 15 recording sessions: 981 bona fide frames from two participants and 869 attack frames spanning four attack families, summarised in Table 3.1. Only the 16×12 thermal arrays are stored; no images, names or audio are retained, consistent with the approved ethics application.

Table 3.1: Composition of the completed dataset.

Class	Type	Sessions	Frames	Detail
Bona fide	–	8	981	2 participants; 30/40/50 cm and natural pose
Print attack	print	2	246	plain photograph, warmed photograph
Screen attack	screen	2	245	phone, laptop
Heat decoy	heat decoy	2	256	warm water cup, water balloon
Combination	combination	1	122	printed photograph, heat source behind
Total		15	1,850	

Figure 3.2 shows the frame count for each session.

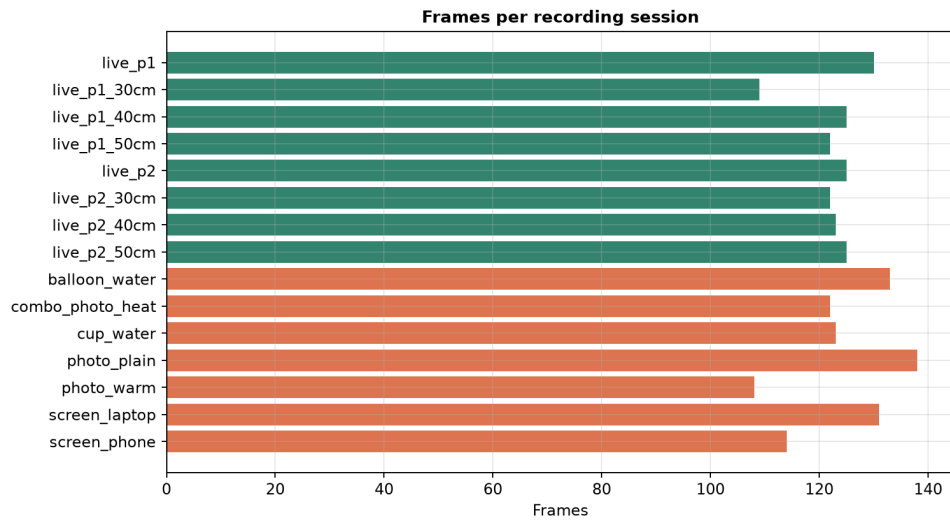


Figure 3.2: Frames per recording session. Sessions are deliberately similar in length so that no single recording dominates the dataset or the grouped cross-validation folds.

Recording the second participant

Recording a second participant was identified, after the first phase of data collection, as the single most important gap in the dataset, since every prior recording and every conclusion drawn from it concerned one individual only. A second participant was recruited, given the Participant Information Sheet, and recorded as participant 2 across four sessions following the same protocol used for participant 1: three at fixed distances of 30, 40 and 50 cm, and one natural-pose session in which the participant was asked to turn and tilt their head, talk and smile, so that the recording would contain realistic movement rather than a static pose.

Each recording was checked numerically before the next was made, rather than waiting until the session ended, and this practice caught a framing fault immediately: the first 40 cm recording gave a peak of 34.01°C , a drop of 1.37°C from the 30 cm session, where participant 1's own recordings had dropped only 0.20°C over the same step. A fall of that size was not credible as a distance effect. Computing the horizontal centre of mass of the warm region located it at column 9.1 of 15, where every other valid recording sits between 6.8 and 7.5, with the left third of the frame flat at ambient: the participant had moved back and sideways simultaneously, so part of the face had left the sensor's field of view and the hottest region was being clipped. The recording was repeated after confirming the live view was centred, giving 123 frames with a peak of 35.07°C – a 0.31°C step from the 30 cm session, consistent with participant 1's data. Only the corrected recording was retained, rejected for a documented framing fault identified before any model was run on it, not because of the score it produced. Checking each recording numerically while the participant is still present costs a minute of the session and avoids collecting a distance curve that disagrees with the reference data for a reason that could not be diagnosed after the participant had left.

Why there are no mask attacks

The original proposal listed masks, and the literature treats them as the most serious attack instrument [9, 3]. None appears here, for reasons of cost: the custom silicone masks used in the study establishing their severity were made to order at roughly USD 4,000 each [9]. A cheap latex substitute was rejected as uninformative, since thermally an unworn low-cost mask is an inert sheet near room temperature – the printed-photograph case already covered – while a worn one is a thin layer over a live face, which the heat-decoy and combination attacks bracket from either side. The combination attack described next was built instead, and is the harder test. The absence of a genuine mask attack remains a limitation, recorded as one in Chapter 6.

The combination attack

The combination attack was designed to defeat the temperature cue directly. The printed photograph used for the plain-print attack was mounted with a heat source behind it, warming the paper’s surface into the live temperature range rather than the 37–39°C reached by the heat decoys, and propped at approximately 40 cm with the operator out of frame. It succeeded as intended: across 122 frames the peak is 36.05°C with a standard deviation of 0.55°C, and 56% of frames fall inside the live temperature band, making it by a wide margin the most convincing spoof in the dataset – the plain photograph sits at 31.63°C and the water balloon at 39.08°C, both far outside the band, while the combination attack straddles it. A detector relying on temperature alone would be close to guessing against this attack, which is precisely why it was constructed.

3.5 Data quality

Before any modelling, integrity checks were run for impossible temperatures, frozen frames, within-session drift, and short sessions. Two corrupt frames were found across the project, both reading over 200°C and both the first frame of a recording, consistent with an I²C read glitch; both were removed, since a single such frame distorts any feature computed from the frame maximum. The final dataset contains none.

Drift analysis behaved as physically expected: the heated spoofs warm measurably during recording – the combination attack drifts +0.92°C and the warmed photograph +1.41°C across their minute – while the fixed-distance live sessions are stable to within 0.30°C (five of six within 0.12°C). The one live session with noticeable drift, the natural-pose recording for participant 2, drifts –0.80°C, consistent with the participant moving their head as instructed rather than indicating a fault.

3.6 A methodological error found and corrected

The first working analysis pipeline split frames randomly into training and test sets and produced an ACER of 0.000 – apparently perfect detection. This result was incorrect. At 2 fps, consecutive frames from a single 60-second recording are near-identical, so a random split places near-copies of the same frame on both sides of the split; the model was not learning to recognise liveness, but to recognise individual recordings it had already seen in a slightly different guise.

The splitting procedure was rewritten so that whole recording sessions are assigned to a single split, and no session ever appears on both sides of a train/test boundary. Under this session-grouped validation, the same model configuration scored approximately 0.51, indistinguishable from chance. Every result reported in this dissertation from Chapter 5 onward uses session-grouped cross-validation, and the analysis code enforces this centrally so that it cannot be bypassed by a later change to any individual script. An honest result of 0.09 is preferable to a meaningless 0.00, and this experience motivated the sensitivity checks reported alongside single headline numbers throughout Chapter 5, rather than reporting point estimates alone.

3.7 Analysis pipeline

With the dataset complete, the analysis code was rewritten as a single coherent suite of eighteen numbered scripts and two shared modules, replacing an earlier version that had grown incrementally as data arrived and that assumed, in places, a single participant and three attack families. Three rules are enforced centrally throughout the suite: validation is always grouped by session (Section 3.6); all random seeds are fixed, since the elliptic-envelope detector’s randomised minimum-covariance estimate was found to move ACER by approximately 0.005 between otherwise identical runs; and no parameter is chosen by reading the evaluation metric, a discipline applied explicitly in the cropping-window experiment of Section 5.8. The full suite runs from a single command in approximately fifty minutes and writes every result to CSV, so that no figure quoted in this dissertation was transcribed by hand and the full analysis can be reproduced or audited from the project repository (Appendix C).

Chapter 4

Design and Implementation

This chapter describes how the detector was designed and built: the features extracted from each thermal frame, the families of model compared against one another, and the design of the final fused detector. Chapter 5 reports the resulting performance in full.

4.1 Design principle: work upward in complexity

Every model compared in this project was required to earn its place against a simpler alternative. Given a 192-value input and, at most, fifteen recording sessions to validate against, a large model is not free: it is a specific hypothesis that more capacity is needed than the data justifies, and that hypothesis has to be tested rather than assumed. Detectors were therefore developed in increasing order of complexity – single hand-set thresholds, classical supervised classifiers, one-class anomaly detectors, a hand-designed two-cue fusion rule, and finally deep neural networks – with each tier required to beat the previous one to be considered a genuine improvement rather than added complexity for its own sake.

4.2 Feature engineering

Two independent families of feature were extracted from every frame, motivated by the two physical properties a live face is expected to exhibit and an inert artefact is not: it is warm, and it is warm in a structured, face-shaped pattern.

4.2.1 Temperature features

The primary temperature feature is the peak pixel temperature in the frame. A *live band* was defined as the mean peak temperature of bona fide frames ± 2 standard deviations, giving an acceptance interval of 34.46–36.24°C (Section 5.1). This is deliberately the

simplest possible temperature feature; the intention was to establish how far a threshold this simple could be pushed before adding complexity, per the design principle above.

4.2.2 Topology features

Twelve hand-designed shape features were computed from the spatial pattern of the warm region in each frame, covering five aspects of its shape: size (warm-pixel area as a fraction of the frame), shape (aspect ratio, height, width of the bounding region), compactness (the ratio of warm area to its bounding perimeter), internal variation (the standard deviation and range of temperature within the warm region), and asymmetry (left-right and top-bottom imbalance of the warm region about its centroid). These features were chosen to be simple, explainable, and computable in closed form from a 16×12 array without a trained detector of any kind – a design constraint discussed further in Section 4.7 below, where a trained face detector was deliberately avoided for the same reason.

4.3 Baseline thresholds

Before any learned model was built, single hand-set thresholds on the strongest individual temperature and topology features were evaluated, to establish a floor that any added complexity would have to beat. This step is reported in full in Section 5.4; the relevant design point here is that the temperature-band threshold is treated throughout this dissertation as the minimum acceptable bar for any subsequent model, precisely because it encodes a physical fact about human skin rather than a pattern learned from a small sample.

4.4 Classical supervised classifiers

Fifteen conventional classifiers – spanning linear and quadratic discriminant analysis, support vector machines [28], random forests [29], and gradient-boosted trees [30], implemented using scikit-learn [31] – were trained on each of eight feature representations (temperature alone, topology alone, raw normalised pixels, and combinations of these), giving 120 model-and-feature-set combinations, each fitted over five session-grouped folds for 600 fits in total. All were validated using the session-grouped cross-validation established in Section 3.6.

4.5 One-class anomaly detection

A deployed liveness detector cannot realistically hold training examples of every attack it will eventually face, since new attack instruments are invented after a system is deployed.

One-class models address this by learning only what a live face looks like, and rejecting anything sufficiently unlike it, so that unseen attacks are handled by construction rather than by chance – the argument developed from Arashloo et al. [19] and Nikisins et al. [20] in Section 2.5. Five anomaly detectors were fitted on bona fide frames only: an isolation forest [32], an elliptic envelope using the minimum covariance determinant estimator [33], a one-class support vector machine [34], a local outlier factor detector [35], and a Gaussian mixture density model, all implemented in scikit-learn [31] and each evaluated on the same eight feature sets as the classical classifiers above, giving forty one-class configurations in total.

4.6 The fused two-cue detector

The per-family breakdown of the temperature and topology baselines (reported in full in Section 5.5) revealed that the two cues fail on almost entirely different attacks: temperature rejects heat decoys and screens reliably but is close to guessing against the combination attack, while topology rejects the combination attack and printed photographs reliably but is weak against heat decoys, whose smooth warm region resembles a face’s shape statistics well enough to pass. This complementary failure pattern motivated a deliberately simple fusion rule: reject a presentation if its peak temperature falls outside the live band, *or* if its topology is judged anomalous by the one-class shape model. Both components are fitted on training-split bona fide frames only, so the fused detector uses no attack data during training and remains a one-class method in the sense of Section 4.5; validation remains grouped by session, so neither component ever scores a session it has seen. This is presented in Chapter 5 as the project’s principal result.

4.7 Face-region cropping

The sensor’s wide field of view includes the neck, shoulders and sometimes clothing at the 30–50 cm working distance, so any area- or shape-based feature is partly a measurement of clothing rather than face. A simple, explainable cropping step was designed to address this: locate ambient temperature, threshold two degrees above it, take the topmost warm row as the top of the head, and take a fixed head-sized window downward centred on the warm region. A trained face detector was deliberately not used, so that the cropping step remains as easy to justify and reproduce as the rest of the pipeline; the one free parameter is the size of the crop window, and its sensitivity was tested explicitly rather than tuned against the evaluation metric, consistent with the discipline established in Section 3.7. The outcome of this experiment, which did not improve on the uncropped detector, is reported in full in Section 5.8.

4.8 Neural networks and transfer learning

Four custom convolutional networks were designed and trained from scratch on the raw 16×12 frames, alongside a convolutional autoencoder used as a reconstruction-error anomaly detector. In addition, eleven pretrained ImageNet backbones were used as frozen feature extractors with a logistic-regression head trained on top, spanning a wide range of architectural families and parameter counts: GoogLeNet [36], ResNet18 and ResNet50 [37], DenseNet121 [38], SqueezeNet1.1 [39], VGG11 [40], EfficientNet-B0 [41], MobileNetV3 [42], ConvNeXt-Tiny [43], and the Vision Transformer [44] and Swin Transformer [45], all pretrained on ImageNet [46]. All networks were implemented in PyTorch [47] and evaluated under the same session-grouped validation as every other model in this dissertation. The purpose of this comparison was not an expectation that deep models would outperform the hand-designed detector – with at most fifteen independent recording sessions, there is little reason to expect a large model to learn a better representation than two physically motivated features – but to test that expectation rather than assert it, consistent with the design principle of Section 4.1. The results, reported in Section 5.9, support the hand-designed design and are treated in this dissertation as a finding rather than a negative result to be minimised.

4.9 Deployment feasibility

Because the fused detector’s trained state is a mean vector and covariance matrix over twelve features, plus the two scalar bounds of the temperature band – approximately 180 numbers in total – its deployment cost was assessed directly rather than assumed. Feature extraction and scoring time were measured per frame, and temporal aggregation over a short window of consecutive frames was evaluated as a way to trade decision latency for accuracy. This design consideration, and the resulting cost and latency figures, are reported in Section 5.11.

Chapter 5

Results and Discussion

Every configuration evaluated is reproduced in Appendix D, and the per-session composition of the dataset in Appendix A; the tables in this chapter report the best rows from those grids, and the complete output is given there so that the selection can be audited.

Every number in this chapter is produced by the analysis suite described in Section 3.7 and traces to a CSV file in `results/tables/`; none was transcribed by hand. Figures are generated by the same suite from those tables.

5.1 Temperature: necessary but not sufficient

Across all 981 bona fide frames, peak face temperature is $35.35 \pm 0.45^\circ\text{C}$. Two standard deviations gives a live band of $34.46\text{--}36.24^\circ\text{C}$, the acceptance interval used throughout this chapter.

5.1.1 Distance

Peak temperature declines only slightly with distance: 35.65°C at 30 cm, 35.41°C at 40 cm, and 35.23°C at 50 cm, pooled across both participants. A linear fit gives a slope of -0.0213°C per centimetre ($r = -0.997$, $p = 0.050$), predicting a total change of just 0.43°C across the working range. For practical purposes, face temperature is stable over 30–50 cm, so a deployed system does not need to estimate user distance (Figure 5.1).

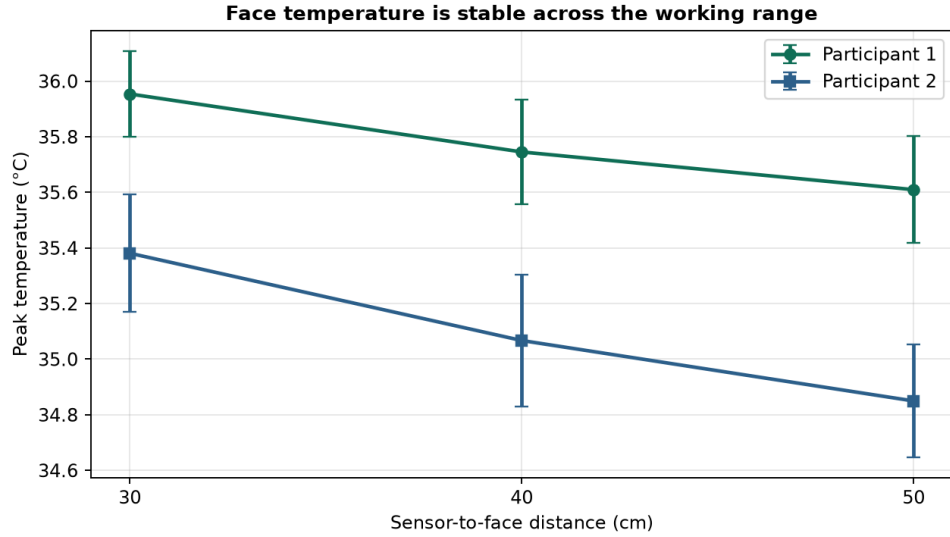


Figure 5.1: Peak face temperature against distance for both participants. Participant 2 sits about 0.6°C cooler throughout but shows the same shallow decline, so the stability of face temperature over the working range is not specific to one individual.

5.1.2 Why temperature alone fails

Live and attack temperature distributions overlap. Live frames average $35.35 \pm 0.45^\circ\text{C}$ and attack frames $35.72 \pm 2.50^\circ\text{C}$; the difference is statistically detectable (Welch $p = 2.6 \times 10^{-5}$) but the effect size is negligible (Cohen’s $d = -0.20$), and the attack standard deviation is more than five times larger, spanning everything from a cold photograph to a hot water balloon. In consequence, 19.4% of attack frames fall inside the live band – and, crucially, the attacks that survive are exactly the ones designed to: 93% of warmed-photograph frames and 56% of combination-attack frames fall inside the band, while the plain photograph, the phone screen and both heat decoys sit almost entirely outside it (Figure 5.2).

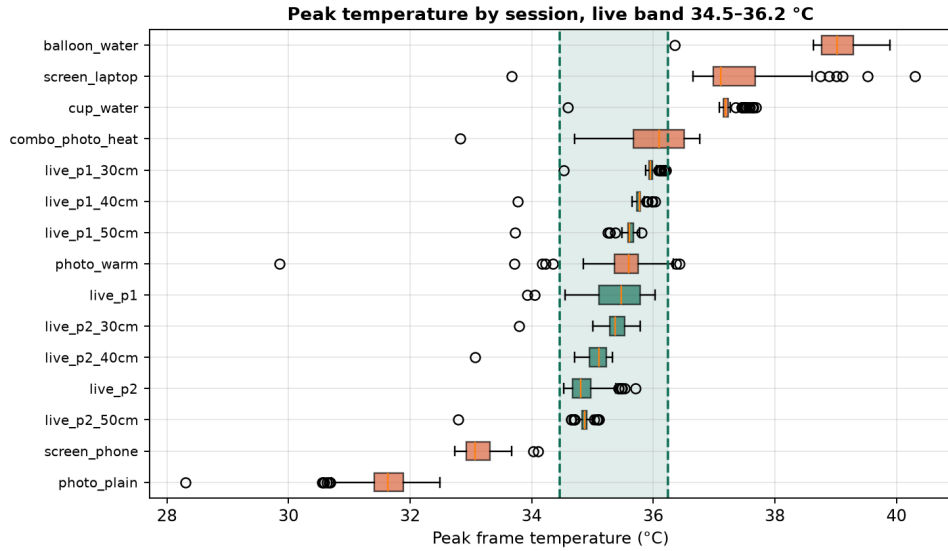


Figure 5.2: Peak temperature by session, with the live band (34.46–36.24°C) shaded. Most spoofs fall clearly outside the band, but the combination attack and the warmed photograph overlap it – which is precisely why a temperature threshold alone is not a sufficient defence.

5.1.3 A difference between participants

The two participants differ by 0.64°C in mean peak temperature (Cohen’s $d = 2.04$, a very large effect). Both fall inside the pooled band on 99% of frames each, so the band itself transfers between them, but this difference is the clearest single explanation for the cross-participant failure reported in Section 5.7: a model fitted tightly around one person’s thermal signature is fitted around the wrong centre for another.

5.2 Topology: where the signal actually is

Twelve hand-designed shape features (Section 4.2) were evaluated individually by AUC. The strongest single feature, warm-region height, separates the classes at AUC 0.901, and three features exceed AUC 0.77. The per-family breakdown is more informative than the pooled figure: against live faces, the combination attack is separable by the range of temperature within the warm region at AUC 0.999, printed photographs by warm-region height at 0.981, screens at 0.912, and heat decoys at 0.906.

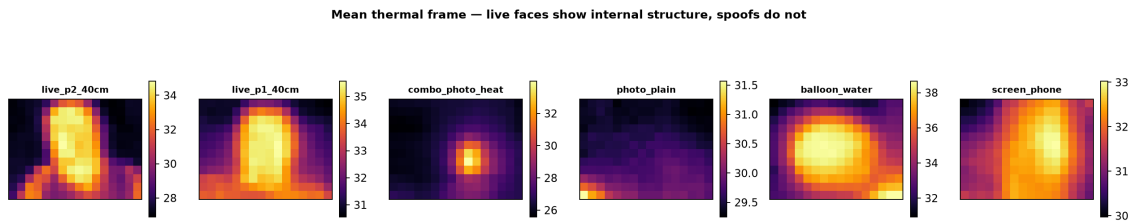


Figure 5.3: Mean thermal frame for six scenarios. A live face shows a large, structured warm region with internal variation. The combination attack is a small compact blob, the unheated photograph is near ambient, and the water balloon is large but smooth and featureless.

The combination attack illustrates the underlying argument well: its warm region occupies only 0.20 of the frame, against 0.46–0.66 for a live face, and forms a single compact blob radiating from one point, with none of the internal structure of a face (Figure 5.3). The spoof that best defeats temperature is the one that shape catches most easily – direct evidence for the fusion design of Section 4.6. Projecting the twelve topology features onto their first two principal components (Figure 5.4) shows the same structure: live frames form a coherent region, and the attack families occupy distinct areas, with the heat decoys sitting closest to the live cluster.

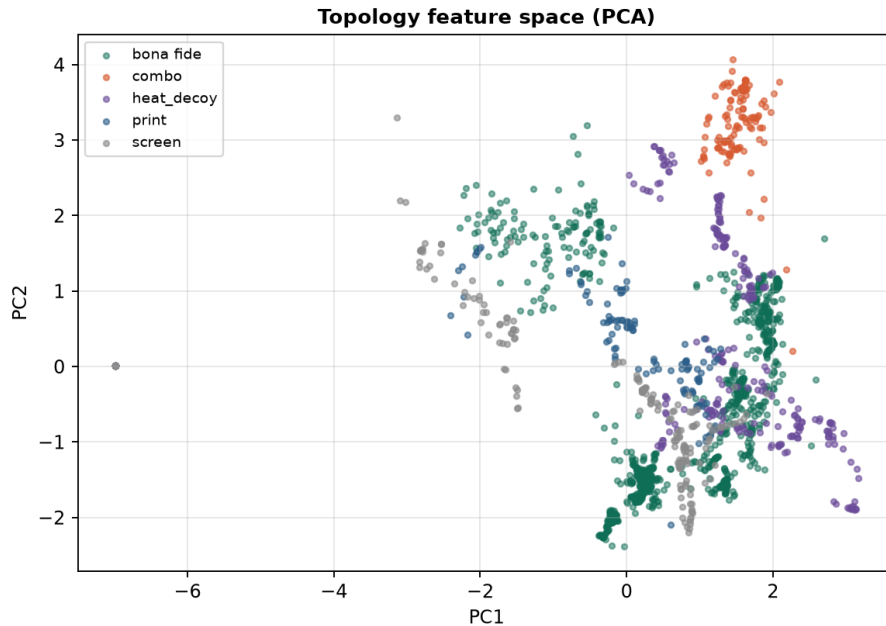


Figure 5.4: The twelve topology features projected onto two principal components. Live frames form a coherent region; the attack families sit in distinct areas, with the heat decoys closest to the live cluster – consistent with topology’s weakness against that family reported in Section 5.5.

5.3 Are the topology features redundant?

The twelve topology features were designed by hand from physical reasoning (Section 4.2), and the AUC figures above report how informative each is *individually*. Individual AUC cannot distinguish an informative feature from a redundant one: a feature can separate the classes well and still add nothing, because it is close to a duplicate of a stronger feature already in the set. Two considerations make this worth checking explicitly. First, the fused detector’s shape component estimates a covariance matrix over these twelve features, and a covariance estimate over strongly correlated inputs is ill-conditioned, so the Mahalanobis distance it produces is disproportionately influenced by whichever near-duplicate directions happen to be noisiest. Second, the elliptic-envelope detector was observed during development to move ACER by approximately 0.005 between otherwise

identical runs (Section 3.7), which is the behaviour an ill-conditioned covariance estimate would produce.

Figure 5.5 shows the Pearson correlation matrix across the twelve features, computed on bona fide frames only – the distribution the one-class model is actually fitted on. Pooling live and attack frames would measure between-class separation rather than the conditioning of the model’s covariance estimate.

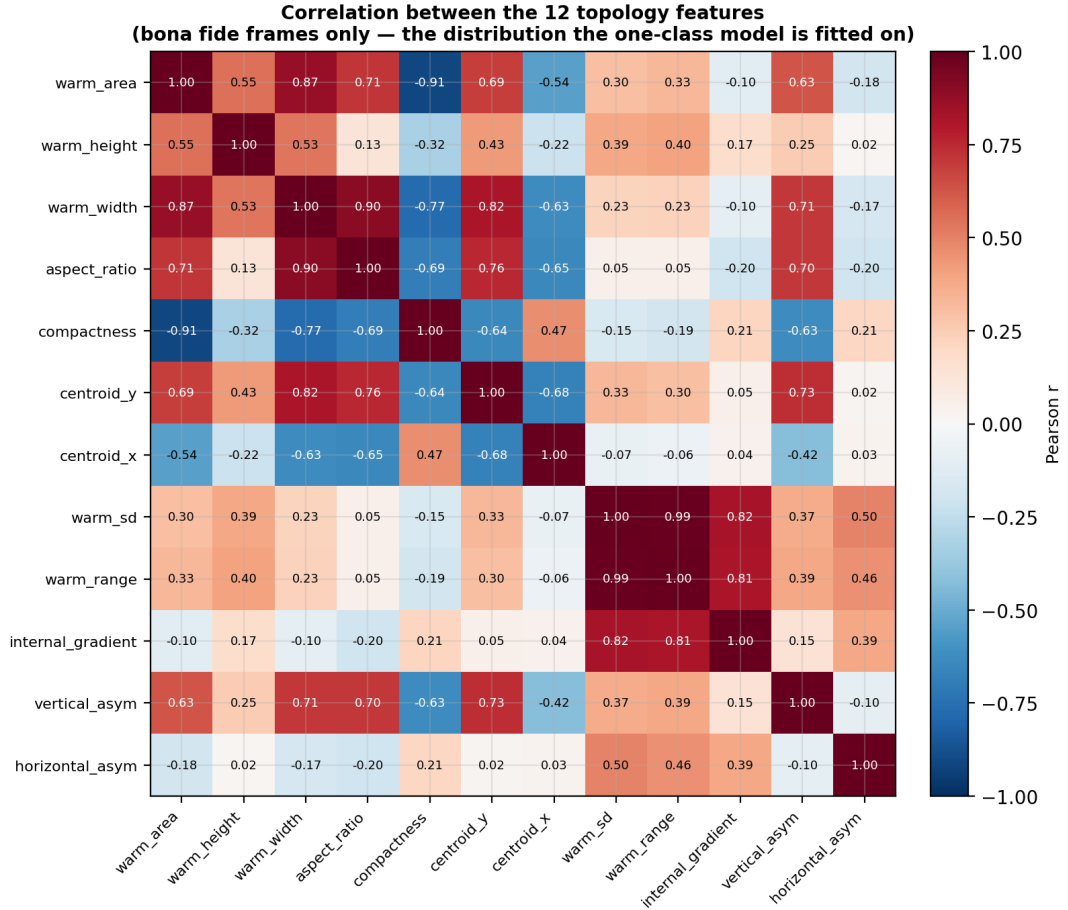


Figure 5.5: Pearson correlation between the twelve topology features, computed over the 981 bona fide frames. Thirteen of the sixty-six distinct pairs reach $|r| \geq 0.7$; warm_sd and warm_range are effectively the same measurement at $r = 0.991$.

The feature set is substantially redundant. Thirteen of the sixty-six distinct pairs reach $|r| \geq 0.7$, and the strongest, warm_sd against warm_range at $r = 0.991$, are effectively the same measurement – both describe the spread of temperature within the warm region, one as a standard deviation and one as a peak-to-peak range. Three further pairs exceed $|r| = 0.85$: warm_area against compactness ($r = -0.909$), which is close to a definitional relationship since compactness is a perimeter-to-area ratio; warm_width against aspect_ratio ($r = 0.901$), unsurprising because aspect ratio is width divided by height; and warm_area against warm_width ($r = 0.872$).

The aggregate picture is summarised in Table 5.1. The correlation matrix has a condition number of 1350, with a smallest eigenvalue of 0.0042, and six of the twelve

principal directions account for 95% of the variance – so the twelve hand-designed features span roughly six effective dimensions.

Table 5.1: Conditioning of the topology feature set (bona fide frames).

Diagnostic	Value
Features	12
Pairs with $ r \geq 0.7$	13 of 66
Strongest pair (warm_sd, warm_range)	$r = 0.991$
Condition number of the correlation matrix	1350
Smallest eigenvalue	0.0042
Directions carrying 95% of variance	6 of 12

Two conclusions follow, and it is worth being careful about which is supported. The first is supported directly: the redundancy explains the elliptic envelope’s run-to-run instability, because inverting a covariance matrix with a smallest eigenvalue of 0.0042 amplifies small changes in the fitted sample into visible changes in the resulting scores. The second is a plausible implication rather than a demonstrated result: a reduced feature set – dropping one of each near-duplicate pair, or using the six principal components – would likely produce a better-conditioned and more stable detector. That was not tested here, because selecting features by re-reading the evaluation metric is the error this project has already been careful to avoid twice (Sections 3.6 and 5.8); doing it properly requires a nested procedure in which feature selection occurs inside the cross-validation loop, on training folds only. It is recommended as future work in Chapter 6 rather than claimed as a result here.

It should be stressed that this redundancy does not invalidate the headline result. The fused detector’s reported ACER of 0.092 is measured under session-grouped cross-validation with fixed seeds, so it is a valid estimate of that specific detector’s performance; the redundancy indicates the design is inefficient and its shape component numerically fragile, not that its measured performance is wrong.

5.4 Baselines and models

5.4.1 Single thresholds

The best single hand-set threshold is the temperature band, at ACER 0.135 (APCER 0.197, BPCER 0.072). The best shape threshold, warm area, reaches 0.188. Both are far better than the 0.500 of random guessing, which sets the bar every subsequent model must clear to justify its added complexity.

5.4.2 Classical supervised models

Of 120 classifier-and-feature-set combinations (Section 4.4), the best is quadratic discriminant analysis on combined normalised-pixel and topology features, at ACER 0.158 – worse than the simple temperature threshold. The most plausible explanation is that supervised models are being asked to learn a decision boundary from four attack families recorded across only seven attack sessions, which is not enough variety to generalise, whereas the threshold encodes a physical fact about human skin rather than a pattern learned from a small sample (Figure 5.6).

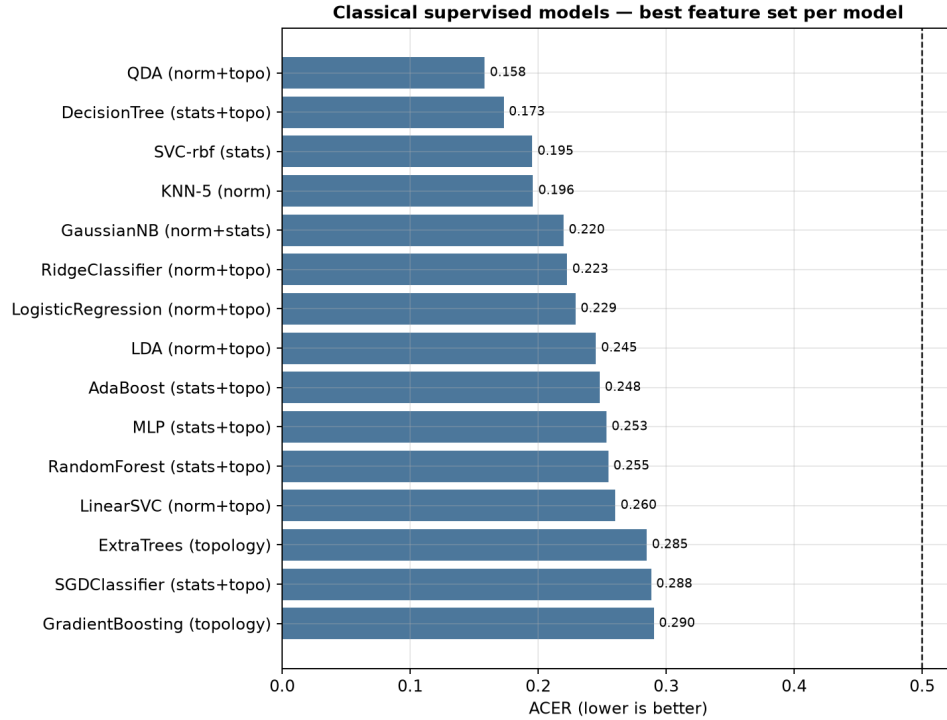


Figure 5.6: Best result for each classical model, using whichever feature set suited it best. None reaches the temperature-threshold baseline of ACER 0.135.

5.4.3 One-class detection

The best one-class model is an isolation forest on topology features, at ACER 0.141, with the elliptic envelope on combined statistics and topology close behind at 0.176 but with a notable APCER of 0.000: it accepted no attack at all, at the cost of rejecting 35% of genuine frames. The one-class framing is more defensible than a supervised classifier even where raw numbers are similar, because it does not assume the attacker uses a spoof already collected – an advantage made concrete in Section 5.6. Figure 5.7 compares the five detectors evaluated.

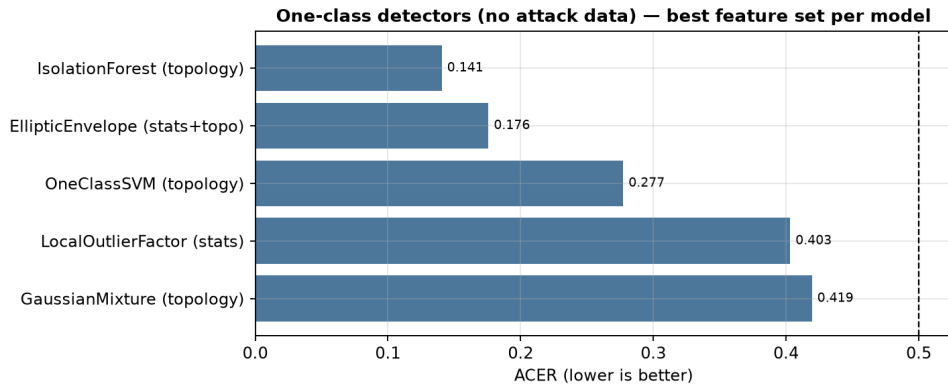


Figure 5.7: The five one-class anomaly detectors, each fitted on bona fide frames only. The isolation forest on topology features is strongest at ACER 0.141; the elliptic envelope trades a perfect APCER of 0.000 against a 35% false-rejection rate.

5.5 Combining two cues: the principal result

Temperature and topology fail on almost entirely different attacks. Temperature rejects 99.6% of heat decoys and 100% of screens but only 44.3% of combination attacks; topology rejects 98.4% of combination attacks and 95.1% of prints but only 9.4% of heat decoys, because a mug or balloon of warm water presents a large, smooth warm region whose shape statistics resemble a face closely enough to pass. Fusing the two cues with the rule described in Section 4.6 – reject if temperature falls outside the live band *or* topology is anomalous – gives the results in Table 5.2.

Table 5.2: Temperature, topology and fused two-cue detector, under session-grouped cross-validation.

Metric	Temperature threshold	Best one-class	Fused two-cue
APCER	0.197	0.134	0.039
BPCER	0.072	0.148	0.146
ACER	0.135	0.141	0.092

The fused detector reaches ACER 0.092 (APCER 0.039, BPCER 0.146), and rejects 99.2% of combination attacks, 99.6% of heat decoys, 100% of screens, and 87.0% of printed photographs – the printed photograph remains the weakest case, since the warmed print sits inside the live band and is flat enough to pass the shape test. This is the project’s best result and its principal contribution (Figure 5.8).

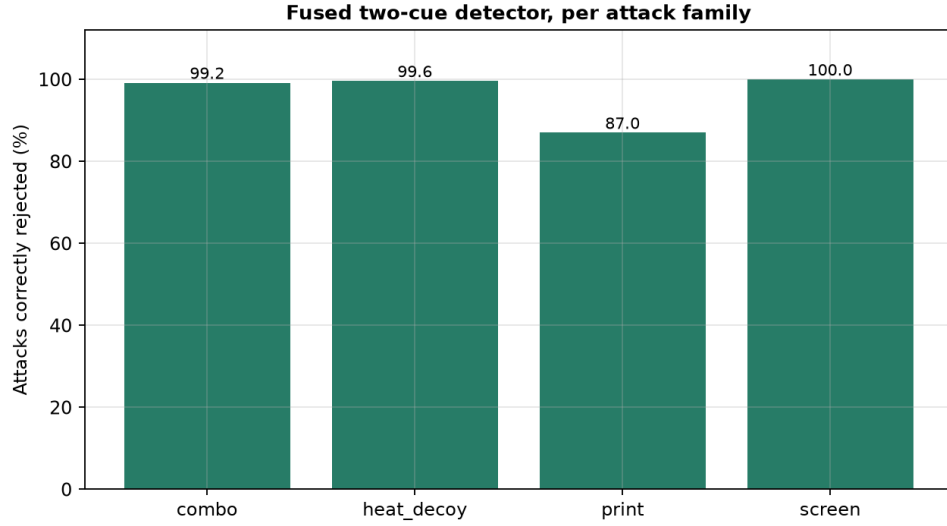


Figure 5.8: Attacks correctly rejected by the fused detector, per family. Printed photographs remain the weakest case, because the warmed print sits inside the live band and is flat enough to pass the shape test.

Sensitivity to the width of the temperature band was checked explicitly: widening it from two to three standard deviations changes ACER by a few hundredths in either direction. The two-standard-deviation result is reported because that interval was chosen in advance as the conventional choice, not because it scored best – the same discipline applied to the cropping-window parameter in Section 5.8.

5.6 Performance against attacks never seen in training

Because the fused detector never sees attack data during fitting, it answers the unseen-attack question by construction. Treating each attack family in turn as the only one withheld from training gives ACER 0.073 on screens, 0.075 on heat decoys, 0.077 on combination attacks, and 0.138 on prints – a mean of 0.091, close to the within-distribution figure of 0.092.

The contrast with supervised training is stark. Training a classifier on three families and testing on the fourth gives a mean ACER between 0.340 and 0.525 depending on the model, with the worst individual configurations reaching 0.773 – worse than guessing. In three of forty-eight supervised configurations the classifier accepted every single instance of the withheld family, and sixteen scored at or near 0.500. This is the strongest practical argument in this dissertation for the one-class design of Section 4.5, and it could only be made once the combination attack existed to serve as a genuinely novel family.

5.7 Cross-participant generalisation: the key limitation

This is the most important experiment in the project, and the one that could not be run until a second participant existed. The detector was fitted on one participant’s bona fide frames only, then tested on the other participant’s bona fide frames together with every attack. If it has learned liveness in the abstract, it should accept the unseen person; if it has learned one individual’s thermal signature, it will reject them.

It largely rejects them. Across thirty configurations of model and feature set, the best cross-participant result is ACER 0.167, against 0.092 within participant. The failure mode is more informative than the average: twenty of the thirty configurations produce a BPCER of 1.000 or above 0.99 – they classify essentially every frame of a real, live human being as an attack (Figure 5.9).

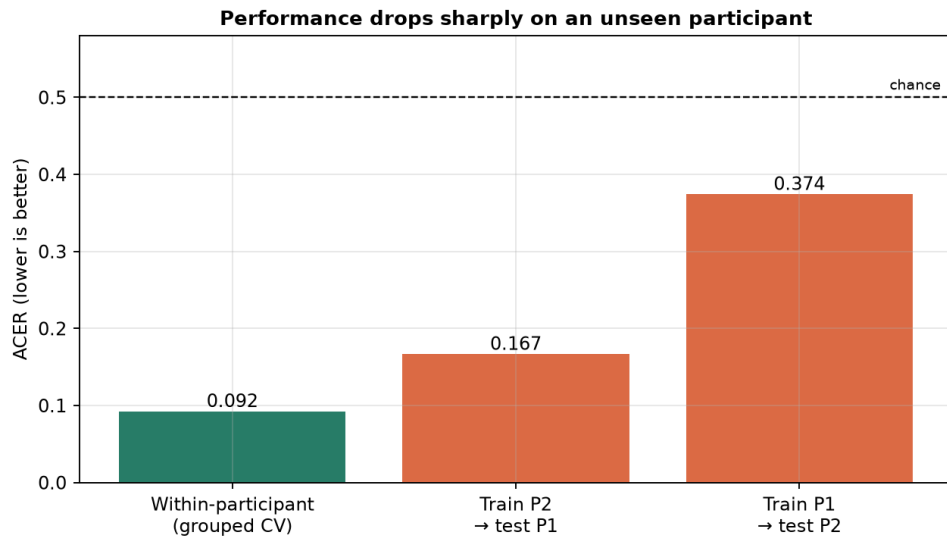


Figure 5.9: Within-participant performance against cross-participant performance. The gap between them is the honest measure of how far this method is from being deployable for users it has not seen.

The direction of transfer matters: training on participant 2 and testing on participant 1 gives ACER 0.167, while the reverse gives 0.374. The most plausible explanation is that participant 1’s own sessions are less varied and were recorded at a higher room ambient, giving a narrower and less representative definition of “normal”.

The consequence should be stated plainly rather than minimised. The within-participant ACER of 0.092 is partly a measure of the detector recognising a specific person’s thermal pattern, not liveness in the abstract. Two participants are not enough to fix this, but they are enough to reveal it, and the 0.64°C mean difference between participants reported in Section 5.1 (Cohen’s $d = 2.04$) is the most likely mechanism.

5.8 Face-region cropping: an attempted fix that did not work

The reasoning behind cropping (Section 4.7) was sound: the sensor’s field of view includes the neck, shoulders and clothing at 30–50 cm, so every area- and shape-based feature is partly a measurement of clothing rather than face. The evidence supporting this concern is concrete: session live_p1, recorded in a sleeveless top, has a warm-area fraction of 0.201, while every other live session sits between 0.463 and 0.655; under grouped validation it is rejected 100% of the time while every other live session sits below 7%.

Cropping did not help. The cropped detector scores ACER 0.106, against 0.092 for the full frame – slightly worse. Testing sensitivity to the one free parameter, the crop window size, across sixteen window sizes gives an ACER range of 0.058 to 0.181, a spread of 0.123 (mean 0.115, standard deviation 0.041). The variation caused by an arbitrary parameter choice exceeds the entire effect being measured, so no single cropped result can be reported as a genuine improvement; the best individual figure of 0.058 was deliberately not selected as a headline result, since choosing a parameter by reading the evaluation metric is the same class of error as the frame-level splitting mistake corrected in Section 3.6. Cropping is therefore reported here as inconclusive, and the torso confound as real and unresolved: at 16×12 there may simply not be enough pixels to crop reliably, which is a further reason the sensor-resolution question raised in Section 3.1 matters. The full sweep is shown in Figure 5.10 and is saved in `results/tables/11_crop_sensitivity.csv` so that the decision not to report a single cropped figure can be audited.

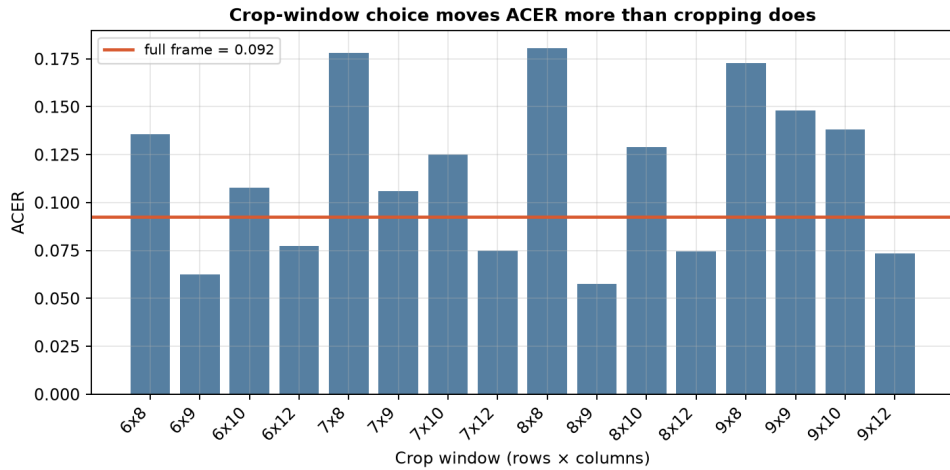


Figure 5.10: ACER for sixteen crop-window sizes against the full-frame baseline. The spread across window sizes (0.058 to 0.181) exceeds any gain from cropping, so no single cropped result can be reported as a genuine improvement.

5.9 Neural networks and transfer learning

Table 5.3 summarises the strongest deep-learning results (Section 4.8), all under the same session-grouped validation used throughout this chapter.

Table 5.3: Best deep-learning results (selected architectures).

Approach	Parameters	APCER	BPCER	ACER
GoogLeNet (frozen)	5,599,904	0.228	0.013	0.121
ResNet50 (frozen)	23,508,032	0.221	0.029	0.125
DenseNet121 (frozen)	6,953,856	0.259	0.014	0.137
ResCNN (from scratch)	37,698	0.262	0.052	0.157
ViT-B/16 (frozen)	85,798,656	0.320	0.025	0.172
Swin-Tiny (frozen)	27,519,354	0.345	0.027	0.186
ConvNeXt-Tiny (frozen)	27,818,592	0.434	0.067	0.251
Convolutional autoencoder	–	0.941	0.101	0.521

The best deep result, GoogLeNet, reaches ACER 0.121; the best of all eleven pretrained and custom networks still does not reach the fused two-cue detector’s 0.092. Two further patterns are worth noting. Larger is not better: the 85-million-parameter Vision Transformer scores 0.172, worse than the 5.6-million-parameter GoogLeNet, and ConvNeXt-Tiny is the weakest backbone tested. The autoencoder essentially failed, at 0.521 – worse than chance – since a 12×16 frame gives it almost no structure to reconstruct, and it reconstructs spoofs about as well as it reconstructs faces.

This is presented as a finding rather than a shortcoming, consistent with Section 4.8: with fifteen independent recording sessions there is not enough data to learn representations from scratch, and generic ImageNet-pretrained features transfer only partially to 192-pixel thermal data. This is a stronger argument for the simple final design than would have been available had the complex alternatives never been tried. Figure 5.11 shows the full comparison.

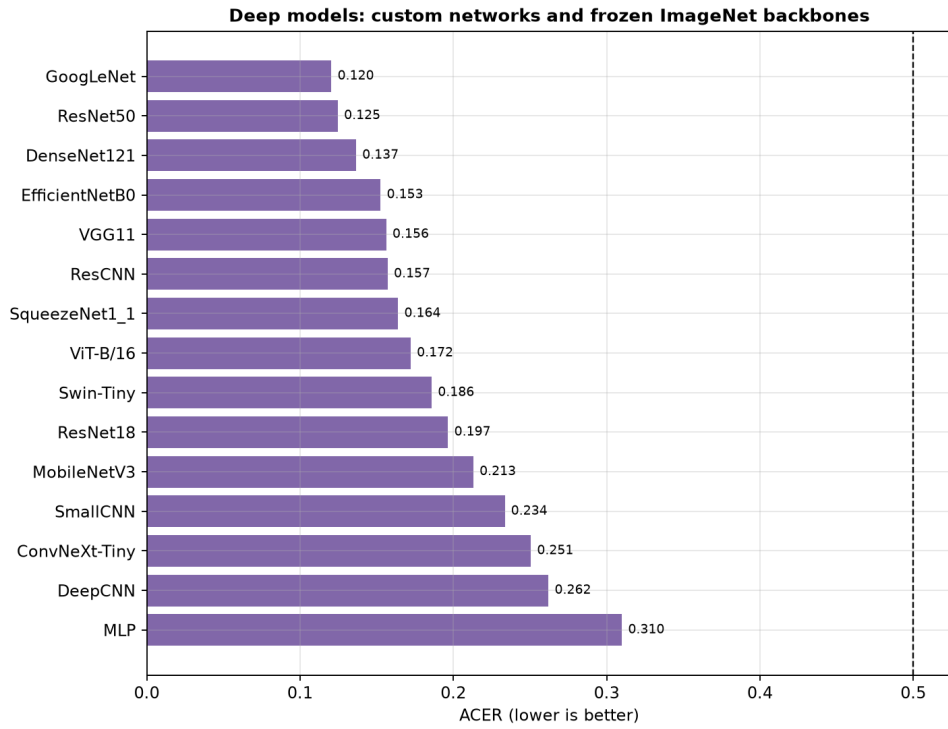


Figure 5.11: Custom networks and frozen ImageNet backbones. The best of them, GoogLeNet at ACER 0.121, still does not reach the fused two-cue detector’s 0.092, and parameter count is not predictive of performance.

5.10 Overall ranking

Table 5.4 summarises the strongest result from each model family evaluated in this dissertation.

Table 5.4: Overall ranking of the strongest result per approach.

Approach	Uses attack data	APCER	BPCER	ACER
Fused two-cue (temperature + topology)	no	0.039	0.146	0.092
GoogLeNet, frozen backbone	yes	0.228	0.013	0.121
ResNet50, frozen backbone	yes	0.221	0.029	0.125
Peak-temperature band alone	no	0.197	0.072	0.135
Isolation forest on topology	no	0.134	0.148	0.141
ResCNN, trained from scratch	yes	0.262	0.052	0.157
QDA, normalised pixels + topology	yes	0.200	0.116	0.158
Convolutional autoencoder	no	0.941	0.101	0.521
Random guessing	—	—	—	0.500

Across seven threshold baselines, 120 classical model-and-feature combinations, forty one-class configurations, four custom networks, an autoencoder and eleven pretrained

architectures, the best result comes from combining two physically motivated cues. It uses no attack data in training, generalises to unseen attack families (Section 5.6), and fits in under a kilobyte of stored parameters.

5.11 Could this be built as working software? Deployment feasibility

Aggregating decisions over a short window of consecutive frames improves matters modestly: a 31-frame window (approximately 15 seconds at 2 fps) gives APCER 0.000 and BPCER 0.133, for ACER 0.066; shorter windows sit between that figure and the single-frame ACER of 0.092. The improvement from averaging, 0.026, is itself informative – if errors were random frame-level noise, averaging thirty-one frames would reduce them sharply, but the curve is nearly flat, indicating the errors are systematic rather than random: the same poses and the same sessions fail consistently, which points toward needing more varied live data rather than a longer decision window (Figure 5.12).

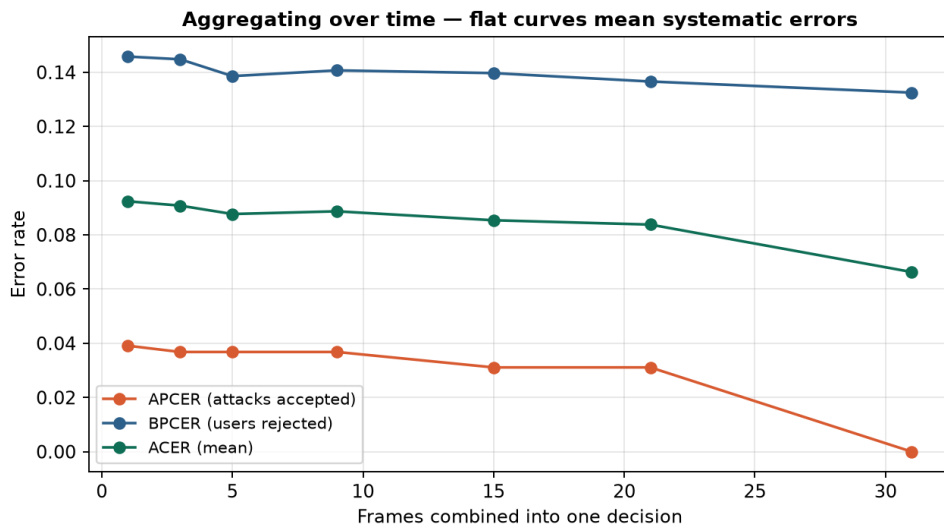


Figure 5.12: Error rates as a function of how many consecutive frames are combined into one decision. The curves are almost flat, indicating that the same frames are misclassified consistently rather than at random.

On cost, the trained fused detector is a mean vector and covariance matrix over twelve features – approximately 180 numbers, well under a kilobyte. Feature extraction takes 0.38 ms per frame and scoring 0.47 ms, so a complete decision costs under one millisecond against a 500 ms budget at 2 fps. No GPU or deep-learning runtime is required; the pipeline – read a frame over I²C, compute twelve features in NumPy, score against the stored model, output a decision – would run comfortably on a Raspberry Pi or a microcontroller, consistent with the deployment feasibility demonstrated at comparable resolution by Vandersteegen et al. [18].

For a security layer, the error asymmetry is the right way round: a false rejection is an inconvenience, while a false acceptance is a security breach. Allowing retries brings the practical failure rate for a genuine user from 13.3% on a first attempt to 1.8% after two attempts and 0.2% after three – though this assumes attempts are statistically independent, which the systematic error pattern identified above suggests is optimistic, and this figure should therefore be read as a best case rather than an expected deployment figure.

5.12 Discussion

Taken together, these results support a coherent central claim: a two-cue detector with a clear physical rationale, validated without the leakage identified in Section 3.6, tested against four attack families including one built specifically to defeat it, generalises to attack families withheld from training, and outperforms every deep model evaluated against it, using under a kilobyte of stored parameters. This claim is qualified, not unconditional: it holds within the two participants recorded for this project, and it is not yet shown to hold for a person the detector has not seen. Section 5.7 treats that qualification as a finding in its own right, since it could only be discovered, not merely suspected, once a second participant existed to test it – and since resolving it is identified in Chapter 6 as the most valuable direction for continued work.

Chapter 6

Conclusions and Recommendations

6.1 Revisiting the research question

This project set out to answer whether a low-resolution thermal sensor array can reliably distinguish a live human face from presentation attacks in a privacy-preserving manner, and which lightweight approach performs best under that resolution constraint. Within the population of participants recorded for this project, the answer is yes: a detector combining peak temperature and the topology of the warm region, fitted only on genuine presentations, achieves an Average Classification Error Rate of 0.092 under session-grouped cross-validation, outperforms fifteen classical classifiers, five one-class anomaly detectors and eleven pretrained or custom deep networks evaluated on the same 192-pixel data, and generalises to attack families withheld entirely from training. It does so using approximately 180 stored numbers and under one millisecond of computation per decision, which supports the deployment feasibility claimed in Section 5.11.

That answer, however, is conditional in a way the original research question did not anticipate: the detector does not yet generalise across people. Trained on one participant and tested on another, the best cross-participant ACER is 0.167, and most configurations reject nearly every genuine frame from the unseen participant. This project therefore answers its research question in two parts – liveness detection at privacy-preserving thermal resolution is achievable for a known population, and remains unproven for an unknown one – and treats the second part as the more important finding, since it could not have been discovered without deliberately recording a second participant to test for it.

6.2 Objectives revisited

Of the five objectives set out in Section 1.3, four were achieved in full: a capture pipeline was built and characterised; a purpose-built dataset of 1,850 frames across four attack families was collected; a broad comparison of lightweight and deep approaches was carried out; and detection performance was evaluated throughout using the ISO/IEC 30107-3

metrics. The fifth objective, establishing baseline behaviour, was also achieved, though it required resolving a hardware discrepancy first: the original proposal specified the MLX90640 (32×24), while every recording made for this project logged 192 values rather than 768. The diagnostic work in Section 3.1 ruled out interpolation and reading-pattern settings as causes and established that the board is populated with the related MLX90641 (16×12). This does not affect the validity of the results, since the pipeline handles either frame size automatically and every reported number was produced at the geometry actually recorded, but it does constrain the interpretation of the cropping experiment in Section 5.8, which has only 192 pixels to work with.

6.3 Limitations

The following limitations apply to the results reported in this dissertation, in order of importance.

1. **The method does not generalise across people.** The best cross-participant ACER is 0.167 against 0.092 within participant, and twenty of thirty cross-participant configurations reject essentially every frame of the unseen person. This is the most important limitation and is measured directly rather than suspected.
2. **Two participants is a very small sample.** A generalisation gap can be demonstrated to exist, but not characterised across a population, and a population-specific cause cannot be ruled out.
3. **The torso confound is real and unresolved.** Shape features partly measure clothing rather than face, and cropping at 16×12 resolution was too sensitive to window-size choice to be trusted (Section 5.8).
4. **One session dominates the false-rejection rate.** The sleeveless-top session for participant 1 is rejected 100% of the time while every other live session sits below 7%; excluding it would roughly halve BPCER, but excluding data because it scores badly is not defensible and it is retained and reported throughout.
5. **Printed photographs remain the weakest attack case,** rejected at 87.0%, because a warmed print sits inside the live temperature band and is flat enough to pass the shape test.
6. **The heat decoys used are warmer than an ideal attack would be** (37.2°C and 39.1°C against a 35.4°C face), making them easier to reject than a well-executed decoy held closer to body temperature.
7. **No mask attack was tested.** The literature treats masks as the most serious attack instrument, but a custom silicone mask costs around USD 4,000 [9], and a cheap

substitute reduces thermally to cases the print and heat-decoy families already cover (Section 3.4). Masks are outside the threat model reported here, and are the clearest gap in attack coverage.

8. **All recordings were made at a room ambient of 26–32°C.** Behaviour at very low or very high ambient temperature is untested, which matters because the detector’s temperature cue is an absolute threshold rather than a differential one.
9. **The topology feature set is partly redundant.** Thirteen of sixty-six feature pairs correlate at $|r| \geq 0.7$, and the twelve features span roughly six effective dimensions (Section 5.3). This makes the shape model’s covariance estimate ill-conditioned and explains its run-to-run instability; it does not invalidate the reported performance, but it means the design is inefficient.
10. **All results are specific to a 16×12 array.** The board used here is populated with the MLX90641 (Section 3.1), and the cropping result in particular is a consequence of that resolution rather than of the method; a 32×24 sensor would leave enough pixels for face-region cropping to be worth revisiting.

6.4 Recommendations for future work

1. **Recruit a third and further participants.** This is the only step that would directly close the generalisation gap identified in Section 5.7, and is the highest-value single action for any continuation of this work.
2. **Repeat the study on a 32×24 array.** The board used here is populated with the MLX90641 (Section 3.1), so every result is specific to 192 pixels. Running the same pipeline on an MLX90640 would establish how much of the performance reported here is limited by resolution rather than by method, and would make face-region cropping worth revisiting.
3. **Exploit the temporal dimension.** Every feature used here is computed from a single frame, so the detector has no access to how a thermal field changes over time. A living face fluctuates continuously as blood flow and ambient conditions vary, whereas a passively heated artefact follows the simpler cooling profile of whatever is warming it (Section 2.3). Features describing that variation across a short window, rather than a majority vote over independently classified frames as used in Section 5.11, are the most promising untested cue identified by this work, and would be difficult for an attacker to reproduce without actively regulating the artefact’s temperature.
4. **Revisit face-region cropping at full (32×24) resolution,** if the sensor question above resolves in favour of that resolution being available; the torso confound is real,

and more pixels may make a reliable, non-arbitrary crop possible where 16×12 did not.

5. **Collect data across a wider range of ambient temperatures**, to test whether the live temperature band established in Section 5.1 holds outside the 26–32°C range in which this project’s data was collected.
6. **Reduce the topology feature set**. Section 5.3 shows the twelve hand-designed features span roughly six effective dimensions, with thirteen pairs at $|r| \geq 0.7$ and one pair at $r = 0.991$, giving a correlation matrix with a condition number of 1350. Removing near-duplicate features, or substituting the leading principal components, should yield a better-conditioned covariance estimate and a more stable shape model. This must be done with feature selection nested *inside* the cross-validation loop and fitted on training folds only, so that the selection does not read the evaluation metric.
7. **Investigate a client-specific or few-shot enrolment step**, following Fatemifar et al. [21], as a practical mitigation for the cross-participant gap if a fully person-independent detector proves difficult to achieve with a larger dataset.
8. **Build the live demonstration**. The laboratory setup is documented in Figure 3.1, but a real-time implementation running the detector on a microcontroller would test the deployment-feasibility argument of Section 5.11 directly rather than by estimate.

6.5 Closing statement

This project set out to test whether a sensor coarse enough to make identification impossible could still make liveness detection possible. The results say yes, with a caveat that turned out to be more interesting than the headline result: a simple, physically motivated, one-class detector beats every deep model tried, generalises to attacks it has never seen, and costs almost nothing to run – but it has, so far, learned as much about the specific people it was shown as about liveness itself. Establishing whether that gap closes with more people, or reflects something fundamental about person-specific thermal signatures at this resolution, is the natural and necessary next question.

Chapter 7

Reflection

7.1 On a result that was too good

The first working version of my analysis pipeline reported an ACER of 0.000. My reaction was not satisfaction but disbelief: there was no way the system was detecting this accurately, not from 192 pixels, and not on the first attempt. The number was suspicious precisely because it was perfect.

I acted on that instinct more slowly than I felt it. The doubt was immediate, but I did not sit down and investigate properly until the following day. In hindsight the delay is the part I would change — not the mistake itself, which was reasonable, but the gap between noticing that something was wrong and doing something about it.

What I found was that my splitting procedure was assigning frames randomly to training and test sets. At two frames per second, consecutive frames from a single sixty-second recording are almost identical, so a random split placed near-copies of the same frame on both sides. The model was not recognising liveness at all. It was recognising individual recordings it had already seen.

I found this more interesting than embarrassing. It is a specific and understandable failure, and it taught me something about my data that I would not have learned from a correct result: that frames within a recording are not independent observations, and that the unit of independence in this dataset is the *session*, not the frame. Every evaluation reported in this dissertation follows from that correction (Section 3.6).

My general position on results has not changed, and I do not think it should. I do not distrust a number by default; I check that it is sound. But if a result looks wrong, my assumption is that something has been missed or has gone wrong somewhere, rather than that the result is simply surprising. Working on that assumption is what found this error.

7.2 On discarding a recording while the participant was still in the room

During participant 2's session I checked each recording numerically before moving to the next one rather than waiting until the end. The first 40 cm take looked wrong immediately: the peak temperature had dropped 1.37°C from the 30 cm session, where my own recordings had dropped only 0.20°C over the same step. The variation was too different from participant 1 for me to accept it. When I looked at where the warm region actually sat in the frame, its centre had moved to the right and part of the face had passed outside the sensor's field of view.

I did consider keeping it. It was usable data in the sense that it existed, and I was conscious that a retake meant asking for more of someone's time. That was the part I found uncomfortable rather than the technical decision — everyone has something going on in their life, and taking more of it than you said you would is not a small thing to ask. I asked anyway, and the second attempt gave a 0.31°C step, consistent with my own data.

What I would emphasise is the order in which the decision was made. The recording was rejected because of a framing fault I could identify and describe, and it was rejected before any model was run on it. That distinction matters: discarding data because of a documented capture problem is legitimate, whereas discarding it because of the score it produces is not, and the two can look identical afterwards if you do not write down which one happened.

7.3 On not taking the better number

The face-cropping experiment produced a result of 0.058 at its best crop window, against 0.092 for the full frame. I considered taking it. The difference was large enough to be tempting, and it would have made the headline result substantially stronger.

I did not take it, for two reasons. The first is about what the number actually measures. That result was obtained by cropping to the face region, but in real use the sensor does not see only a face — at 30 to 50 cm it captures the neck and shoulders as well, and a system that reports 0.058 is quietly assuming a cropping step that would have to exist at runtime. Producing that step reliably at 16×12 resolution is precisely the problem the cropping experiment tried, and failed, to solve (Section 5.8). Reporting the full-frame figure is reporting the number that corresponds to how the system would actually be used.

The second reason is that 0.058 was the best of sixteen windows, and the spread across those windows ran from 0.058 to 0.181 — wider than any improvement cropping offered. Selecting the window that produced the best score would have been choosing a parameter by reading the evaluation metric, which is the same class of error as the frame-level split

that produced my ACER of 0.000. Recognising that I was about to make the same mistake in a different form is the reason I reported the whole sweep instead of the best cell.

7.4 On expecting the big networks to win

I expected the pretrained architectures to outperform my hand-designed features. That expectation seemed reasonable: these are the models that dominate image classification, and it did not seem likely that twelve numbers I had defined myself would beat a network with eighty-five million parameters pretrained on the 1.3 million labelled images of ImageNet-1k.

They did not win. The best of eleven pretrained backbones, GoogLeNet, reached ACER 0.121 against 0.092 for the fused two-cue detector, and the Vision Transformer managed only 0.172 despite being by far the largest model tested. The four networks I trained from scratch did worse still.

I think there are two main reasons, and I am not certain they are the only ones. The first is that a 16×12 frame gives a deep network very little to work with. A network's advantage is finding fine spatial structure, and at 192 pixels there is almost none to find — a face occupies perhaps forty pixels. Upsampling to 224×224 before feeding a backbone does not create information the sensor never captured. The second is that fifteen independent recording sessions is not enough data to learn a representation from scratch. These models are designed to discover which features matter from large and varied datasets, and I gave them a small and fairly uniform one.

There may be more to it than that. What I am confident of is the practical consequence: at this resolution and this dataset size, features derived from what is physically known about the problem outperformed features learned from data. That is the finding that justifies my final design, and I would not have been able to claim it if I had assumed the answer instead of testing it.

7.5 On finding closely related work late

Late in the project I ran a systematic search to test the novelty claim my literature review was making, and found Imam [25], which applies a low-cost thermal sensor array to face presentation attack detection at Nottingham Trent University, using the same evaluation board I had been lent.

My first reaction was anxiety. From the title and abstract it looked as though it might match my project entirely, and for a short while I thought the work I had spent the summer on had already been done by someone else, with more data and at my supervisor's own institution.

Reading it properly changed that. The overlap is real and I have stopped claiming otherwise, but the differences are substantive. That work pairs the thermal array with a conventional RGB camera, which reintroduces precisely the privacy liability that choosing a coarse sensor was meant to avoid; it reports accuracy and F1 rather than the APCER, BPCER and ACER that ISO/IEC 30107-3 specifies; it trains supervised networks on attack examples where my detector requires none; and it does not report whether performance holds for a person the system has not seen. My contribution is narrower than I originally claimed, but it is still a contribution (Section 2.7.1).

What the overlap taught me is that novelty is rarely a single claim that is either true or false. My original framing treated it as a binary question — either this had been done or it had not — when the useful question was narrower: which specific combination of sensor, metric and evaluation protocol had not been reported. Reading the work carefully was what turned an apparent duplication into a clearly stated difference, and that reading was more valuable to the dissertation than the search that produced it.

The lesson is about sequencing rather than about the finding itself. I ran the novelty search near the end, when the claim was already written into two chapters. Running it at the start would have cost an afternoon and would have shaped the framing from the beginning rather than requiring it to be corrected.

7.6 On the participants I did not have

The limitation that troubles me most is the one I could not fix: two participants is not many, and I am aware throughout Chapter 5 that every claim about generalisation rests on a sample of two.

What makes it harder is that the direction of travel has not been encouraging. My first result was an ACER of 0.000, which turned out to be an artefact of leaked validation. Correcting the split moved it to roughly chance. Adding a second participant then exposed a further gap, between 0.092 within a participant and 0.167 across. Each time the evaluation became more honest or the data became more varied, the reported performance got worse. Extrapolating that trend, my instinct is that more participants would push the numbers down again.

But I hold that instinct loosely, because the opposite argument is just as reasonable: models generally improve with more and more varied training data, and a one-class model fitted on eight sessions from two people has a very narrow definition of normal to work from. It is entirely possible that ten participants would produce a detector that generalises properly and scores better than either figure I report. I genuinely do not know which of these is true, and I think saying so is more useful than guessing.

I did try to recruit more people. Two declined, and the reason they gave was that they felt uneasy about sharing this kind of data. I find that worth recording, because

this project's central claim is that a 16×12 thermal array cannot identify anyone — 192 temperature values contain nothing a person could be recognised from. The technical privacy argument is sound, but it did not transfer to the people I asked. Perceived privacy and actual privacy are not the same thing, and a system that is genuinely privacy-preserving can still fail to feel that way. If this work were continued, that is a problem to design for rather than to argue away.

Given the time available I would not do much differently, other than starting recruitment earlier. What I want to be clear about is the shape of my uncertainty: I am not uncertain about the 0.092. That number is honestly obtained and I would defend it. What I do not know is how far it travels.

References

- [1] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. DOI: 10.1109/TPAMI.2022.3215850.
- [2] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2019. Introduces the WMCA database; DOI: 10.1109/TIFS.2019.2916652, preprint arXiv:1909.08848.
- [3] Marek Kowalski, Arkadiusz Grudzień, and Kacper Ciężyński. A study on presentation attack detection in thermal infrared. *Sensors*, 20(14):3988, 2020. DOI: 10.3390/s20143988.
- [4] Abdallah Naser, Ahmad Lotfi, and Caixia Zhong. Towards human distance estimation using a thermal sensor array. *Neural Computing and Applications*, 35(32):23357–23367, 2021. DOI: 10.1007/s00521-021-06193-2.
- [5] Abdallah Naser, Ahmad Lotfi, and Caixia Zhong. Calibration of low-resolution thermal imaging for human monitoring applications. *IEEE Sensors Letters*, 6(3):1–4, 2022. DOI: 10.1109/LSENS.2022.3155936.
- [6] Abdallah Naser, Ahmad Lotfi, and Caixia Zhong. Multiple thermal sensor array fusion towards enabling privacy-preserving human monitoring applications. *IEEE Internet of Things Journal*, 9(17):16677–16688, 2022. doi: 10.1109/jiot.2022.3150566.
- [7] Abdallah Naser, Ahmad Lotfi, and Caixia Zhong. Adaptive thermal sensor array placement for human segmentation and occupancy estimation. *IEEE Sensors Journal*, 21(2):1993–2002, 2020. doi: 10.1109/jsen.2020.3020401.
- [8] International Organization for Standardization. ISO/IEC 30107-3: Information technology – Biometric presentation attack detection – Part 3: Testing and reporting. International Organization for Standardization, Geneva, 2023.
- [9] Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing deep face recognition with custom silicone masks. IEEE 9th International Conference on

Biometrics Theory, Applications and Systems (BTAS), 2018. Introduces the CSMAD dataset.

- [10] Raghavendra Ramachandra and Christoph Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys*, 2017. See also the RGB-camera-focused survey, arXiv:2010.04145.
- [11] Margarita N. Favorskaya. Face presentation attack detection: Research opportunities and perspectives. *Intelligent Decision Technologies*, 2023. DOI: 10.3233/IDT-220197.
- [12] Vincent A. Weidlich. Thermal infrared face recognition. *Cureus*, 2021. doi: 10.7759/cureus.13736.
- [13] Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020. Introduces HQ-WMCA; arXiv:2009.09703.
- [14] Anjith George and Sébastien Marcel. A comprehensive evaluation on multi-channel biometric face presentation attack detection, 2022. arXiv:2202.10286.
- [15] Nishat Tasnim Newaz and Eisuke Hanada. A low-resolution infrared array for unobtrusive human activity recognition that preserves privacy. *Sensors*, 24(3):926, 2024. doi: 10.3390/s24030926.
- [16] Xie Zhang and Chenshu Wu. Tadar: Thermal array-based detection and ranging for privacy-preserving human sensing. In *Proceedings of the 25th International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, pages 11–20, 2024. doi: 10.1145/3641512.3686357. arXiv:2409.17742; code: github.com/aiot-lab/TADAR.
- [17] Cunyi Yin, Jing Chen, Xiren Miao, Hao Jiang, and Deying Chen. Device-free human activity recognition with low-resolution infrared array sensor using lstm. *Sensors*, 21(10):3551, 2021. doi: 10.3390/s21103551. 8x8 Grid-EYE array; PMC8161224.
- [18] Maarten Vandersteegen, Wouter Reusen, Kristof Van Beeck, and Toon Goedemé. Person detection using an ultra low-resolution thermal imager on a low-cost mcu, 2022. arXiv:2212.08415.
- [19] Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, vol. 5, pp. 13868–13882, 2017.

- [20] Olegs Nikisins, Amir Mohammadi, André Anjos, and Sébastien Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. *International Conference on Biometrics (ICB)*, 2018.
- [21] Soroush Fatemifar, Shervin Rahimzadeh Arashloo, Muhammad Awais, and Josef Kittler. Client-specific anomaly detection for face presentation attack detection. *Pattern Recognition*, 2021. DOI: 10.1016/j.patcog.2020.107696.
- [22] R. Vardasca, A. R. Marques, J. Diz, A. Seixas, J. Mendes, and E. F. J. Ring. The influence of angles and distance on assessing inner-canthi of the eye skin temperature. *Thermology International*, 27(4):130–135, 2017.
- [23] Eleonora Micheletti, Nevin W. El-Nimri, Robert N. Weinreb, and John H. K. Liu. Relative stability of regional facial and ocular temperature measurements in healthy individuals. *Translational Vision Science & Technology*, 2022. doi: 10.1167/tvst.11.12.15. PMC9804027.
- [24] Yangling Zhou, Pejman Ghassemi, Michelle Chen, David McBride, Jon P. Casamento, T. Joshua Pfefer, and Quanzeng Wang. Clinical evaluation of fever-screening thermography: Impact of consensus guidelines and facial measurement location. *Journal of Biomedical Optics*, 25(9):097002, 2020. doi: 10.1117/1.JBO.25.9.097002. PMC7486803.
- [25] Abdulrafiu Musa Imam. Biometric presentation attack detection. *American Journal of Computing and Engineering*, 8(1):32–56, 2025. doi: 10.47672/ajce.2631.
- [26] Melexis NV. EVB90640-41 evaluation board user manual. Melexis NV, Ypres, Belgium, 2019. URL <https://www.melexis.com/en/product/EVB90640-41/>.
- [27] Melexis NV. MLX90641 16 × 12 far infrared thermal sensor array datasheet. Melexis NV, Ypres, Belgium, 2019. URL <https://www.melexis.com/en/product/MLX90641/>.
- [28] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 1995. DOI: 10.1007/BF00994018.
- [29] Leo Breiman. Random forests. *Machine Learning*, 2001. DOI: 10.1023/A:1010933404324.
- [30] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001. DOI: 10.1214/aos/1013203451.
- [31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2011. arXiv:1201.0490.

- [32] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, 2008. DOI: 10.1109/ICDM.2008.17.
- [33] Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 1999. DOI: 10.1080/00401706.1999.10485670.
- [34] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 2001. DOI: 10.1162/089976601750264965.
- [35] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000. DOI: 10.1145/335191.335388.
- [36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *IEEE CVPR*, 2015. GoogLeNet; arXiv:1409.4842.
- [37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE CVPR*, 2016. ResNet; arXiv:1512.03385.
- [38] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *IEEE CVPR*, 2017. DenseNet; arXiv:1608.06993.
- [39] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016. arXiv:1602.07360.
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. VGG; arXiv:1409.1556.
- [41] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*, 2019. arXiv:1905.11946.
- [42] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019.
- [43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *IEEE CVPR*, 2022. ConvNeXt; arXiv:2201.03545.

- [44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. ViT; arXiv:2010.11929.
- [45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. IEEE ICCV, 2021. arXiv:2103.14030.
- [46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. DOI: 10.1007/s11263-015-0816-y.
- [47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS), 2019. arXiv:1912.01703.

Appendices

Appendix A

Dataset Composition and Reproducibility

A.1 Per-session breakdown

Table A.1 lists every recording session in the dataset, with its class, participant, capture distance where fixed, frame count, and mean and standard deviation of peak temperature. This is the full version of the summary given in Table 3.1.

Table A.1: Every recording session in the dataset.

Session	Class	Type	Part.	Dist.	Frames	Peak °C
live_p1	bona fide	–	1	–	130	35.42 ± 0.38
live_p1_30cm	bona fide	–	1	30 cm	109	35.95 ± 0.15
live_p1_40cm	bona fide	–	1	40 cm	125	35.75 ± 0.19
live_p1_50cm	bona fide	–	1	50 cm	122	35.61 ± 0.19
live_p2	bona fide	–	2	–	125	34.86 ± 0.24
live_p2_30cm	bona fide	–	2	30 cm	122	35.38 ± 0.21
live_p2_40cm	bona fide	–	2	40 cm	123	35.07 ± 0.24
live_p2_50cm	bona fide	–	2	50 cm	125	34.85 ± 0.20
photo_plain	attack	print	–	–	138	31.63 ± 0.52
photo_warm	attack	print	–	–	108	35.51 ± 0.70
screen_phone	attack	screen	–	–	114	33.12 ± 0.25
screen_laptop	attack	screen	–	–	131	37.33 ± 0.70
cup_water	attack	heat decoy	–	–	123	37.20 ± 0.27
balloon_water	attack	heat decoy	–	–	133	39.08 ± 0.45
combo_photo_heat	attack	combination	–	–	122	36.05 ± 0.55
Total					1,850	

The three participant-1 sessions recorded at fixed distance sit consistently around 0.6°C warmer than the corresponding participant-2 sessions, which is the between-participant difference discussed in Section 5.1 and the mechanism proposed for the cross-participant failure in Section 5.7.

A.2 Repository layout

Table A.2: Project repository layout.

Location	Contents
thermal_pad_dataset/recordings/	Raw sensor CSV exports (source of truth)
thermal_pad_dataset/processed/	1,850 .npz frames used by the models
thermal_pad_dataset/manifest.csv	Index of every frame with metadata and split
thermal_pad_dataset/analysis/	18-script analysis suite plus README
thermal_pad_dataset/results/tables/	29 result tables (CSV) plus 17_summary.md
thermal_pad_dataset/results/figures/	17 figures plus FIGURE_INDEX.md
thermal_pad_dataset/results/logs/	Full console output of every script
thermal_pad_dataset/scripts/	Sensor import, manifest builder, batch runners
thermal_pad_dataset/analysis_v1_july/	Superseded first-round suite, kept as a record
dissertation/	L ^A T _E X source, figures and bibliography

A.3 Reproducing any result

Running the full analysis suite:

```
conda activate thesis
cd thermal_pad_dataset/analysis
python run_all.py
```

Useful variants:

```
python run_all.py --skip-deep # skips scripts 12-14 (no PyTorch needed)
python run_all.py --quick-transfer # script 14 uses 4 backbones, not 11
python run_all.py --only 08 10 # just the fused detector and the
                                # cross-participant test
```

Adding a new recording:

```
cd thermal_pad_dataset/scripts
python import_sensor_csv.py "<file>.csv" --label bona_fide --participant N
--fresh
python build_manifest.py
```

Every number quoted in this dissertation traces to a CSV file in results/tables/, and results/tables/17_summary.md collects the headline figures in one place. All

random seeds are fixed (Section 3.7), so re-running the suite reproduces every reported value exactly within the same environment. The full run takes approximately fifty minutes.

Appendix B

Ethics Approval and Participant Materials

B.1 Approval

An application for ethical approval was submitted using the University's *Application for Ethical Approval* form for undergraduate and taught postgraduate programmes, under student number 2428134, and full ethical approval was granted before any data collection took place. No recording of any kind was made prior to approval.

No reference number was issued for this application; approval was recorded by signature on the returned application form, which is reproduced in the electronic appendix.

B.2 Participant materials

The following participant-facing documents were used, and are reproduced in the electronic appendix submitted alongside this dissertation:

- Participant Information Sheet
- Participant Consent Form

Every participant was given the Information Sheet and provided written informed consent on the approved Consent Form before recording began.

Blank templates of both documents are reproduced in the electronic appendix. Completed consent forms carry participant signatures and are therefore not reproduced; the signed originals are retained securely by the researcher and are available to the examination board on request, in line with the approved application.

A note on the resolution stated in the participant materials

Both the approved application and the Participant Information Sheet describe the sensor as a Melexis MLX90640 producing 32×24 (768-pixel) frames. That was the honest understanding when those documents were written and approved, and it is what participants were told; the board carries no external marking distinguishing the two sensors it accepts, and the discrepancy surfaced only later, when every recording was found to log 192 values (Section 3.1).

The participant materials have deliberately not been altered after the fact, since they record what participants were actually given; the correction is disclosed here instead. It works in participants' favour: every frame was collected at a *lower* resolution than the one they consented to, so the privacy protection promised was exceeded rather than fallen short of, and the non-identifiability argument on which approval rested holds more strongly at 16×12 than at 32×24 .

B.3 Data protection measures

The following measures were applied throughout, consistent with the approved application:

- Participant 1 is the researcher, as stated in Section 3.3. Self-inclusion was covered by the approved application, and the researcher completed the same Information Sheet and Consent Form process as participant 2. The methodological consequence – that a detector fitted largely on one person may be learning that person rather than liveness – is measured directly rather than assumed away, in Section 5.7.
- No names, photographs, video or audio were collected *as data* at any point; the dataset consists solely of 16×12 temperature arrays. The laboratory setup photographs in Figure 3.1 were taken separately for documentation, show the researcher only, and form no part of the dataset.
- All thermal recordings are indexed by participant number only.
- No name-to-number key was retained, so the stored data is non-identifiable rather than merely pseudonymised.
- The sensor's own resolution (16×12 pixels, 192 temperature values) means no stored frame is capable of revealing a participant's identity. This is not an administrative control but a property of the hardware, and it forms part of the privacy argument the project makes (Section 2.4).
- Data is stored on the researcher's own machine, backed up regularly, and will be deleted in accordance with the retention period stated in the approved application.

Appendix C

Analysis Code Index

The analysis suite is submitted separately via the code-submission link referenced in the MSc Project Handbook. It consists of eighteen numbered scripts and two shared modules, listed in Table C.1, forming a single pipeline that runs end to end from one command.

Table C.1: The analysis suite.

#	Script	What it answers
–	<code>common.py</code>	Shared loading, feature extraction, metrics, grouped cross-validation, seed control
–	<code>face_crop.py</code>	Head-region crop helper used by script 11
01	<code>01_dataset_overview.py</code>	What is in the dataset, and is the split clean?
02	<code>02_data_quality.py</code>	Corrupt frames, frozen frames, drift, short sessions
03	<code>03_temperature_analysis.py</code>	The temperature cue – and why it is not enough
04	<code>04_topology_analysis.py</code>	The shape cue – which features actually separate
05	<code>05_threshold_baselines.py</code>	What a single hand-set threshold achieves
06	<code>06_classical_models.py</code>	15 supervised models $\times$ 8 feature sets
07	<code>07_oneclass_models.py</code>	5 anomaly detectors, no attack data in training
08	<code>08_fused_detector.py</code>	The headline result – two cues combined
09	<code>09_unseen_attack.py</code>	Performance against an attack family never trained on
10	<code>10_cross_participant.py</code>	The key limitation – transfer to a new person
11	<code>11_face_crop_study.py</code>	The torso confound, and why cropping is inconclusive
12	<code>12_neural_networks.py</code>	Custom MLP / CNN / ResNet trained from scratch
13	<code>13_autoencoder.py</code>	Reconstruction-error anomaly detection
14	<code>14_transfer_learning.py</code>	Eleven frozen pretrained ImageNet backbones
15	<code>15_deployment_simulation.py</code>	Windowing, retries, model size, latency
16	<code>16_figures.py</code>	Every figure, generated from the result tables
17	<code>17_report_tables.py</code>	Ranked comparison table and headline numbers
18	<code>18_feature_correlation.py</code>	Are the twelve topology features redundant?

C.1 Key parameters

- **Sensor.** Melexis EVB90640-41 evaluation board fitted with an MLX90641 (16×12); emissivity 1.000, frame rate 2 fps, ADC resolution 2, interleaved reading pattern, de-interlace, IIR and TGC filters enabled (Section 3.1).
- **Capture.** Approximately 60 seconds per scenario at 30–50 cm, yielding roughly 120 frames, one CSV per scenario (Section 3.2).
- **Live band.** Mean peak temperature of bona fide frames ± 2 standard deviations, giving 34.46–36.24°C, chosen a priori rather than tuned (Section 5.1).

- **Validation.** Session-grouped k -fold cross-validation (`GroupKFold` on the session label) enforced centrally in `common.py`; no session appears in both a training and a test fold in any reported result (Section 3.6).
- **Seeds.** Fixed via `common.set_seed()`, which pins `numpy`, `random` and `torch` (Section 3.7).

C.2 The three rules the suite enforces

1. **Validation is grouped by session, always.** Frames within one 60-second recording are highly correlated; a random frame-level split leaks information between train and test and produces an ACER near 0.000, which is meaningless. This was a real error made early in the project and is the most important methodological point in the work (Section 3.6).
2. **Seeds are fixed.** `EllipticEnvelope` uses a randomised minimum-covariance estimate and `IsolationForest` subsamples; without a fixed seed the reported ACER moves by about 0.005 between runs – enough to change a ranking. Section 5.3 identifies feature redundancy as the likely reason that instability is as large as it is.
3. **Parameters are not chosen by reading the evaluation metric.** Script 11 sweeps the crop-window size and reports the whole sweep rather than the best cell (Section 5.8); script 08 uses a two-standard-deviation band chosen a priori and reports sensitivity to that choice separately.

Appendix D

Complete Result Grids

This appendix reproduces every configuration evaluated, rather than the best rows quoted in Chapter 5. It is included so that the selection of reported results can be audited: the tables below are the complete output of the analysis suite, and each is generated directly from the corresponding CSV file in `results/tables/`.

All results use session-grouped five-fold cross-validation with fixed random seeds.

D.1 Classical supervised models

Fifteen classifiers evaluated on each of eight feature representations, 120 combinations in total, ranked by ACER.

Table D.1: Complete classical supervised model grid.

Model	Feature set	Dim.	APCER	BPCER	ACER	F1
QDA	norm+topo	204	0.200	0.116	0.158	0.828
QDA	norm+stats	215	0.217	0.127	0.172	0.812
DecisionTree	stats+topo	35	0.228	0.118	0.173	0.810
QDA	all	227	0.229	0.123	0.176	0.807
DecisionTree	stats	23	0.216	0.155	0.186	0.800
QDA	norm	192	0.266	0.119	0.193	0.786
SVC-rbf	stats	23	0.280	0.111	0.195	0.780
KNN-5	norm	192	0.228	0.164	0.196	0.789
GaussianNB	norm+stats	215	0.173	0.267	0.220	0.777
KNN-5	norm+topo	204	0.312	0.132	0.222	0.749
RidgeClassifier	norm+topo	204	0.354	0.091	0.223	0.739
KNN-5	raw	192	0.275	0.177	0.226	0.753
LogisticRegression	norm+topo	204	0.353	0.105	0.229	0.733
DecisionTree	topology	12	0.219	0.245	0.232	0.759

(continued overleaf)

(continued from previous page)

Model	Feature set	Dim.	APCER	BPCER	ACER	F1
GaussianNB	norm+topo	204	0.174	0.290	0.232	0.767
GaussianNB	all	227	0.189	0.282	0.235	0.762
KNN-5	norm+stats	215	0.340	0.135	0.237	0.729
LDA	norm+topo	204	0.391	0.099	0.245	0.708
AdaBoost	stats+topo	35	0.370	0.125	0.248	0.711
GaussianNB	stats+topo	35	0.350	0.147	0.248	0.716
GaussianNB	stats	23	0.346	0.155	0.251	0.715
MLP	stats+topo	35	0.366	0.141	0.253	0.707
GaussianNB	topology	12	0.354	0.153	0.254	0.710
DecisionTree	norm+topo	204	0.285	0.223	0.254	0.727
RandomForest	stats+topo	35	0.364	0.146	0.255	0.707
LinearSVC	norm+topo	204	0.426	0.095	0.260	0.683
RandomForest	stats	23	0.357	0.164	0.260	0.704
KNN-5	all	227	0.402	0.133	0.267	0.685
QDA	raw	192	0.412	0.125	0.269	0.680
AdaBoost	stats	23	0.421	0.120	0.271	0.675
LDA	raw	192	0.249	0.294	0.271	0.722
QDA	stats	23	0.419	0.123	0.271	0.676
RidgeClassifier	raw	192	0.267	0.276	0.272	0.717
QDA	stats+topo	35	0.407	0.140	0.274	0.677
LinearSVC	all	227	0.463	0.086	0.274	0.658
RandomForest	topology	12	0.312	0.249	0.280	0.699
LDA	norm+stats	215	0.476	0.091	0.284	0.644
SVC-rbf	stats+topo	35	0.433	0.135	0.284	0.660
RidgeClassifier	norm	192	0.367	0.202	0.284	0.680
ExtraTrees	topology	12	0.356	0.214	0.285	0.683
RidgeClassifier	norm+stats	215	0.466	0.104	0.285	0.647
LogisticRegression	all	227	0.459	0.111	0.285	0.649
RidgeClassifier	all	227	0.482	0.091	0.286	0.639
SGDClassifier	stats+topo	35	0.405	0.171	0.288	0.665
LinearSVC	stats+topo	35	0.419	0.158	0.288	0.661
GradientBoosting	topology	12	0.299	0.281	0.290	0.694
LDA	norm	192	0.346	0.234	0.290	0.681
MLP	stats	23	0.458	0.124	0.291	0.644
LogisticRegression	stats+topo	35	0.440	0.144	0.292	0.651
ExtraTrees	stats+topo	35	0.441	0.143	0.292	0.650

(continued overleaf)

(continued from previous page)

Model	Feature set	Dim.	APCER	BPCER	ACER	F1
GradientBoosting	stats+topo	35	0.442	0.146	0.294	0.648
GaussianNB	norm	192	0.319	0.269	0.294	0.686
LDA	all	227	0.515	0.073	0.294	0.618
SVC-rbf	norm	192	0.455	0.135	0.295	0.643
DecisionTree	norm	192	0.340	0.251	0.295	0.680
MLP	norm+stats	215	0.431	0.159	0.295	0.650
GradientBoosting	stats	23	0.442	0.149	0.295	0.647
SVC-rbf	norm+topo	204	0.456	0.142	0.299	0.639
LogisticRegression	stats	23	0.476	0.121	0.299	0.631
LinearSVC	norm+stats	215	0.505	0.095	0.300	0.618
SVC-rbf	topology	12	0.352	0.251	0.301	0.671
SVC-rbf	norm+stats	215	0.463	0.144	0.303	0.632
LinearSVC	raw	192	0.431	0.177	0.304	0.643
MLP	norm+topo	204	0.418	0.195	0.306	0.646
GradientBoosting	all	227	0.423	0.193	0.308	0.643
LogisticRegression	norm+stats	215	0.494	0.125	0.309	0.615
MLP	norm	192	0.407	0.213	0.310	0.647
SVC-rbf	all	227	0.475	0.146	0.310	0.621
LogisticRegression	raw	192	0.405	0.221	0.313	0.645
GradientBoosting	norm+stats	215	0.429	0.205	0.317	0.633
QDA	topology	12	0.481	0.155	0.318	0.613
RidgeClassifier	stats	23	0.495	0.143	0.319	0.606
LDA	stats	23	0.488	0.156	0.322	0.607
MLP	topology	12	0.351	0.303	0.327	0.652
MLP	all	227	0.463	0.194	0.328	0.612
RandomForest	norm	192	0.542	0.115	0.329	0.577
KNN-5	stats+topo	35	0.529	0.131	0.330	0.582
AdaBoost	norm+stats	215	0.521	0.140	0.331	0.585
AdaBoost	all	227	0.522	0.139	0.331	0.585
RidgeClassifier	stats+topo	35	0.503	0.159	0.331	0.593
ExtraTrees	stats	23	0.526	0.137	0.331	0.582
LinearSVC	stats	23	0.532	0.132	0.332	0.579
LDA	stats+topo	35	0.483	0.182	0.333	0.600
LogisticRegression	norm	192	0.475	0.194	0.335	0.602
AdaBoost	topology	12	0.312	0.360	0.336	0.657
KNN-5	topology	12	0.397	0.278	0.338	0.629

(continued overleaf)

(continued from previous page)

Model	Feature set	Dim.	APCER	BPCER	ACER	F1
SGDClassifier	topology	12	0.322	0.357	0.340	0.651
SGDClassifier	norm+topo	204	0.562	0.121	0.341	0.557
SGDClassifier	norm+stats	215	0.499	0.184	0.342	0.586
ExtraTrees	norm	192	0.560	0.128	0.344	0.555
LinearSVC	norm	192	0.553	0.144	0.349	0.555
LDA	topology	12	0.375	0.324	0.350	0.628
LinearSVC	topology	12	0.340	0.361	0.350	0.639
RidgeClassifier	topology	12	0.376	0.329	0.353	0.625
SGDClassifier	all	227	0.513	0.193	0.353	0.571
LogisticRegression	topology	12	0.341	0.371	0.356	0.635
RandomForest	norm+topo	204	0.588	0.133	0.360	0.528
RandomForest	all	227	0.596	0.144	0.370	0.516
RandomForest	norm+stats	215	0.604	0.141	0.372	0.509
KNN-5	stats	23	0.617	0.132	0.374	0.500
ExtraTrees	norm+topo	204	0.620	0.132	0.376	0.497
SGDClassifier	stats	23	0.613	0.147	0.380	0.498
SGDClassifier	norm	192	0.540	0.241	0.390	0.532
GradientBoosting	norm+topo	204	0.574	0.207	0.391	0.513
DecisionTree	all	227	0.451	0.333	0.392	0.570
ExtraTrees	all	227	0.643	0.146	0.395	0.469
MLP	raw	192	0.551	0.241	0.396	0.522
AdaBoost	norm+topo	204	0.578	0.218	0.398	0.506
SGDClassifier	raw	192	0.545	0.252	0.399	0.523
ExtraTrees	norm+stats	215	0.659	0.146	0.403	0.453
DecisionTree	norm+stats	215	0.534	0.313	0.423	0.512
DecisionTree	raw	192	0.659	0.208	0.434	0.432
RandomForest	raw	192	0.740	0.167	0.454	0.359
GradientBoosting	norm	192	0.673	0.237	0.455	0.410
ExtraTrees	raw	192	0.766	0.151	0.459	0.333
SVC-rbf	raw	192	0.753	0.184	0.469	0.340
AdaBoost	norm	192	0.703	0.237	0.470	0.379
AdaBoost	raw	192	0.761	0.226	0.493	0.320
GaussianNB	raw	192	0.687	0.325	0.506	0.373
GradientBoosting	raw	192	0.740	0.432	0.586	0.298

D.2 One-class detectors

Five anomaly detectors across eight feature representations, fitted on live frames only. Feature sets wider than forty dimensions are reduced by PCA fitted on the training live frames, since a full covariance estimate is unreliable at that width.

Table D.2: Complete one-class detector grid.

Model	Feature set	Dim.	Reduced	APCER	BPCER	ACER
IsolationForest	topology	12	none	0.134	0.148	0.141
EllipticEnvelope	stats+topo	35	none	0.000	0.352	0.176
IsolationForest	stats+topo	35	none	0.004	0.446	0.225
OneClassSVM	topology	12	none	0.046	0.509	0.277
EllipticEnvelope	stats	23	none	0.000	0.568	0.284
EllipticEnvelope	topology	12	none	0.459	0.133	0.296
IsolationForest	stats	23	none	0.011	0.585	0.298
OneClassSVM	stats	23	none	0.009	0.629	0.319
EllipticEnvelope	norm+stats	215	PCA20	0.528	0.115	0.322
OneClassSVM	stats+topo	35	none	0.004	0.674	0.339
IsolationForest	norm+stats	215	PCA20	0.464	0.217	0.340
EllipticEnvelope	norm+topo	204	PCA20	0.548	0.134	0.341
EllipticEnvelope	norm	192	PCA20	0.557	0.125	0.341
EllipticEnvelope	all	227	PCA20	0.570	0.132	0.351
IsolationForest	all	227	PCA20	0.406	0.307	0.356
OneClassSVM	norm+topo	204	PCA20	0.228	0.559	0.393
OneClassSVM	norm	192	PCA20	0.270	0.534	0.402
LocalOutlierFactor	stats	23	none	0.025	0.781	0.403
LocalOutlierFactor	topology	12	none	0.101	0.705	0.403
IsolationForest	norm+topo	204	PCA20	0.553	0.283	0.418
GaussianMixture	topology	12	none	0.000	0.839	0.419
GaussianMixture	all	227	PCA20	0.000	0.890	0.445
LocalOutlierFactor	stats+topo	35	none	0.021	0.871	0.446
GaussianMixture	norm+stats	215	PCA20	0.000	0.899	0.450
IsolationForest	norm	192	PCA20	0.712	0.189	0.450
OneClassSVM	all	227	PCA20	0.247	0.655	0.451
OneClassSVM	norm+stats	215	PCA20	0.253	0.651	0.452
GaussianMixture	norm	192	PCA20	0.000	0.953	0.477
GaussianMixture	stats	23	none	0.000	0.958	0.479
GaussianMixture	norm+topo	204	PCA20	0.000	0.997	0.498

(continued overleaf)

(continued from previous page)

Model	Feature set	Dim.	Reduced	APCER	BPCER	ACER
GaussianMixture	stats+topo	35	none	0.000	1.000	0.500
GaussianMixture	raw	192	PCA20	0.083	0.921	0.502
IsolationForest	raw	192	PCA20	0.636	0.388	0.512
LocalOutlierFactor	all	227	PCA20	0.335	0.725	0.530
EllipticEnvelope	raw	192	PCA20	0.855	0.244	0.549
LocalOutlierFactor	norm+topo	204	PCA20	0.450	0.683	0.567
LocalOutlierFactor	norm+stats	215	PCA20	0.447	0.700	0.573
LocalOutlierFactor	raw	192	PCA20	0.331	0.835	0.583
OneClassSVM	raw	192	PCA20	0.515	0.685	0.600
LocalOutlierFactor	norm	192	PCA20	0.585	0.665	0.625

D.3 Cross-participant generalisation

Every combination of model and feature set, fitted on one participant’s live frames and tested on the other participant together with all attacks. BPCER here is the share of an unseen genuine user’s frames that would be rejected; a value of 1.000 means every frame of a real person was rejected.

Table D.3: Complete cross-participant grid.

Train	Test	Feature set	Model	APCER	BPCER	ACER
2	1	all	EllipticEnvelope	0.000	0.333	0.167
2	1	norm+topo	EllipticEnvelope	0.000	0.354	0.177
2	1	norm+topo	IsolationForest	0.000	0.418	0.209
2	1	all	IsolationForest	0.000	0.716	0.358
1	2	norm+topo	IsolationForest	0.306	0.442	0.374
2	1	topology	IsolationForest	0.013	0.759	0.386
1	2	all	IsolationForest	0.238	0.683	0.461
1	2	norm+topo	EllipticEnvelope	0.486	0.471	0.478
1	2	stats	EllipticEnvelope	0.000	1.000	0.500
1	2	stats	OneClassSVM	0.000	1.000	0.500
1	2	stats+topo	EllipticEnvelope	0.000	1.000	0.500
1	2	stats+topo	OneClassSVM	0.000	1.000	0.500
1	2	all	OneClassSVM	0.000	1.000	0.500
2	1	stats	EllipticEnvelope	0.000	1.000	0.500
2	1	stats	IsolationForest	0.000	1.000	0.500

(continued overleaf)

(continued from previous page)

Train	Test	Feature set	Model	APCER	BPCER	ACER
2	1	stats	OneClassSVM	0.000	1.000	0.500
2	1	topology	OneClassSVM	0.000	1.000	0.500
2	1	norm+topo	OneClassSVM	0.000	1.000	0.500
2	1	stats+topo	EllipticEnvelope	0.000	1.000	0.500
2	1	stats+topo	IsolationForest	0.000	1.000	0.500
2	1	stats+topo	OneClassSVM	0.000	1.000	0.500
2	1	all	OneClassSVM	0.000	1.000	0.500
1	2	topology	OneClassSVM	0.004	1.000	0.502
1	2	stats+topo	IsolationForest	0.008	1.000	0.504
2	1	topology	EllipticEnvelope	0.008	1.000	0.504
1	2	stats	IsolationForest	0.011	1.000	0.506
1	2	norm+topo	OneClassSVM	0.030	1.000	0.515
1	2	topology	IsolationForest	0.048	0.982	0.515
1	2	all	EllipticEnvelope	0.370	0.721	0.546
1	2	topology	EllipticEnvelope	0.138	1.000	0.569

D.4 Unseen attack families

Leave-one-attack-family-out evaluation. The one-class fused rows require no attack data at all, so every family is unseen by construction; the supervised rows are trained on the remaining three families.

Table D.4: Complete unseen-attack grid.

Approach	Features	Held-out family	APCER	BPCER	ACER
one-class fused	—	screen	0.000	0.146	0.073
one-class fused	—	heat_decoy	0.004	0.146	0.075
one-class fused	—	combo	0.008	0.146	0.077
one-class fused	—	print	0.130	0.146	0.138
GaussianNB	all	heat_decoy	0.000	0.756	0.378
GaussianNB	all	screen	0.143	0.764	0.453
GaussianNB	stats	heat_decoy	0.000	0.909	0.455
GaussianNB	all	combo	0.000	0.917	0.459
GaussianNB	stats	combo	0.000	1.000	0.500
GaussianNB	topology	combo	0.000	1.000	0.500
GaussianNB	all	print	0.077	0.941	0.509

(continued overleaf)

(continued from previous page)

Approach	Features	Held-out family	APCER	BPCER	ACER
GaussianNB	topology	heat_decoy	0.035	0.992	0.513
GaussianNB	stats	screen	0.029	1.000	0.514
GaussianNB	topology	print	0.146	1.000	0.573
GaussianNB	topology	screen	0.261	1.000	0.631
GaussianNB	stats	print	0.272	1.000	0.636
LogisticRegression	stats	combo	0.008	0.000	0.004
LogisticRegression	all	combo	0.000	0.018	0.009
LogisticRegression	stats	screen	0.306	0.002	0.154
LogisticRegression	all	screen	0.351	0.040	0.196
LogisticRegression	stats	print	0.423	0.002	0.212
LogisticRegression	all	print	0.496	0.012	0.254
LogisticRegression	topology	heat_decoy	0.957	0.000	0.478
LogisticRegression	stats	heat_decoy	1.000	0.000	0.500
LogisticRegression	topology	combo	0.000	1.000	0.500
LogisticRegression	all	heat_decoy	1.000	0.008	0.504
LogisticRegression	topology	print	0.008	1.000	0.504
LogisticRegression	topology	screen	0.526	1.000	0.763
RandomForest	topology	heat_decoy	0.723	0.006	0.364
RandomForest	all	combo	0.762	0.036	0.399
RandomForest	all	print	0.382	0.551	0.467
RandomForest	stats	heat_decoy	0.941	0.002	0.472
RandomForest	all	heat_decoy	0.984	0.004	0.494
RandomForest	topology	combo	0.008	1.000	0.504
RandomForest	all	screen	0.465	0.568	0.516
RandomForest	stats	screen	0.106	0.998	0.552
RandomForest	topology	print	0.114	1.000	0.557
RandomForest	stats	combo	0.541	0.618	0.580
RandomForest	stats	print	0.268	0.980	0.624
RandomForest	topology	screen	0.547	1.000	0.773
SVC-rbf	all	combo	0.000	0.553	0.277
SVC-rbf	all	screen	0.200	0.475	0.337
SVC-rbf	stats	screen	0.065	0.784	0.425
SVC-rbf	topology	heat_decoy	0.879	0.006	0.443
SVC-rbf	stats	combo	0.000	1.000	0.500
SVC-rbf	topology	combo	0.000	1.000	0.500
SVC-rbf	all	heat_decoy	1.000	0.000	0.500

(continued overleaf)

(continued from previous page)

Approach	Features	Held-out family	APCER	BPCER	ACER
SVC-rbf	topology	print	0.037	1.000	0.518
SVC-rbf	all	print	0.557	0.487	0.522
SVC-rbf	stats	print	0.276	0.774	0.525
SVC-rbf	stats	heat_decoy	0.480	0.707	0.594
SVC-rbf	topology	screen	0.404	0.992	0.698

D.5 Crop-window sensitivity sweep

Sixteen crop window sizes against the full-frame baseline. This table is reproduced in full precisely because the best cell must not be selected from it: the spread across window sizes exceeds any gain from cropping.

Table D.5: Complete crop-window sweep.

Crop window	APCER	BPCER	ACER	Accuracy
full frame	0.039	0.146	0.092	0.904
6x8	0.083	0.189	0.136	0.861
6x9	0.008	0.117	0.063	0.934
6x10	0.007	0.209	0.108	0.886
6x12	0.033	0.121	0.077	0.920
7x8	0.153	0.203	0.178	0.821
7x9	0.074	0.139	0.106	0.892
7x10	0.077	0.173	0.125	0.872
7x12	0.043	0.107	0.075	0.923
8x8	0.220	0.142	0.181	0.822
8x9	0.002	0.113	0.058	0.939
8x10	0.127	0.132	0.129	0.871
8x12	0.028	0.121	0.074	0.923
9x8	0.190	0.156	0.173	0.828
9x9	0.113	0.183	0.148	0.850
9x10	0.131	0.145	0.138	0.862
9x12	0.011	0.136	0.073	0.923

D.6 Pretrained backbones

All eleven frozen ImageNet backbones with a logistic-regression head trained on the thermal data. Backbone weights were not updated.

Table D.6: Complete pretrained-backbone grid.

Backbone	Embed dim.	Parameters	APCER	BPCER	ACER
GoogLeNet	1024	5,599,904	0.228	0.013	0.120
ResNet50	2048	23,508,032	0.221	0.029	0.125
DenseNet121	1024	6,953,856	0.259	0.014	0.137
EfficientNetB0	1280	4,007,548	0.281	0.025	0.153
VGG11	25088	9,220,480	0.220	0.093	0.156
SqueezeNet1_1	86528	722,496	0.238	0.090	0.164
ViT-B/16	768	85,798,656	0.320	0.025	0.172
Swin-Tiny	768	27,519,354	0.345	0.026	0.186
ResNet18	512	11,176,512	0.373	0.020	0.197
MobileNetV3	576	927,008	0.403	0.023	0.213
ConvNeXt-Tiny	768	27,818,592	0.434	0.067	0.251