Toward Reliable and Interpretable Tactile Material Classification

Bridging Human Perception and Deep Learning

D. Hogendoorn

TU Delft

Toward Reliable and Interpretable Tactile Material Classification

Bridging Human Perception and Deep Learning

by

D. Hogendoorn

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday April 23, 2026 at 09:30 AM

Student number: 4880137
Project duration: August 25, 2025 – April 23, 2026
Thesis committee:

Dr. Y. Vardar,	TU Delft, Chair and Supervisor
Dr. J. C. F. de Winter,	TU Delft, External member
Dr. L. Zou,	External member

TU Delft

Preface

Touch is one of the most informative human senses, yet many of today's interfaces remain largely limited to flat, uniform surfaces. During a lab tour at TU Delft I was introduced to an electrovibration touchscreen that could modulate friction and create tactile sensations on glass. That moment made haptics feel both technically fascinating and surprisingly tangible, and it planted the seed for the topic of this thesis.

I am a Robotics MSc student at TU Delft, and throughout my studies I increasingly gravitated toward machine perception: how sensors capture the world, how signals can be fused, and how learning-based models can extract structure from messy, real data. Courses in machine learning and deep learning provided the foundation that ultimately shaped the approach taken in this work. This thesis sits at the intersection of those interests: tactile sensing, multimodal signals, and deep-learning models that can both perform well and be interpreted.

Writing this thesis was an intense and sometimes unpredictable process. There were phases where experiments ran smoothly and results were clear, but also weeks where small implementation details, data issues, and long training or interpretability runs slowed progress more than expected. Those moments were also where I learned the most: to work systematically, to document decisions, and to keep pushing until the story became coherent.

Outside of university, frisbee was my consistent way to reset. Whenever I was in a dip, a training or a game could lift my mood in an instant. I would like to thank all my frisbee friends for the energy, and fun during this period. In particular, I want to thank Lenny, Elise, Jop, and Annemarijn for always being a listening ear and for supporting me in whatever I needed at the time. I am also deeply grateful to my family for their patience and continued support throughout my studies. A special thank you goes to Mia, who kept me cheerful throughout the entire process and was probably my biggest believer. Whenever things felt overwhelming, she was always ready to remind me that I could do it, and that encouragement made a bigger difference than I can properly put into words.

Finally, I would like to thank my supervisors and lab group. Thank you to Yasemin for the guidance, feedback, and trust throughout the project, and for helping steer the work toward a clear research direction. A huge thank you to Li as well, whose help and input were of great value during the beginning stages of the thesis. Lastly, thank you to everyone in Hitlab for the useful discussions we had during our weekly meetings.

D. Hogendoorn
Delft, April 2026

Abstract

Deep learning has led to strong performance in recognizing materials from touch signals, but it is often difficult to understand which parts of those signals drive a model's decision. This thesis studies accurate and interpretable tactile material classification on the public SENS3 dataset using three dominant cue sources: thermal transients during static contact, deformation dynamics during pressing, and friction/vibration cues during sliding. An adjective-mediated pipeline that first predicts psychophysical rating distributions and then classifies materials is compared against a direct multimodal classifier that fuses modality-specific encoders. The direct multimodal model shows strong generalization, reaching 0.896 test accuracy across seven retained material classes. To better understand the learned decision process, Integrated Gradients was used as the main explanation method, combined with Temporal Saliency Rescaling for models with a single classification output. The resulting attribution maps were then summarized into contributions at the level of signal type, interaction phase or bin, and individual measurement channel. The resulting explanations reveal class-conditional cue usage aligned with interaction structure: transient thermal phases dominate for metal-like materials, pressing dynamics contribute strongly for compliant/textile-like classes, and sliding evidence concentrates in low-force regimes where friction and vibration cues are most informative. Overall, the results demonstrate that high-performing tactile material recognition can be combined with interpretable, physically grounded attribution summaries, improving trust in model decisions for haptic interfaces and embodied systems.

Contents

1

Introduction

Digital interfaces have become increasingly rich in vision and sound. Images are highly realistic, video is cinematic, and spatial audio can create a strong sense of immersion. Yet when a user physically interacts with most devices, the experience is still largely limited to smooth glass surfaces. Touch therefore remains one of the least developed senses in modern interfaces. This matters because tactile information helps distinguish not only what an object looks like, but also how it feels. Recent work shows that tactile signals can increase perceived presence and user engagement, and can enhance memory for the experience (Gibbs et al., 2022).

However, providing such benefits in a material-specific way requires more than adding arbitrary haptic feedback. We need to capture and interpret the tactile interaction signals produced by finger–material contact and motion, because important material properties depend on contact mechanics and heat transfer and are not reliably determined from vision or audio alone. If we want interactive systems that feel as informative as they look and sound, we must understand how to capture and interpret the tactile signatures of real materials.

When a fingertip contacts and moves across a surface, the recorded interaction signals capture the skin–material processes that underlie tactile perception. Three sources dominate: (i) mechanical signals (deformation, compliance, micro-geometry), (ii) thermal signals (transient heat flow at contact), and (iii) friction and vibration signals during sliding (frictional regimes and texture-induced vibrations). In the first few hundred milliseconds after contact, temperature and heat-flux change rapidly and provide a strong, time-localised signal; during sustained contact and sliding, friction and vibration provide complementary information about surface texture (Bergmann Tiest, 2010; Felicetti et al., 2023).

Modelling these interaction signals is difficult with hand-crafted rules because the measurements depend on many interacting factors. The same material can produce different traces depending on contact force, sliding speed, and individual finger properties, and these exploration conditions can emphasise different texture signals, as shown by Fishel and Loeb (2012).

In such settings, writing rules for all relevant combinations of force, speed, and user-dependent effects quickly becomes impractical. In general, when the correct mapping from measurements to outcomes is complex, it is often more effective to learn it from examples than to hand-code rules for all possible cases (Jordan and Mitchell, 2015). Deep learning supports this by learning useful features directly from the raw, high-dimensional signals and has proven effective across many perception domains (LeCun et al., 2015).

In tactile perception research, several deep-learning backbone families are commonly used for time-series interaction data, each with a different inductive bias. One-dimensional convolutional neural networks are efficient and well suited to capturing local temporal patterns such as short-lived transients and vibration bursts. Long Short-Term Memory models explicitly model sequential dependencies and can integrate information over longer time spans. For this reason, both backbone families are considered in this thesis as part of the broader model-development process. However, the main focus is on developing and interpreting multimodal classifiers for tactile material recognition, with the backbone comparison serving as a supplementary architectural check rather than as the central contribution. Much prior work in tactile perception focuses on vision-style tactile modalities, such as GelSight, which

emphasize surface geometry but under-represent finger–material interaction dynamics (Devillard et al., 2023; Luo et al., 2018; Zhang et al., 2021).

To connect model decisions to what people actually perceive, psychophysical descriptors are also needed. However, much of the prior work either binarizes such attributes, for example “soft” versus “not soft” (Cao et al., 2023), or predicts continuous or distributional ratings from vision alone or from robot–material interaction rather than from finger–material interaction signals themselves (Hassan et al., 2023; Richardson and Kuchenbecker, 2020). This leaves open the question of how well human perceptual ratings can be recovered directly from the same interaction signals that are used for tactile material recognition.

A further challenge is interpretability. Deep neural networks are often criticized as black boxes, and a range of interpretability techniques has therefore been proposed to identify which parts of a time series, which time segments, or which features influence a model’s decision (Bento et al., 2021; Ismail, Gunady, Bravo, and Feizi, 2020; Pham et al., 2022; Siddiqui et al., 2020; Wang et al., 2025). At the same time, Nauta et al. (2023) warn that current time-series explanations can be misleading when they rely on raw saliency or attention alone. Saliency may spread across whole timesteps, and attention weights do not necessarily reflect true importance. Explanations should therefore be validated and connected to physically meaningful events. Following recent work (Räuker et al., 2023), interpretability is treated here as a post-hoc, behaviour-grounded check: explanations should be faithful to the model and useful for understanding finger–material physics. This thesis therefore focuses explicitly on post-hoc analysis rather than on probing inner mechanisms during training.

The broader aim of this thesis is to understand how material identity and perceived material qualities are encoded in finger–material interaction signals. This is a necessary step before such signals can be used reliably in applications such as haptic rendering, tactile interfaces, or embodied systems. To address this aim, the thesis studies three connected problems: material recognition from interaction signals, prediction of psychophysical descriptors from those same signals, and interpretability of the resulting models in terms of meaningful phases and signal channels. Concretely, this requires (a) robust material classifiers from interaction signals, (b) prediction of psychophysical descriptors such as rough–smooth and hot–cold as full rating distributions, and (c) interpretable models that reveal which parts of the interaction drive the predictions.

Concretely, this thesis addresses the following research questions:

RQ1. To what extent can signals from different interaction modalities (thermal, pressing and sliding data) be used to classify materials from the SENS3 dataset?

RQ2. How accurately can full psychophysical adjective rating distributions (e.g. *hot–cold*, *hard–soft*) be predicted from these interaction signals?

RQ3. Which temporal phases and signal channels across these modalities most strongly drive the models’ decisions, and do these patterns correspond to physically meaningful events in finger–material interaction?

Within this scope, the thesis makes three main contributions. First, it provides a cross-validated benchmark for tactile material classification on SENS3, evaluating unimodal and late-fusion multimodal models under a consistent training and evaluation protocol. As part of the broader architecture exploration, both 1D-CNN and LSTM backbones are considered, with the final main models based on modality-specific 1D-CNN encoders. Second, it develops a distributional modelling approach for psychophysical adjective ratings, predicting full rating histograms from interaction signals and training these predictions against empirical rating distributions using Kullback–Leibler divergence. Third, it introduces a phase- and channel-aware attribution analysis pipeline based on Integrated Gradients with temporal saliency rescaling, aggregating time–channel attributions into structured summaries that reveal which modalities, contact phases or bins, and sensor channels drive the model predictions.

2

Background

This chapter introduces the main concepts that form the foundation of this thesis. Section 2.1 presents 1D-convolutional neural networks (1D-CNNs) and explains why they are well suited to modelling local temporal structure in tactile signals. Section 2.2 describes Long Short-Term Memory (LSTM) networks as an alternative sequence-modeling approach that can capture longer-range temporal dependencies. Finally, Section 2.3 introduces interpretability for time-series models, with a focus on Integrated Gradients and Temporal Saliency Rescaling, which are used later in this thesis to analyse the learned models.

2.1. 1D-Convolutional Neural Networks (1D-CNNs) for time-series signals

A one-dimensional convolutional neural network (1D CNN) is a specialized deep-learning architecture designed to operate on sequential data, such as time series, signals, or other one-dimensional data streams. In contrast to the 2D convolutions used for image data, where a kernel moves over two dimensions (height and width), a 1D convolutional layer slides a kernel along a single dimension (e.g., time or spatial sequence) to detect temporal or sequential patterns. More specifically, a vector of learnable weights is moved over the input sequence, performing a dot product at each position. This produces a transformed feature map (another sequence), highlighting local patterns such as spikes, oscillations, or repetitive motifs in the input. For a multichannel sequence $x \in \mathbb{R}^{T\times C}$ and a kernel $W^{(k)} \in \mathbb{R}^{K\times C}$ with bias $b^{(k)}$, the k-th feature map at time index t is

$$y_t^{(k)} = \sigma\Big(\sum_{i=0}^{K-1}\sum_{c=1}^{C} W_{i,c}^{(k)}\, x_{t+i-\Delta,\,c} + b^{(k)}\Big), \tag{2.1}$$

where K is the kernel width, Δ sets the padding, and $\sigma(\cdot)$ is a non-linearity (e.g., ReLU). Because the filter weights are shared across positions (the kernel is applied repeatedly across the input), 1D CNNs are efficient in terms of the number of parameters and naturally handle multivariate time-series data. Stacking layers lets the network build up from brief transients to longer dynamics. This combination makes 1D-CNNs effective for a wide range of sensor streams and practical for real-time use (Cacciari and Ranfagni, 2024; Ige and Sibiya, 2024).

In practice, temporal convolutions are paired with normalization, dropout, and pooling to stabilize learning and improve generalization. Batch Normalization is one of the more popular normalization techniques. It simply rescales each feature map using batch statistics,

$$\hat{z} = \frac{z-\mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2+\varepsilon}}, \qquad \mathrm{BN}(z) = \gamma\,\hat{z} + \beta, \tag{2.2}$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}^2$ are the batch mean and variance, and (γ, β) are learned parameters (Ioffe and Szegedy, 2015). This reduces covariate shift and enables deeper stacks to train reliably. Dropout

works by randomly setting a fraction p of the activations to zero during training. If h is the vector of activations, dropout produces

$$\tilde{h} = d \odot h,$$

where d is a vector of 0s and 1s in which approximately a fraction p of the entries are 0 (dropped) and the rest are 1 (kept). This helps prevent co-adaptation and improves generalization (Srivastava et al., 2014). Pooling reduces the temporal resolution of feature maps while retaining their most salient information and increasing the effective receptive field. For example, Global Average Pooling (GAP) replaces fully connected layers with a per–channel average,

$$h_{\text{GAP}}[k] = \frac{1}{T} \sum_{t=1}^{T} y_t^{(k)}, \tag{2.3}$$

which reduces parameters and limits overfitting (Lin et al., 2014). Max pooling, in contrast, takes the highest activation within each window,

$$h_{\text{max}}[k, t] = \max_{i \in \mathcal{W}(t)} y_i^{(k)},$$

preserving short, high-amplitude events, while average pooling smooths over longer trends (Boureau et al., 2010).

Compared with hand-engineered features or purely recurrent models, 1D-CNNs offer (i) parameter efficiency via weight sharing, (ii) stable, highly parallel training, and (iii) precise receptive-field control via kernel/stride/dilation choices. These properties suit thermo-tactile signals where short-lived, high-amplitude events coexist with slower trends: early layers specialize on localized transients, while deeper layers integrate evidence over hundreds of milliseconds.

Design rationale for this thesis. The tactile signals used in this thesis combine rapid, local events (e.g., heat-flux peaks at initial contact and vibration bursts during sliding) with slower trends (e.g., thermal equilibration and force ramps). A shallow 1D-CNN with small kernels provides a parameter-efficient way to detect such local motifs while progressively expanding the receptive field through stacking and pooling. This motivates the convolutional encoders used in the modality branches of Models 1 and 3.

2.2. Long Short-Term Memory (LSTM) networks for time-series signals

Recurrent neural networks (RNNs) are a class of sequence models that process inputs one time step at a time while maintaining a hidden state that serves as a dynamic memory of past inputs. At each step, the hidden state is updated as

$$h_t = \phi(W x_t + U h_{t-1} + b),$$

allowing the network to model temporal dependencies that extend beyond the local receptive fields of 1D-CNNs. RNNs are therefore worth mentioning because they provide a complementary mechanism for handling long-range structure in time-series data. However, plain RNNs suffer from vanishing and exploding gradients when dependencies span many steps. Long Short-Term Memory (LSTM) networks address this by introducing gated memory that regulates which information is stored, forgotten, or exposed, enabling stable learning over long horizons, important for sensor data where fast transients and slow dynamics coexist.

At each time t, given the input x_t, previous hidden state h_{t-1}, and previous cell state c_{t-1}, the LSTM computes

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad \text{(input gate: how much new information to write)} \tag{2.4}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad \text{(forget gate: how much past information to keep)} \tag{2.5}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad \text{(output gate: how much memory to reveal)} \tag{2.6}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad \text{(candidate memory update)} \tag{2.7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad \text{(new cell state: retained + added information)} \tag{2.8}$$

$$h_t = o_t \odot \tanh(c_t) \quad \text{(new hidden state: gated exposure of memory)} \tag{2.9}$$

where $\sigma(\cdot)$ is the logistic sigmoid and $\odot$ denotes elementwise multiplication. The forget gate f_t allows the cell state to carry information forward almost unchanged, providing a stable path for gradients and thus mitigating the vanishing-gradient problem (Hochreiter and Schmidhuber, 1997; Mienye et al., 2024).

Design rationale for this thesis. LSTMs provide a complementary inductive bias by explicitly modeling order and longer-range dependencies (e.g., press–hold–release dynamics or slow thermal drift). For this reason, LSTM backbones are evaluated alongside 1D-CNNs in this thesis. However, given the limited dataset size and the prominence of local transients in several modalities, convolutional encoders are used as the primary backbone in the final models, with recurrent variants retained as comparisons.

2.3. Interpretability for time-series

Deep models for sequential data often behave as a "black box," so we require post-hoc methods that (i) assign responsibility to *when* and *which* input features drove a prediction and (ii) remain faithful to the model's actual decision function. Two complementary tools used in this thesis are *Integrated Gradients* (IG; Sundararajan et al., 2017) and *Temporal Saliency Rescaling* (TSR; Ismail, Gunady, Corrada Bravo, and Feizi, 2020).

2.3.1. Integrated Gradients (IG).

Given a scalar target output $F(\mathbf{x})$ and a baseline $\mathbf{x}'$, IG attributes importance to each input dimension by integrating input gradients along the straight-line path between $\mathbf{x}'$ and $\mathbf{x}$:

$$\mathrm{IG}_i(\mathbf{x}) \;=\; (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F\big(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\big)}{\partial x_i}\, d\alpha. \tag{2.10}$$

In practice, the path integral is approximated by a Riemann sum with m steps. IG satisfies desirable axioms such as *implementation invariance* and *completeness* ($\sum_i \mathrm{IG}_i(\mathbf{x}) = F(\mathbf{x}) - F(\mathbf{x}')$), making it a principled choice for time-series attribution (Sundararajan et al., 2017). For multivariate sequences $x_{t,c}$ (time t, channel c), we obtain a time–channel attribution map $\mathrm{IG}_{t,c}$ and can aggregate it into phase- or channel-level summaries by averaging absolute attributions over sets of indices.

2.3.2. Temporal Saliency Rescaling (TSR).

Benchmarking on synthetic and real time-series tasks shows that many gradient-based saliency methods (including IG, SmoothGrad, DeepLIFT) often highlight whole time steps instead of showing which channels are responsible, resulting in widely distributed maps rather than clear, focused explanations (Ismail, Gunady, Corrada Bravo, and Feizi, 2020). TSR addresses this by a two-stage rescaling:

1. **Temporal scoring:** compute a timestep score τ_t (e.g., summing or norming saliency across channels), which estimates *when* the model finds evidence,

$$\tau_t \;\propto\; \phi\big(\{a_{t,c}\}_{c=1}^{C}\big), \qquad a_{t,c} \in \mathbb{R} \text{ (base saliency, e.g., IG).} \tag{2.11}$$

2. **Feature refinement:** redistribute per-timestep saliency across channels to emphasize *which* features matter within important times. One simple form is the normalized reweighting

$$\tilde{a}_{t,c} \;=\; \frac{|a_{t,c}|}{\sum_{c'} |a_{t,c'}| + \epsilon} \;\cdot\; \tau_t, \tag{2.12}$$

   which sharpens within-step attribution while preserving the global temporal profile.

Ismail et al. show that TSR improves precision/recall of signal-of-interest recovery on controlled benchmarks and yields crisper, more informative maps on real data, enabling nuanced temporal interpretations beyond raw saliency (Ismail, Gunady, Corrada Bravo, and Feizi, 2020).

Design rationale for this thesis. The interpretability goal is to link model evidence to *when* and *which* channels matter during finger–material interaction. IG provides time–channel attributions with the completeness property, while TSR is applied for scalar classification targets to sharpen channel responsibility within salient time steps.

3

Methods

This chapter describes the dataset, preprocessing steps, and model setups used in this thesis. Section 3.1 introduces the SENS3 data and explains how thermal, pressing, and sliding signals are represented. Section 3.2 describes the preprocessing pipeline, including filtering, normalization, and segmentation into fixed-length inputs. Section 3.3 presents the model architectures and training protocols for distribution prediction (Model 1), material classification (Model 3), and the intermediate adjective-to-material classifier (Model 2). Finally, Sections 3.3.5–3.3.7 describe the attribution-based interpretability approach and how attributions are summarized at the modality, channel, and phase/bin level.

3.1. Data

3.1.1. SENS3 dataset, materials, and participants

All experiments are based on the SENS3 dataset (Balasubramanian et al., 2024), which contains tactile and thermal recordings collected during multiple interaction tasks. In this thesis, three of those tasks, or 'modalities' are used:

(i) static contact for thermal equilibration,

(ii) controlled sliding for friction and vibration, and

(iii) normal-force pressing for indentation dynamics.

For thermal experiments, the skin temperature and heat-flux signals during static contact are used. For sliding, the normal force F_z, an effective friction coefficient $\mu(t)$ derived from the in-plane forces, and band-pass filtered accelerations (A_x, A_y, A_z) are used. For pressing, the indentation and normal force recorded during the press–release interaction are used.

The dataset contains recordings for 50 individual material instances (indexed 1–50). Each index is mapped to one of ten semantic material classes via a fixed lookup table (e.g., indices 1–10 correspond to Fabric; 11–13 and 16 to Sandpaper; 17–20 to Vinyl; etc.). For all models, we retain the seven reasonably well-represented classes *{Fabric, Foam, Metal, Paper, Sandpaper, Vinyl, Wood}* and exclude *{Rubber, Plastic, Leather}*, because these three classes are under-represented (see Figure 3.1) and led to unstable training and low prediction accuracy when included in initial experiments.

Thermal, pressing, and sliding recordings were not available at equal coverage. While thermal and pressing were collected for seven participants, only three participants completed the full sliding protocol across the retained material classes. For the main multimodal experiments in this thesis, and in particular for attribution-based interpretability, we therefore restrict training and evaluation to the subset of data where all three modalities are present. This way, each sample has the same set of modalities, so modalities can be compared fairly and attributions can be computed consistently.

In addition to this main setting, an exploratory performance-only experiment was performed using six participants. Because sliding was only fully available for two participants at that stage, sliding segments were distributed across the remaining participants to increase coverage. This approach is not suitable for interpretability analysis, since it breaks the strict correspondence between a participant's interaction

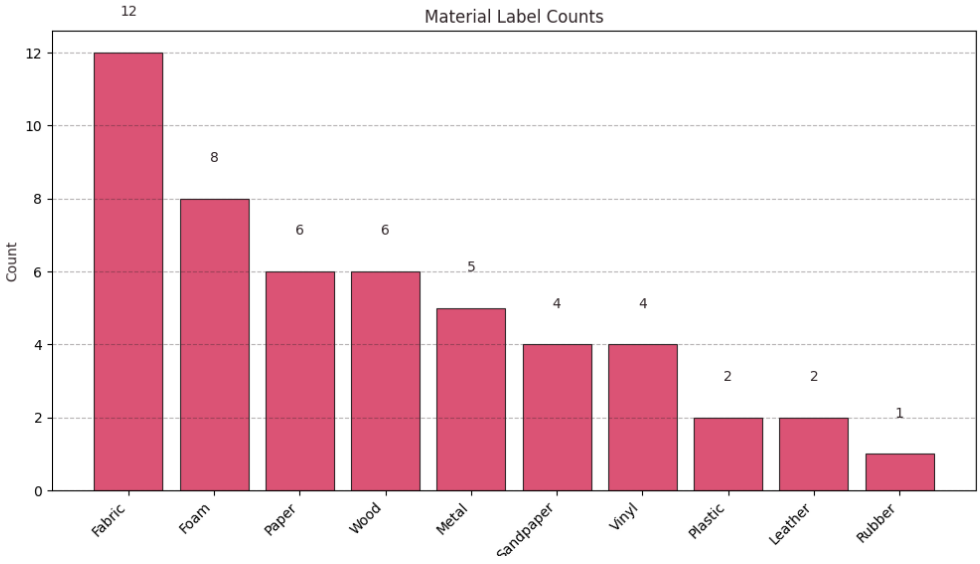


Figure 3.1: Classes in the Sens3 dataset of 50 materials before exclusion of the classes Plastic, Leather and Rubber

and their multimodal signals. However, it provides a useful indication of how classification performance changes when more participant data is included, and is therefore reported only as a supplementary performance comparison.

3.1.2. Psychophysical ratings

In addition to the sensor data, SENS3 includes psychophysical ratings collected in a separate experiment. Each material (indices 1–50) was rated by 20 participants on a set of adjective pairs using a discrete scale from 1 to 15.

The raw rating files contain eight columns corresponding to the adjective pairs rough–smooth, flat–bumpy, sticky–slippery, hot–cold, regular–irregular, fine–coarse, hard–soft, and wet–dry.

Although each participant rates the same materials, some people systematically give higher (or lower) scores. We reduce this rater bias by centering each participant's ratings to a shared reference mean before we pool them into a distribution.

Let $r_{u,m}$ denote the rating given by participant u for material m on a particular adjective, and let $\bar{r}_u$ be the mean of all ratings that participant u gave for that adjective. Let $\bar{r}_{\text{global}}$ be the mean of the rating scale (here, the mean of $\{1, \dots, 15\}$). Each individual rating is shifted so that the participant's average aligns with the global scale mean:

$$r_{u,m}^{\text{adj}} = \text{clip}\big(r_{u,m} - \bar{r}_u + \bar{r}_{\text{global}},\, 1, 15\big), \tag{3.1}$$

which recentres each participant's ratings around $\bar{r}_{\text{global}}$ while keeping values within the allowed range and preserving the relative ordering of materials for that participant. For example, if a participant tends to give very high scores with an average of $\bar{r}_u = 12$ and the global mean is $\bar{r}_{\text{global}} = 8$, then a rating of $r_{u,m} = 14$ is adjusted to $r_{u,m}^{\text{adj}} = 14 - 12 + 8 = 10$, and a rating of 8 is adjusted to $8 - 12 + 8 = 4$. Materials that this participant rated higher than others remain relatively higher after the shift, but their overall tendency to use the top end of the scale is removed.

For each material m, we combine the mean-corrected ratings from all participants and count how often each score $k \in \{1, \dots, 15\}$ occurs. Let $c_k(m)$ be the number of ratings equal to k. To avoid empty bins (which can cause numerical issues later), we add a small constant to every count:

$$\tilde{c}_k(m) = c_k(m) + \alpha, \tag{3.2}$$

with $\alpha = 0.1$. We then convert these counts into a probability distribution over the 15-point scale by normalising:

$$p_k(m) = \frac{\tilde{c}_k(m)}{\sum_{j=1}^{15} \tilde{c}_j(m)}. \tag{3.3}$$

This results in one empirical distribution $p(m) \in \Delta^{15}$ per material and adjective. This distribution is used as the training target for the adjective-prediction model (Model 1), and it reduces the effect of participants who consistently use higher or lower parts of the scale.

The final set of adjective pairs was selected after an exploratory inspection of the empirical class-level rating distributions, before training the models. Two criteria were used: (i) the adjective should show meaningful between-material variation in its target distributions, so that it could plausibly support the downstream material classification task, and (ii) the adjective should be reasonably related to the measured interaction signals. Based on these criteria, we retained *hot–cold*, *hard–soft*, *rough–smooth*, and *sticky–slippery*. We excluded *fine–coarse*, *flat–bumpy*, and *regular–irregular* because their empirical distributions were dominated by a similar central peak across nearly all retained material classes, with only limited class-specific variation. Although such targets might still be predictable for Model 1, they provide little discriminative structure for Model 2. We also excluded *wet–dry*, since this adjective was less clearly grounded in the available sensing modalities and was therefore harder to justify as a meaningful target for the present signal set. Examples of the target distributions for all adjective pairs are shown in Appendix A.

3.2. Preprocessing

3.2.1. Normalisation of thermal signals

Normalisation is used to make the thermal signals comparable across participants and channels and to stabilise training. Three strategies are considered: no normalisation, global normalisation, and per-participant normalisation.

In the global variant, a Keras Normalization layer is adapted on the training data only. This layer subtracts a single global mean and divides by a single global standard deviation for each channel, so that all inputs have approximately zero mean and unit variance per channel across all participants. The same transformation is then applied to the validation data.

In the per-participant variant, which yields the best results and is used in the final thermal models (see Appendix B), the normalisation is tailored to each participant. The idea is that different participants can have different baselines (e.g. skin temperature or contact behaviour), and these global offsets are not informative for material discrimination. Per-participant normalisation removes these individual baselines while preserving the temporal dynamics within each trial.

Let $x_{n,t,c}^{(p)}$ denote the value of channel c at time t in trial n belonging to participant p, and let $\mathcal{I}_p$ be the set of training trials from participant p. For each participant p and channel c, the mean and standard deviation are estimated over all timesteps of all training trials of that participant:

$$\mu_{p,c} = \frac{1}{|\mathcal{I}_p|\, T_{\text{therm}}} \sum_{n \in \mathcal{I}_p} \sum_{t=1}^{T_{\text{therm}}} x_{n,t,c}^{(p)}, \tag{3.4}$$

$$\sigma_{p,c} = \sqrt{\frac{1}{|\mathcal{I}_p|\, T_{\text{therm}}} \sum_{n \in \mathcal{I}_p} \sum_{t=1}^{T_{\text{therm}}} \left(x_{n,t,c}^{(p)} - \mu_{p,c}\right)^2}. \tag{3.5}$$

All trials from participant p (both training and validation) are then normalised using that participant's statistics:

$$\tilde{x}_{n,t,c}^{(p)} = \frac{x_{n,t,c}^{(p)} - \mu_{p,c}}{\sigma_{p,c} + \varepsilon}, \tag{3.6}$$

with a small ε to avoid division by zero.

This per-participant normalisation ensures that, for each participant, the thermal channels are centred and scaled relative to that participant's own distribution. As a result, the models are encouraged to focus on how the thermal signal evolves over time for a given material, rather than on absolute level differences between participants.

3.2.2. Sliding signals (friction and vibration)

For the sliding modality, we use the normal force $F_z(t)$, tri-axial accelerations $(A_x(t), A_y(t), A_z(t))$, and tangential fingertip speed $v(t)$. Since sliding trials can contain large within-trial variation in both force

and speed, we preprocess the signals and then extract only time intervals that satisfy controlled sliding conditions.

First, to reduce the effect of short-lived spikes and measurement noise, we smooth the normal force and speed signals using a short moving average filter. This smoothing is used both to stabilize the segmentation step and as input to the model, making the sliding branch less sensitive to brief outliers in $F_z(t)$ and $v(t)$.

Second, to isolate texture-related vibration content, we band-pass filter the acceleration signals using a Butterworth filter (Butterworth, 1930), keeping frequencies between 20 Hz and 1 kHz. Frequencies below 20 Hz mainly reflect slow hand and finger motion, while frequencies above 1 kHz are dominated by sensor noise and are outside the most relevant range for tactile vibration signals (Shultz et al., 2018). This filtering removes slow drift and high-frequency noise and retains vibration components that are more likely to arise from texture-dependent friction.

Next, a segmentation algorithm is applied to obtain a subset of the data that includes only samples recorded within predefined force and speed bounds. Let $f(t)$ denote the normal force (i.e., $f(t) = F_z(t)$) and $v(t)$ the tangential speed at time t. The preprocessed sliding signal is decomposed into N valid segments x_i, each spanning a time interval $[t_i, T_i]$, and the final training signal is defined as the concatenation

$$x = [x_1, x_2, \dots, x_N],$$

subject to the following constraints for all $i \in \{1, \dots, N\}$ and all $t \in [t_i, T_i]$:

$$v(t) \in [v_{\min}, v_{\max}], \tag{3.7}$$

$$f(t) \in [f_{\min}, f_{\max}], \tag{3.8}$$

$$t_{i+1} - T_i < \delta, \tag{3.9}$$

$$T_N - t_0 > \psi. \tag{3.10}$$

In this work, $[f_{\min}, f_{\max}] = [0.2, 0.8]$ N and $[v_{\min}, v_{\max}] = [33, 167]$ mm/s. Here, δ is a tolerance that allows short interruptions during which force or speed may briefly leave the admissible ranges while still maintaining continuity of the overall interaction, and ψ is a minimum total duration that ensures the signal contains enough data to form at least one valid training sample. The parameters $[f_{\min}, f_{\max}]$, $[v_{\min}, v_{\max}]$, δ, and ψ are treated as design choices: looser bounds retain more data but increase variability, while tighter bounds yield more consistent segments at the cost of dataset size.

After segmentation, each valid interval is cut into fixed-length windows to match the input length used for training. This produces a set of sliding segments that (i) are band-pass filtered in acceleration, (ii) lie within controlled ranges of force and speed, and (iii) are temporally standardized, making them suitable as input to the sliding branch of the models.

3.2.3. Pressing signals (indentation and normal force)

For the pressing modality, the normal force $F_z(t)$ and the indentation (stage-position) signal are used. In the raw recordings, the force/torque data are sampled at 10 kHz and the stage-position signal at 10 Hz over a 12 s press–release interaction. To obtain a consistent input representation, both signals are resampled and aligned to a common sampling rate equal to the lower of the two (10 Hz), and fixed-length windows of T_{press} samples per trial are extracted.

As in the thermal branch, per-participant, per-channel z-score normalisation is applied, with means and standard deviations estimated on the training split and a global fallback for participants that do not appear in the training data of a given fold.

3.2.4. Splits and cross-validation

All data were divided into training, validation, and test sets. In this thesis, a 64/16/20 split was used, meaning that 64% of the data was used for model fitting, 16% for validation during model development, and the remaining 20% was held out for final evaluation. This type of three-way split is a common practical choice in supervised learning, because it separates model fitting, model selection, and final performance assessment. The exact proportions are heuristic rather than fixed by a universal rule, but the underlying principle is standard: the validation set is used to guide model choices, while the test set is kept untouched until the very end to provide an unbiased estimate of generalization performance.

To make performance estimates less dependent on one particular partition of the development data, stratified k-fold cross-validation was used. In stratified k-fold cross-validation, the data are divided into

k folds such that each fold preserves, as closely as possible, the class proportions of the full dataset. The model is then trained k times, each time using $k-1$ folds for training and the remaining fold for validation, after which the results are averaged across folds. In this work, $k = 5$ was used. Five-fold cross-validation is a common default in modern machine-learning practice, and stratification is especially useful in classification problems because it reduces the risk that individual folds become unrepresentative due to class imbalance (scikit-learn developers, 2026).

3.3. Models

3.3.1. Overall modelling setup

The modelling pipeline in this thesis consists of three main models, which are summarized in Figure 3.2. Model 1 maps the input interaction signals to psychophysical adjective rating distributions, Model 2 takes these predicted adjective distributions as input and maps them to a material class label, and Model 3 maps the input signals directly to the material class. In this way, Models 1 and 2 together define an adjective-mediated classification route, while Model 3 serves as the direct signal-based benchmark.

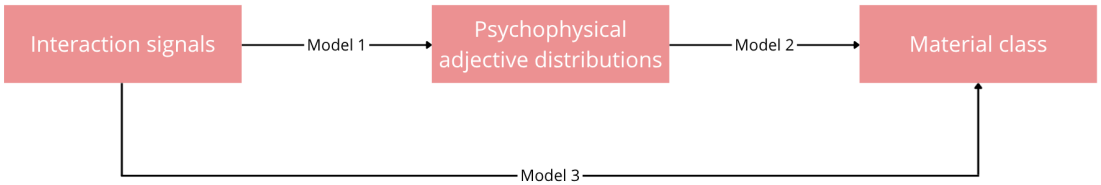


Figure 3.2: Overall modelling pipeline used in this thesis. Model 1 maps the input interaction signals to psychophysical adjective distributions, Model 2 maps these predicted adjective distributions to a material class, and Model 3 maps the input signals directly to the material class.

More detailed pipeline diagrams for the individual models are provided in Appendix C, where the corresponding internal processing blocks are shown in more detail.

Model 1 predicts, for each adjective pair (e.g. hot–cold or hard–soft), a discrete probability distribution over the 1–15 psychophysical rating scale. Model 2 uses these predicted adjective distributions to test how much material information can be recovered from perceptual descriptors alone. Model 3, in contrast, bypasses the adjective layer and predicts the material class directly from the interaction signals.

Model 1 and Model 3 operate on one or more of the available sensor modalities: thermal, sliding, and pressing (see Section 3.1). Model 2 does not take raw sensor signals as input, but instead operates on the adjective distributions predicted by Model 1.

3.3.2. Model 1: signal to adjective rating distributions

Architecture Model 1 reuses the modality-specific encoders and late-fusion structure from Model 3, but replaces the final material-classification layer with four adjective-specific distribution heads, as shown in Figure 3.3. The goal is to map an input signal (thermal, sliding, or pressing) to a discrete rating distribution over the 1–15 psychophysical scale, while keeping the architecture as close as possible to the material-classification network for comparability. Because the input signals are recorded in separate trials and are not temporally aligned, the modalities are combined using late fusion: each modality is encoded independently into a fixed-length embedding, and the embeddings are fused at the feature level. A more detailed version of the full pipeline is provided in Appendix C (Figure C.2).

Modality encoders. For each modality $m \in \{\mathrm{T}, \mathrm{S}, \mathrm{P}\}$ we compute an embedding vector

$$\mathbf{f}_m = E_m(x^{(m)}) \in \mathbb{R}^{64},$$

where $x^{(m)} \in \mathbb{R}^{T_m \times C_m}$ is the input sequence for modality m. The thermal encoder uses two temporal convolution layers (64 filters, kernel size 5; then 128 filters, kernel size 5) with ReLU activations and

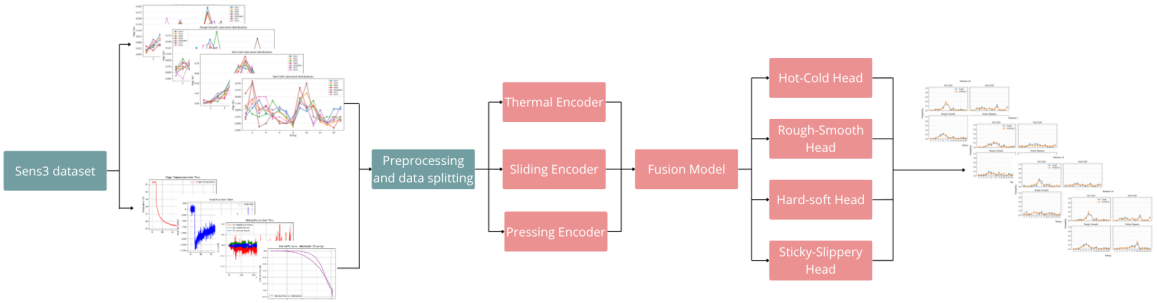


Figure 3.3: High-level pipeline overview for Model 1. The input interaction signals are preprocessed and split into training and evaluation sets, encoded separately per modality, and fused into a shared representation. Four adjective-specific output heads then predict 15-bin rating distributions for the psychophysical adjective pairs. A more detailed version is provided in Appendix C (Figure C.2).

batch normalisation, followed by dropout, flattening, and a Dense(64) layer with batch normalisation to produce $\mathbf{f}_{\mathrm{T}}$. The sliding and pressing encoders follow a similar 1D-CNN structure but incorporate temporal max-pooling and global average pooling to obtain compact representations. Both apply a Conv1D(64,7) block with batch normalisation and pooling, followed by a Conv1D(96,5) block with batch normalisation (with pooling in the sliding branch), and a Conv1D(128,3) block. Global average pooling is then used, followed by a Dense(64) layer and dropout to produce $\mathbf{f}_{\mathrm{S}}$ and $\mathbf{f}_{\mathrm{P}}$. Dropout is applied within each encoder to reduce overfitting.

Late fusion and multi-head prediction. The modality embeddings are concatenated into a fused representation

$$\mathbf{f} = [\mathbf{f}_{\mathrm{T}}; \mathbf{f}_{\mathrm{S}}; \mathbf{f}_{\mathrm{P}}], \tag{3.11}$$

which is passed through a shared fusion block consisting of Dense(128) with ReLU activation, batch normalisation, and dropout. On top of this shared representation, Model 1 uses one output head per adjective pair. Each head consists of a Dense(64) layer with ReLU activation followed by dropout, and a final Dense(15) softmax layer that outputs the predicted rating distribution $\hat{\mathbf{q}}^{(a)}(x) \in \Delta^{15}$.

This multi-head late-fusion design lets each modality encoder specialise to its own signal characteristics (e.g., thermal transients, vibration/friction structure, indentation dynamics), while the shared fusion block learns how to combine these signals for predicting perceived attributes.

KLD-only distribution loss. Since Model 1 predicts a rating distribution over a 15-bin scale for each adjective pair, training is formulated as a distribution-matching problem. For each material class and adjective pair, a target rating distribution is constructed from the available psychophysical ratings and normalized to sum to 1. These material-level target distributions are then reused for all sensor samples belonging to that material class. Let $p_n^{(a)} \in \mathbb{R}^R$ denote the target distribution assigned to sample n for adjective pair a, and let $\hat{q}_n^{(a)}$ be the corresponding predicted distribution after the model's softmax layer, with $R = 15$ bins. The loss for adjective pair a is the Kullback–Leibler divergence (KLD):

$$\mathcal{L}_{\mathrm{KL}}^{(a)} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{R} p_{n,k}^{(a)} \log \left(\frac{p_{n,k}^{(a)}}{\hat{q}_{n,k}^{(a)}} \right). \tag{3.12}$$

The final Model 1 is trained as a joint multi-head model over all four adjective pairs, so the total training objective is the sum of the four head-specific KLD losses.

Interpretation. KLD measures how well the predicted probability mass aligns with the empirical rating histogram: it is zero only when the predicted distribution exactly matches the target. Intuitively, KLD penalizes assigning probability to bins that humans rarely chose and rewards concentrating probability where humans often chose. Because KLD can be interpreted as an excess log-loss when using $\hat{q}$ to represent data generated from p, a compact intuition is given by $\exp(-D)$ for a divergence value D. For example, $D = 0.336$ corresponds to $\exp(-D) = 0.715$, meaning the model achieves about 71.5%

of the ideal geometric-mean likelihood of a perfect distribution match (see Chapter 4 for the reported KLD values).

Evaluation and qualitative inspection. Model 1 is evaluated using the average KLD on the validation set, computed per sample using Eq. 3.12 with the material-specific target distribution assigned to that sample, and then averaged across all validation samples. For qualitative inspection, we visualize the predicted rating distributions at the sample level. For each validation sample, a single figure is generated that contains four panels, one for each adjective pair (Hot–Cold, Hard–Soft, Rough–Smooth, Sticky–Slippery). Each panel overlays the target histogram and the predicted distribution over the 15-bin rating scale. This provides an intuitive check of whether the model captures the main properties of the human ratings, such as the peak location, spread, and skew, and whether it assigns probability mass to plausible neighboring bins rather than placing mass far from the target.

In the broader pipeline, the predicted adjective distributions from Model 1 are passed to Model 2 to evaluate how much material information is recoverable from psychophysical descriptors alone. Since Model 1 shares the same time-series encoder structure as Model 3, it can also be analysed with attribution-based methods.

3.3.3. Model 2: adjective-to-material classifier

Model 2 is a lightweight material classifier that operates on the psychophysical descriptors predicted by Model 1. For each sample, Model 1 outputs a 15-bin distribution for each adjective pair. These distributions are concatenated into a single feature vector

$$\mathbf{x} = [\hat{\mathbf{q}}^{(\text{hot–cold})};\ \hat{\mathbf{q}}^{(\text{hard–soft})};\ \hat{\mathbf{q}}^{(\text{rough–smooth})};\ \hat{\mathbf{q}}^{(\text{sticky–slippery})}] \in \mathbb{R}^{60},$$

which forms the input to Model 2. The goal of Model 2 is then to predict the 7-way material label from this compact, human-interpretable representation. As in the rest of the thesis, we exclude Rubber, Plastic, and Leather and keep the seven retained material classes.

Architecture. Model 2 is implemented as a small multilayer perceptron (MLP). It consists of a dense hidden layer (default 32 units) with batch normalisation and dropout, followed by a softmax output layer over the seven material classes. Additional hidden layers are supported but disabled in the default configuration. L2 regularisation is applied to the dense layer weights, and label smoothing is used during training. An overview of the model architecture can be found in Figure 3.4.

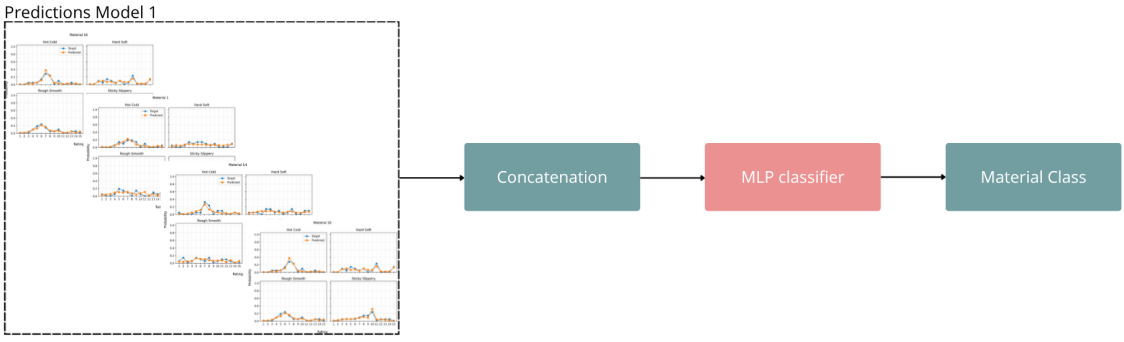


Figure 3.4: High-level overview of Model 2. The four 15-bin adjective distributions predicted by Model 1 are concatenated into a 60-dimensional feature vector, which is passed to a lightweight MLP classifier to predict the material class.

Training objective and optimisation. Model 2 is trained using cross-entropy with label smoothing (implemented as a smoothed one-hot target when sparse label smoothing is not directly supported). Optimisation uses Adam with the tuned learning rate from the configuration, along with early stopping and learning-rate reduction on plateau. Training uses a batch size of 8.

Cross-validation and evaluation. To avoid optimistic estimates, Model 2 is evaluated using a stacking-style protocol that mirrors the fold structure of Model 1: for each fold, Model 2 is trained on Model 1 validation predictions from the other folds and evaluated on the held-out fold predictions. Performance is reported using accuracy and macro-F_1, and confusion matrices are computed per fold for later analysis.

Interpretability. Because Model 2 operates on concatenated adjective distributions rather than on raw time-series signals, its interpretability analysis differs from that of Model 1 and Model 3. The corresponding procedure is described in Section 3.3.6.

3.3.4. Model 3: signal-based material classifier

Backbone candidates considered during development. Because tactile interaction signals are sequential, two backbone families were considered during model development: 1D convolutional networks and recurrent LSTM networks. The 1D-CNN was included as a strong baseline for capturing local temporal patterns such as thermal transients and vibration bursts, while the LSTM was considered because of its ability to model longer-range and order-sensitive temporal dependencies. This comparison formed part of the broader architecture exploration and is reported in the supplementary results. The final Model 3 configuration used in the main experiments is described separately below.

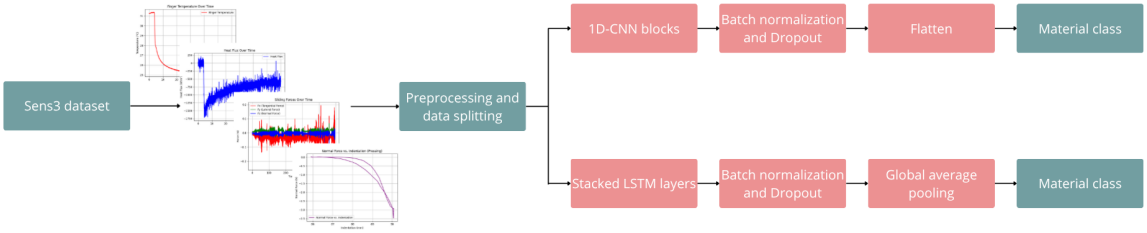


Figure 3.5: Backbone families considered during model development for sequence modelling: a 1D-CNN encoder and a stacked LSTM encoder. These architectures were explored as candidate sequence backbones; the final Model 3 used modality-specific 1D-CNN encoders in a late-fusion configuration.

Final selected architecture. For the final Model 3 experiments reported in the main results, a multimodal late-fusion architecture with modality-specific 1D-CNN encoders was used, as shown in Figure 3.6. Each selected modality is processed by its own dedicated encoder, after which the resulting modality embeddings are concatenated and mapped to the final material prediction through a shared fusion head. This architecture supports both unimodal and multimodal settings by applying the same fusion head either to a single modality embedding or to the concatenation of multiple embeddings.

For each modality $m \in \{\mathrm{T}, \mathrm{S}, \mathrm{P}\}$, an encoder network produces an embedding vector

$$\mathbf{f}_m = E_m(x^{(m)}) \in \mathbb{R}^{64},$$

where $x^{(m)} \in \mathbb{R}^{T_m \times C_m}$ denotes the input sequence for modality m. The thermal branch uses two temporal convolution layers with ReLU activations and batch normalization, followed by dropout, flattening, and a dense projection to 64 dimensions. The sliding and pressing branches use deeper convolutional stacks with intermediate pooling and global average pooling before the final dense projection. This design allows each modality branch to specialize to its own signal characteristics while producing embeddings of equal dimensionality for fusion.

When multiple modalities are used, the embeddings are concatenated into a fused representation

$$\mathbf{f} = [\mathbf{f}_\mathrm{T}; \mathbf{f}_\mathrm{S}; \mathbf{f}_\mathrm{P}],$$

which is passed to the fusion head for final classification. When only one modality is used, the same classification head is applied directly to that single modality embedding. In this way, the unimodal and multimodal configurations differ only in whether modality embeddings are concatenated before the

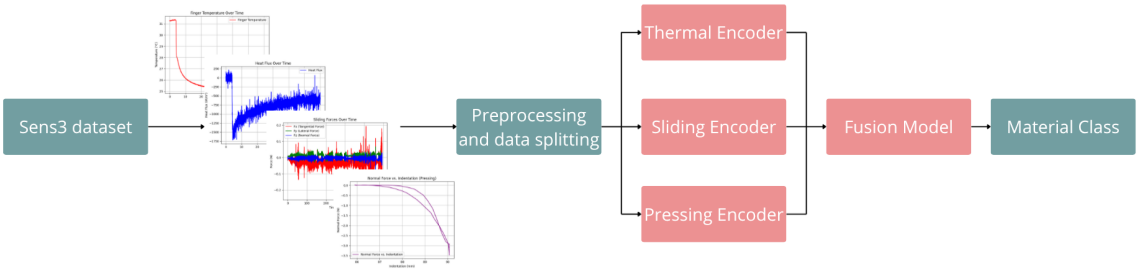


Figure 3.6: High-level pipeline overview for Model 3. The input interaction signals are preprocessed and split into training and evaluation sets, encoded separately per modality, and fused into a joint representation that is used to predict the material class. A more detailed version is provided in Appendix C (Figure C.1).

final classifier. In this thesis, the main Model 3 results are obtained with all three available modalities: thermal, pressing, and sliding. A more detailed version of the full pipeline is provided in Appendix C (Figure C.1).

Fusion/classification head. The fused representation is mapped to the final class probabilities by a shared classification head consisting of a Dense(128) layer with ReLU activation, followed by batch normalization, dropout, and a final Dense(7) softmax layer over the retained material classes. The hidden dense layer allows the model to re-weight and combine the modality embeddings before classification, while batch normalization and dropout act as regularizers. Using the same head structure for both unimodal and multimodal variants ensures that the main architectural difference lies in the input representation rather than in the classifier itself.

Training objective and optimisation. Model 3 is trained with sparse categorical cross-entropy,

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N} \log p_\theta(y_n \mid x_n),$$

where x_n is the input signal, $y_n \in \{0, \dots, 6\}$ is the corresponding material label, and p_θ denotes the softmax output of the network. Optimization uses Adam with a tuned learning rate from the final hyperparameter configuration. Regularization is provided by dropout in both the modality encoders and the fusion head, together with batch normalization after intermediate layers. During training, validation performance is monitored across epochs and the trained model is saved after training for later evaluation and analysis. Final hyperparameters and training settings are reported in Appendix E.

Preprocessing and normalization. Normalization is applied as a preprocessing step rather than as a layer inside the network. For each split, the required mean and standard deviation statistics are computed from the training data only and then applied to the corresponding training, validation, and test arrays using the same parameters. This avoids leakage of information from validation or test data into the preprocessing stage and ensures that the reported performance reflects the true evaluation protocol.

Evaluation protocol. Model 3 is evaluated using a two-stage split protocol. First, the full dataset is divided into a fixed 80/20 train+validation/test split. The held-out test set remains untouched during model selection and is used only for final evaluation. Second, within the train+validation portion, stratified 5-fold cross-validation is performed to assess model stability and to guide model selection.

For each fold, a fresh instance of Model 3 is trained on the fold-specific training split and evaluated on the corresponding validation split. Normalization statistics are computed on the training portion of that fold only and then applied to the validation and test data using the same parameters. Performance is reported using accuracy and macro-F_1, and confusion matrices are computed per fold for later analysis. After cross-validation, the trained models are also evaluated on the fixed held-out test set to quantify final generalization performance.

3.3.5. Attribution-based interpretability

To understand which parts of the learned representation drive the model predictions, we apply attribution-based interpretability to the trained models. Since the models in this thesis operate on different types of input, the exact form of the explanation depends on the model under consideration. For Model 1, Model 3, and the composite Model 1→Model 2 pathway, attribution is computed with respect to raw temporal sensor inputs. For Model 2 itself, attribution is computed with respect to the concatenated adjective-distribution input vector. In all cases, explanations are computed on held-out evaluation data within the same cross-validation setup as the reported performance results, so that the explanations reflect model behaviour under evaluation conditions.

For the signal-based models, the primary method is Integrated Gradients (IG) (Section 2.3.1), which assigns relevance to individual timestep–channel inputs by integrating gradients from a baseline input to the actual sample. We use an all-zero baseline in the normalised input space. Since raw time-series saliency can be diffuse, we apply Temporal Saliency Rescaling (TSR) (Section 2.3.2) to obtain sharper and more stable attribution maps when the target is scalar. For each sample, this produces an attribution tensor $\tilde{\mathbf{A}} \in \mathbb{R}^{T \times C}$ with the same shape as the input sequence.

3.3.6. Interpretability of Model 2

Unlike Model 1 and Model 3, Model 2 does not operate on raw temporal sensor signals. Instead, its input consists of the concatenated adjective-distribution outputs of Model 1. The interpretability analysis for Model 2 therefore addresses a different question. Rather than identifying which parts of the sensor signals were most important, it examines which predicted adjective pairs contributed most to the final material decision. As a result, no temporal, phase-, bin-, or channel-level analysis is available for this model.

The input to Model 2 is a concatenation of four predicted adjective distributions,

[*hot–cold* | *hard–soft* | *rough–smooth* | *sticky–slippery*],

where each adjective pair is represented by a 15-bin probability distribution. The full input vector therefore has dimensionality $4 \times 15 = 60$. To explain Model 2 decisions at the adjective-pair level, two complementary techniques were used: Integrated Gradients and block ablation.

Interpretability was computed on held-out evaluation samples reconstructed from the same Model 1 outputs that were used during Model 2 training. In the fold-based setting, explanations were generated on the held-out validation fold of each run, after which the results were also combined into a single summary across all held-out folds. Since the current Model 2 pipeline does not save a separate test-split artifact, this interpretability analysis should therefore be understood as operating on held-out validation predictions rather than on a separately stored test set.

The IG computation was performed on the target class logit rather than on the post-softmax probability, in order to avoid distortions introduced by the normalisation of the softmax layer. For each sample, feature-level attributions were then summed within each adjective block, yielding one signed attribution value per adjective pair. In addition, positive-only and absolute normalised percentages were computed to express the relative contribution of each adjective pair within a sample.

To complement the gradient-based analysis, a block-ablation procedure was also applied. For each adjective pair, the corresponding 15-bin block in the input vector was replaced by its baseline value, while all other adjective blocks were kept unchanged. The classifier was then evaluated again, and the change in the target class score was recorded. More specifically, if $x^{(-a)}$ denotes the input after replacing adjective block a by the baseline, then the ablation effect for adjective pair a was measured as

$$\Delta_a^{\text{logit}} = F_c(x) - F_c(x^{(-a)}),$$

with an analogous quantity also computed for the target class probability. A larger positive drop indicates that the removed adjective block was more important for the original prediction. As with IG, these raw drops were further converted into normalised positive-only and absolute percentage contributions across adjective pairs.

Explanations were generated with respect to the predicted class by default, although the implementation also supports explanations with respect to the true class or a manually specified target class. For the experiments reported in this thesis, the adjective-pair contributions were aggregated across samples to obtain summary statistics at three levels: overall across all held-out evaluation samples, per

true material class, and separated by correct versus incorrect predictions. This produced a compact overview of which adjective pairs were most influential for Model 2 decisions, both globally and for specific classes.

It is important to note that the resulting explanations do not indicate which raw sensor signals were most important, but only which intermediate adjective-distribution blocks were most influential for the final material classifier. Model 2 interpretability should therefore be interpreted as a decision-level explanation of the adjective-based pathway, complementing the signal-level analyses performed for Model 1 and Model 3.

3.3.7. From signal-level attribution maps to modality, phase, and channel importance

Per-sample attribution maps are useful for qualitative inspection, but the main results in this thesis are based on aggregated summaries that quantify importance at three levels: modality (thermal, pressing, sliding), channel (e.g., heat flux vs. temperature, friction coefficient vs. IMU axes), and phase/bin within each modality. All summaries are computed from absolute attribution magnitudes $|\tilde{\mathbf{A}}|$, so positive and negative attributions do not cancel when aggregating.

For a given group g (a modality, channel, or phase/bin), we compute its attribution mass by summing absolute attributions over all timestep–channel pairs that belong to the group. This yields a per-sample group mass S_g. We then convert these masses into percentage shares s_g by normalising by the total attribution mass of the sample.

Because different modalities and phases can have different durations and different numbers of channels, we report normalised summaries to make comparisons fair. For modality- and phase-level results, we use time-normalised shares (i.e., normalising by the number of timesteps) so that longer segments do not automatically appear more important simply because they contain more samples. For channel-level results, we normalise by the number of contributing elements (timesteps $\times$ channels) so that higher-dimensional signals do not dominate purely due to having more entries. We also store raw (unnormalised) attribution sums and alternative normalisations (e.g., per-timestep) for later sensitivity analysis.

Finally, we aggregate per-sample shares across the dataset in two complementary ways. The primary results use a per-sample mean, where each sample contributes equally to the final percentages. As a robustness check, we also report pooled summaries, where attribution masses are summed across samples before converting to percentages; this highlights whether a small number of high-magnitude samples disproportionately influence the overall picture. All summaries are reported overall and stratified by material class and participant to analyse class-conditional and participant-dependent differences in signal usage.

4

Results

This chapter reports predictive performance and interpretability results for the models developed in this thesis. Section 4.1 summarizes performance for all models and provides additional detail for the models used in the final pipeline. Section 4.2 focuses on interpretability, using nested donut charts to visualize which modalities and signal parts drive the predictions. Finally, Section 4.3 summarizes how sensitive the interpretability conclusions are to different summarization settings.

Method definitions (metrics, interpretability algorithms, and summarization settings) are described in Chapter 3. To keep this chapter focused, extended results and implementation details are placed in the appendices: participant-wise performance and normalisation effects (Appendix B), pipeline schematics and stacking protocol details (Appendix C), hyperparameter/architecture tuning outcomes (Appendix D), full interpretability tables (Appendix F), full per-class donut-chart grids (Appendix G), attribution sensitivity analyses (Appendix H), and final training settings/hyperparameters (Appendix E).

4.1. Performance

Table 4.1 provides a compact overview of performance across all models. Model 1 predicts perceptual distributions and is evaluated using KL divergence (KLD). Model 3, and the composite Model (1+2), are evaluated on 7-class material classification using accuracy and macro-F1 across folds. For Model 1, lower KLD indicates better agreement between the predicted and ground-truth rating distributions, whereas for Model (1+2) and Model 3, higher accuracy and macro-F1 indicate stronger material classification performance. Within the classification models, macro-F1 is particularly important because it gives equal weight to all classes and is therefore more sensitive to weak classes than accuracy. Under this interpretation, Model 3 is clearly the strongest overall classifier, while the composite Model (1+2) shows substantially weaker and less stable performance. The relatively small standard deviations of Model 3 further indicate that this performance is consistent across folds. Final training settings and selected hyperparameters are listed in Appendix E.

Table 4.1: Overall performance summary (mean $\pm$ std over 5 folds). For Model 1, lower KLD indicates predicted perceptual distributions closer to the ground truth. For Model (1+2) and Model 3, higher accuracy and macro-F1 indicate better classification performance. Standard deviations indicate how consistent performance is across folds.

Model	Task	Metric	Validation	Test
1	Perceptual distribution prediction	KLD (overall)	0.341 ± 0.022	0.336 ± 0.020
1+2	Material classification	Accuracy	0.585 ± 0.220	0.600 ± 0.115
		Macro-F1	0.463 ± 0.247	0.463 ± 0.137
3	Material classification	Accuracy	0.834 ± 0.059	0.896 ± 0.017
		Macro-F1	0.807 ± 0.057	0.885 ± 0.020

4.1.1. Performance of Model 1

Model 1 predicts probability distributions over rating bins for four adjective pairs (Hot–Cold, Hard–Soft, Rough–Smooth, Sticky–Slippery). Performance is evaluated using KL divergence between predicted and ground-truth distributions (see Chapter 3). Table 4.2 reports mean test KLD across folds, both overall and per adjective pair. Additional performance breakdowns and supporting analyses (including participant-wise effects) are provided in Appendix B, while final training settings are listed in Appendix E.

Table 4.2: Model 1 test performance measured with KL divergence (KLD) per adjective pair (mean $\pm$ std over 5 folds). Lower values indicate predicted rating distributions that are closer to the ground truth.

Adjective pair	Test KLD
Overall	0.336 ± 0.020
Hot–Cold	0.294 ± 0.036
Hard–Soft	0.371 ± 0.015
Rough–Smooth	0.348 ± 0.021
Sticky–Slippery	0.331 ± 0.017

Table 4.2 shows that the adjective pairs are not equally easy to predict. Because lower KLD is better, Hot–Cold is the easiest adjective pair for the model, while Hard–Soft is the most difficult. This suggests that thermal perceptual structure is captured more reliably than compliance-related perceptual structure in the current signal representation. The relatively small spread across folds indicates that these differences are systematic rather than caused by one unusually good or poor split. A compact likelihood-style intuition is given by $\exp(-\mathrm{KLD})$; for $\mathrm{KLD} = 0.336$, $\exp(-0.336) \approx 0.71$, meaning the predicted rating distributions have roughly a 71% likelihood-style overlap with the ground truth.

4.1.2. Performance of the composite Model (1+2)

The composite Model (1+2) performs 7-class material classification. Overall results show moderate accuracy but lower macro-F1, indicating that performance is limited by weaker classes. Table 4.3 reports overall validation and test performance, and Table 4.4 reports per-class F1. A schematic of the composite pipeline and the stacking-style evaluation protocol is provided in Appendix C. Final hyperparameters are listed in Appendix E.

Table 4.3: Composite Model (1+2) overall classification performance (mean $\pm$ std over 5 folds). Higher values indicate better performance. Macro-F1 is especially informative here because it highlights whether performance is balanced across classes rather than dominated by the easiest classes.

Split	Accuracy	Macro F1
Validation	0.585 ± 0.220	0.463 ± 0.247
Test	0.600 ± 0.115	0.463 ± 0.137

Table 4.3 shows that the composite model achieves only moderate classification performance. The gap between accuracy and macro-F1 indicates that the model does not perform equally well across all classes: some materials are recognized reasonably well, while others remain difficult. The relatively large standard deviations, especially on validation, further suggest that the model is sensitive to the specific fold and therefore less stable than desired. This already points to a limitation of the adjective-based route before inspecting the class-wise results in more detail.

To complement the per-class metrics, Figure 4.1 shows where the classification errors are concentrated. Because the matrix is row-normalized, each row should be read as the distribution of predictions for one true class.

The confusion matrix makes clear that the composite model does not fail uniformly. Instead, a limited set of class pairs accounts for much of the performance loss. This is particularly useful when read together with Table 4.4: low per-class F1 scores correspond to rows with more mass away from the diagonal, indicating systematic confusion rather than isolated mistakes.

Table 4.4: Composite Model (1+2) per-class F1 scores (mean ± std over 5 folds). Higher F1 indicates better balance between precision and recall for that class. Supports are the total numbers of samples across folds and provide context for how much data each class contributes to the reported average.

Class	Val support	Val F1	Test support	Test F1
Fabric	29	0.821 ± 0.158	35	0.799 ± 0.041
Foam	19	0.456 ± 0.445	25	0.541 ± 0.336
Metal	12	0.733 ± 0.435	15	0.900 ± 0.224
Paper	14	0.164 ± 0.289	20	0.358 ± 0.214
Sandpaper	10	0.333 ± 0.471	10	0.160 ± 0.358
Vinyl	10	0.248 ± 0.341	10	0.067 ± 0.149
Wood	14	0.487 ± 0.338	20	0.418 ± 0.284

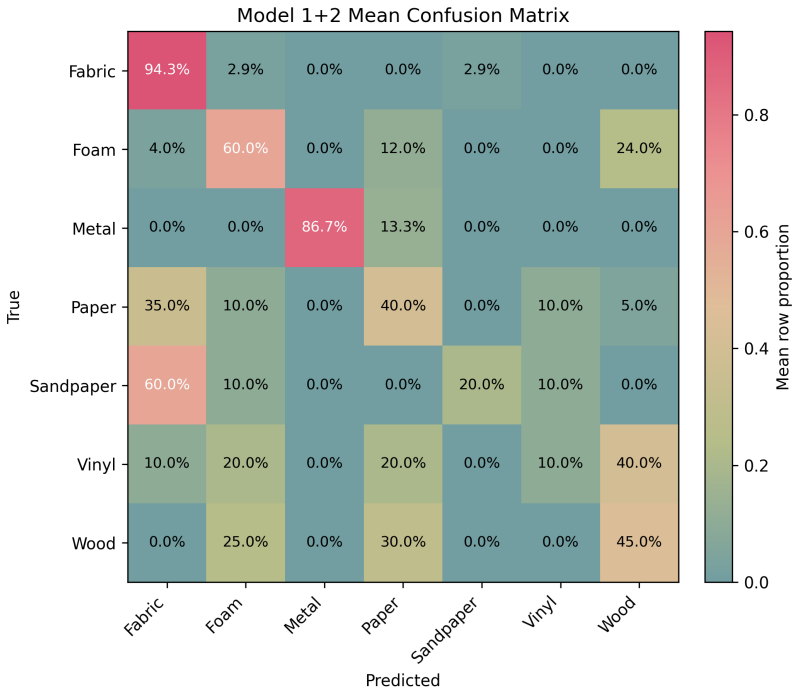


Figure 4.1: Model 1+2 mean confusion matrix on test (row-normalized, aggregated across folds).

4.1.3. Performance of Model 3

Model 3 is the strongest classifier in this study. It achieves high validation performance and remains strong on the held-out test set. Table 4.5 reports overall results, and Table 4.6 reports per-class F1.

Table 4.5: Model 3 overall classification performance (mean $\pm$ std over 5 folds). Higher values indicate better performance, and the small standard deviations indicate stable performance across folds.

Split	Accuracy	Macro-F1
Validation	0.834 ± 0.059	0.807 ± 0.057
Test	0.896 ± 0.017	0.885 ± 0.020

Table 4.5 shows that Model 3 performs strongly on both accuracy and macro-F1. The high macro-F1 indicates that the performance is not only high on average, but also well balanced across classes. In addition, the small standard deviations show that this result is stable across folds, which increases confidence that the model is learning a robust decision strategy rather than benefiting from a favorable split. Compared to the composite model, Model 3 is both more accurate and more consistent.

Table 4.6: Model 3 per-class F1 scores (mean $\pm$ std over 5 folds). Higher F1 indicates better class-wise performance. Supports are the total sample counts across folds and provide context for the reliability of each class-level estimate.

Class	Val support	Val F1	Test support	Test F1
Fabric	29	0.927 ± 0.124	35	1.000 ± 0.000
Foam	19	0.855 ± 0.135	25	0.841 ± 0.102
Metal	12	1.000 ± 0.000	15	1.000 ± 0.000
Paper	14	0.501 ± 0.340	20	0.798 ± 0.087
Sandpaper	10	1.000 ± 0.000	10	0.867 ± 0.183
Vinyl	10	0.714 ± 0.165	10	0.853 ± 0.145
Wood	14	0.653 ± 0.384	20	0.836 ± 0.120

Table 4.6 shows that Model 3 performs strongly across nearly all classes, although the classes are not equally easy. Fabric and Metal are classified almost perfectly, while Foam, Sandpaper, Vinyl, and Wood also achieve strong test F1 scores. Paper remains the most challenging class, but even here the test performance is substantially stronger than in the composite model. Overall, the table indicates that the direct multimodal route does not only improve aggregate performance, but also reduces the number of weak classes.

For Model 3, Figure 4.2 shows the remaining confusion structure on the test set. As with the composite model, the row-normalized format should be read class by class.

Compared with Figure 4.1, the diagonal is visibly stronger for most classes, which is consistent with the higher macro-F1 reported in Table 4.5. The remaining confusions are therefore more limited and class-specific, rather than reflecting a broad weakness of the model. This supports the conclusion that the direct multimodal model preserves more discriminative information than the adjective-based route.

4.2. Interpretability

This section reports interpretability results using nested donut charts. The outer ring shows the relative contribution of thermal, pressing, and sliding modalities, while the inner ring breaks down contributions within each modality by phases/bins and by channels. The interpretability method and the exact summarization settings used for percentages are defined in Chapter 3. Full numerical interpretability tables are reported in Appendix F, and full per-class donut grids (validation and test) are provided in Appendix G. Sensitivity checks for summarization choices (TSR examples, normalisation and aggregation comparisons) are provided in Appendix H.

4.2.1. Interpretability of Model 1

Model 1 predicts full rating distributions for each adjective pair rather than a single class score. For interpretability, IG attributions are reported and summarized in the same way as the other models (outer ring: modality share; inner ring: phase/bin or channel share). TSR is not applied for Model 1 because its

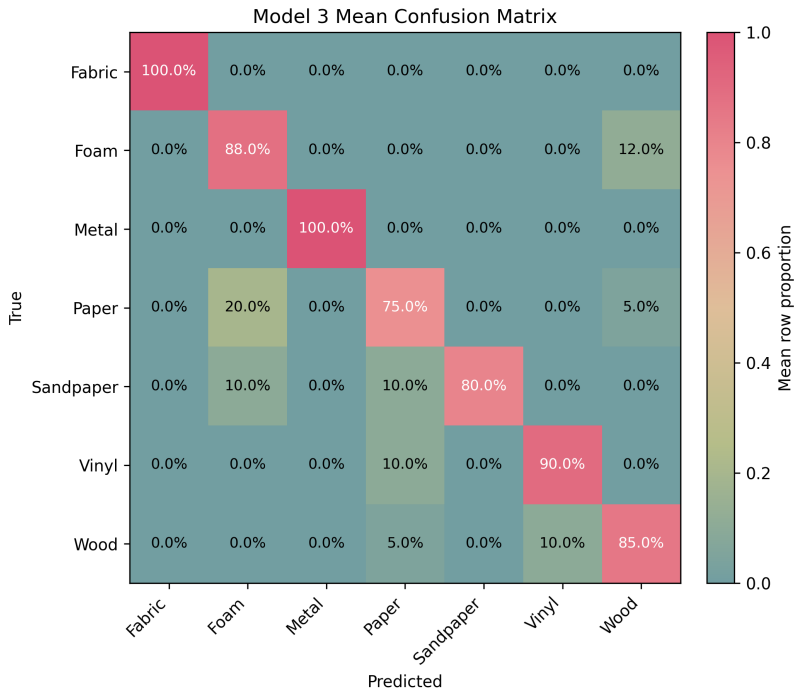


Figure 4.2: Model 3 mean confusion matrix on the test set, row-normalized and averaged across folds. Each row sums to 1, so diagonal values indicate correct recognition for each true class and off-diagonal values indicate systematic confusion with other materials.

output is a 15-bin distribution: applying TSR would require collapsing the distribution to a single scalar target (e.g., selecting one bin or using an expected rating), which would change what the attribution represents and introduce an extra design choice.

Across adjective pairs, Model 1 shows a consistent multimodal profile. Thermal is the largest contributor for all four adjectives, typically accounting for roughly half of the attribution mass, while sliding contributes about one third and pressing about one fifth. The exact balance shifts slightly between adjective pairs: Rough–Smooth is most thermal-driven, while Sticky–Slippery shows the strongest pressing contribution. The modality splits are stable between validation and test, suggesting that Model 1 relies on similar signals under both evaluation splits (see Figure 4.3).

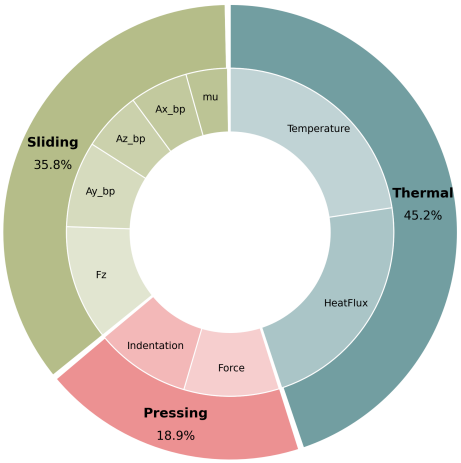
Within each modality, the inner ring highlights which parts of the interaction matter most for a given adjective. For thermal, attribution concentrates in transient contact phases (peak and half-equilibration) rather than purely steady contact. For pressing, plateau and lift-off generally dominate over early loading. For sliding, the most salient evidence is concentrated in the low-force regime, spread over low-to-mid speeds, consistent with the dataset being restricted to controlled force–speed ranges.

A full numerical breakdown of modality shares per adjective pair (validation and test), as well as phase/channel summaries, is provided in Appendix F.

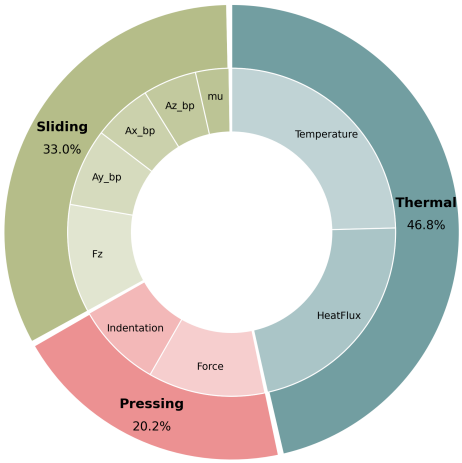
4.2.2. Interpretability of the composite Model (1+2)

The composite model shows a distinctive explanation profile in which sliding dominates the outer-ring attribution on both validation and test. On validation, sliding accounts for 65.2% of the attribution on average, compared to 21.5% thermal and 13.3% pressing. On the held-out test set the same pattern remains, with sliding at 62.5%, thermal at 24.2%, and pressing at 13.4%. This dominance is visible in the overall donut charts in Figure 4.4 and remains present across most classes. The full table values for these percentages are provided in Appendix F.

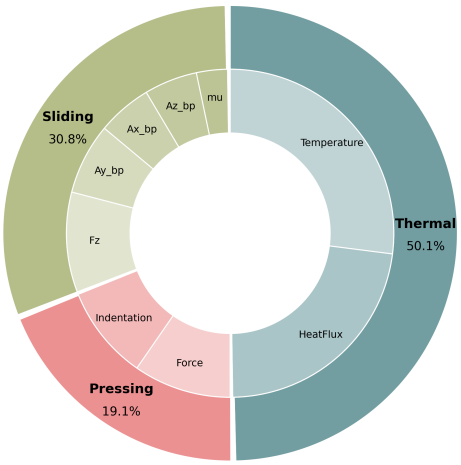
In the per-class donut charts, sliding remains dominant for nearly all materials. The inner ring further shows how the model uses evidence within each modality. For thermal, attribution is concentrated primarily in the transient parts of thermal contact (e.g., peak and half-equilibration) rather than purely in long steady contact. For pressing, attribution is more concentrated in lift-off and loading compared



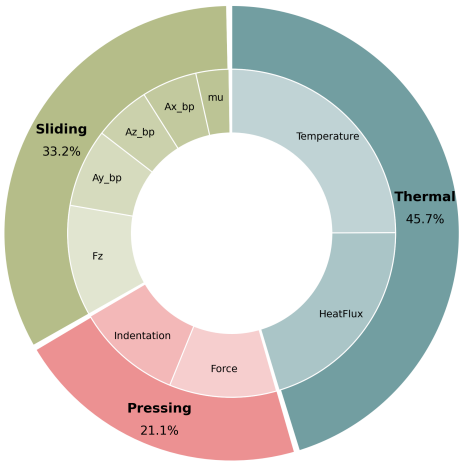
(a) Hot–Cold



(b) Hard–Soft

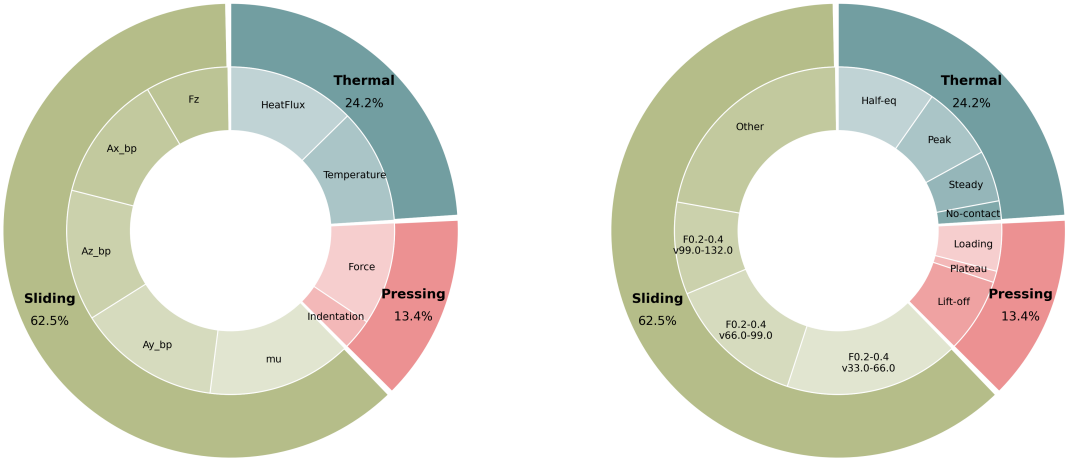


(c) Rough–Smooth



(d) Sticky–Slippery

Figure 4.3: Model 1 interpretability on the test set (outer ring: modality attribution; inner ring: phase/bin attribution) for each adjective pair.



(a) inner circle channels (b) inner circle phases

Figure 4.4: Composite Model (1+2) overall nested donut charts on test set.

to plateau, suggesting that pressing evidence is used in a narrower part of the interaction. For sliding, the most salient bins occur at relatively low forces across low-to-mid speeds, indicating that the model extracts much of its sliding evidence from the light-contact regime.

Channel-detailed donuts complement the phase/bin view by indicating which sensor streams dominate. For the composite model, sliding attribution is spread across friction and IMU channels rather than being dominated by normal force alone. Thermal attribution is shared between heat flux and temperature, and pressing attribution is dominated by one of the pressing features.

A full grid of per-class phase/bin donuts and channel-detailed donuts is provided in Appendix G, while the corresponding numerical tables (per-class modality shares, phase/bin shares, and channel breakdowns) are provided in Appendix F. Pipeline details for the composite setup are provided in Appendix C.

4.2.3. Interpretability of Model 2

The interpretability analysis of Model 2 showed a highly consistent adjective explanation pattern across both Integrated Gradients and block ablation. Overall, hard-soft was the most influential adjective pair for the final material decision, contributing approximately 34.0% of the positive IG share and 33.9% of the positive ablation share across all evaluation samples. The remaining contribution was distributed more evenly across rough-smooth (23.4% IG; 23.8% ablation), sticky-slippery (24.0% IG; 23.8% ablation) and hot-cold (18.6% IG; 18.4% ablation), see Figure 4.5. At the class level, this pattern remains largely stable, hard-soft was the dominant adjective pair for most true classes, with particularly strong contributions for Metal, Foam, Vinyl, and Wood. Fabric showed a more balanced dependency between hard-soft and rough-smooth, while Sandpaper was the clearest exception, relying most strongly on sticky-slippery (see Figure 4.6). The close agreement between the IG- and ablation-based summaries suggests that these adjective-level importance patterns are robust rather than an artifact of one specific explanation method.

4.2.4. Interpretability of Model 3

Model 3 shows a more balanced and class-conditional explanation profile than the composite model, as illustrated in Figure 4.7. On validation, the overall modality split is 40.2% thermal, 34.6% pressing, and 25.2% sliding. On the held-out test set, this pattern remains nearly unchanged, with 39.9% thermal, 34.0% pressing, and 26.1% sliding. Compared to the composite model, Model 3 relies more strongly on thermal and pressing information, while sliding remains an important but more selective source of evidence.

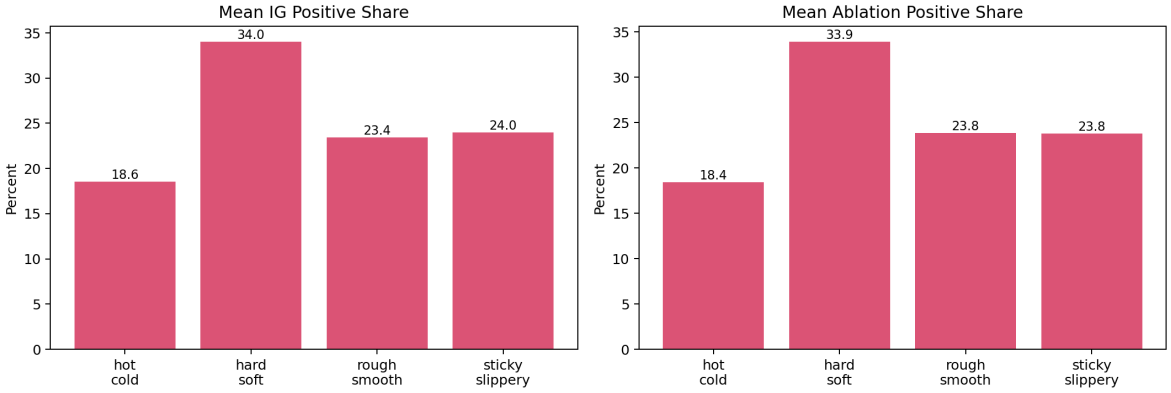


Figure 4.5: Overall adjective-pair contribution to Model 2 decisions across all evaluation samples, shown for both Integrated Gradients (left) and block ablation (right). Hard–soft is the dominant adjective pair overall, while hot–cold contributes the least.

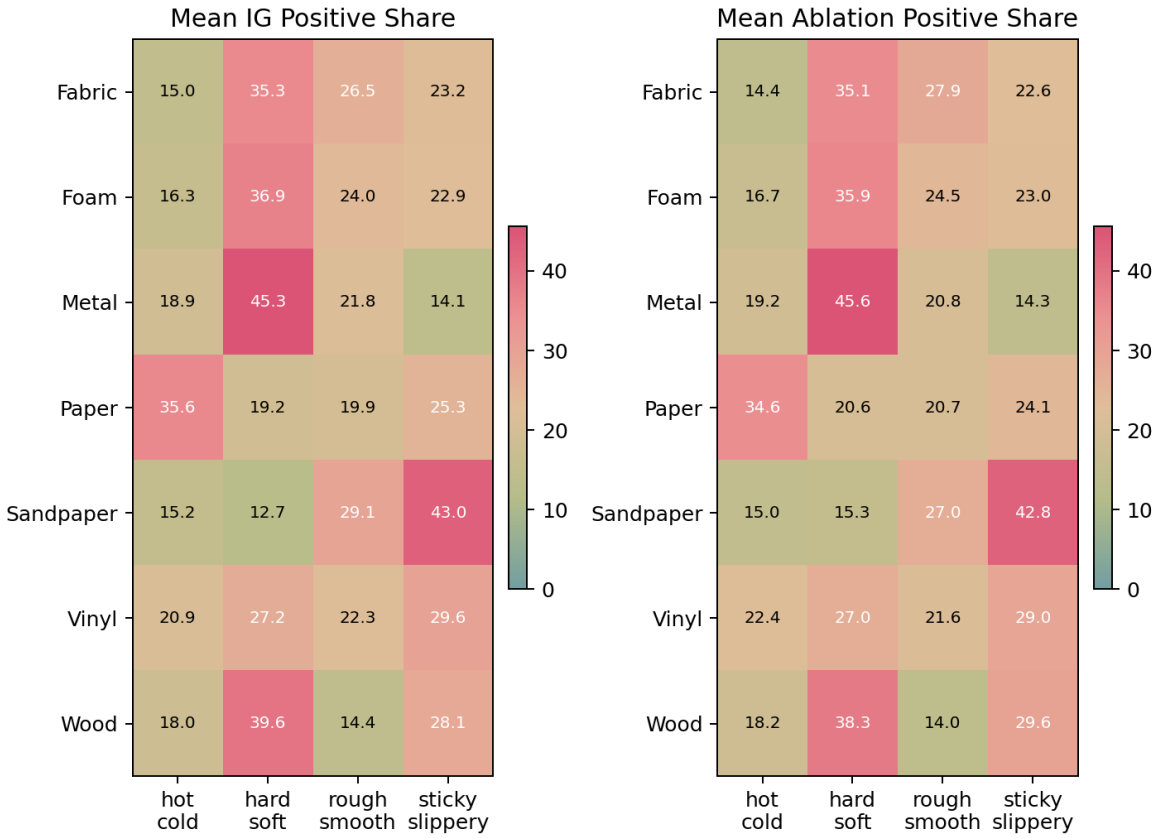
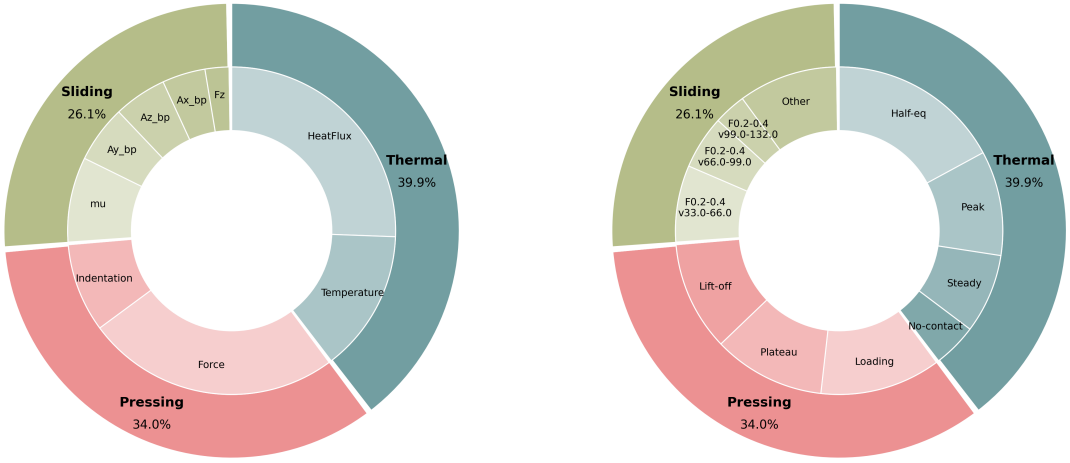


Figure 4.6: Mean adjective-pair contribution to Model 2 decisions per true material class, shown for both Integrated Gradients (left) and block ablation (right). Most classes are dominated by hard–soft, while Sandpaper shows the strongest relative reliance on sticky–slippery.



(a) inner circle channels

(b) inner circle phases

Figure 4.7: Model 3 overall nested donut charts on test set.

The per-class donut charts reveal clear specialization. Metal is strongly thermal-driven, consistent with its very high classification performance. Foam shows large sliding and thermal contributions, while Fabric and Paper are more pressing-driven. Wood, Vinyl, and Sandpaper display a more mixed modality usage. This mixed profile likely contributes to confusion: these classes are not separable by one dominant modality, so the model must rely on a combination of signals that also appear in other materials.

Within-modality patterns provide a more detailed view of how Model 3 uses the input signals. Within thermal, attribution concentrates in the transient heat-transfer phases, particularly peak and half-equilibration, rather than in steady-state contact. Within pressing, attribution is distributed across loading, plateau, and lift-off. Within sliding, the dominant bins occur mainly at low forces across multiple speed ranges. At channel level, thermal attribution is dominated by heat flux over temperature, pressing is driven primarily by normal force over indentation, and sliding depends most strongly on friction coefficient μ and the band-pass IMU axes, while normal force F_z contributes the least.

Full channel and phase/bin breakdown tables are provided in Appendix F. The complete set of interpretability figures, including validation and test variants, is provided in Appendix G, while sensitivity checks for the summarization choices are provided in Appendix H.

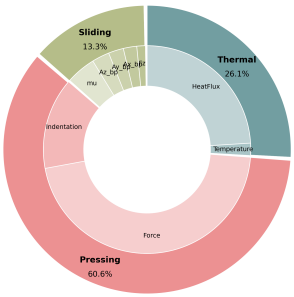
4.3. Sensitivity analysis

Interpretability percentages depend on how attributions are summarized. This thesis uses a single primary setting for the main figures in Section 4.2 and reports sensitivity analyses in Appendix H.

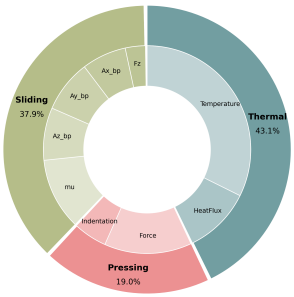
First, modality importance depends on normalization. Raw sums tend to increase the apparent importance of modalities with longer signals. Per-timestep normalization reduces the effect of modality duration but retains channel-count effects (for example, sliding typically contains more channels), which can increase sliding's relative share. Per-element normalization provides the fairest cross-modality comparison and is therefore used for the main results (see Chapter 3). A direct comparison of normalization strategies is provided in Appendix H.

Second, modality percentages depend on aggregation across samples. Per-sample mean aggregation describes a "typical" sample, while pooled aggregation can be dominated by a subset of samples with unusually large attribution magnitudes. In this thesis, pooled aggregation is used as a robustness check and the per-sample mean is used for the main narrative. A direct comparison is provided in Appendix H.

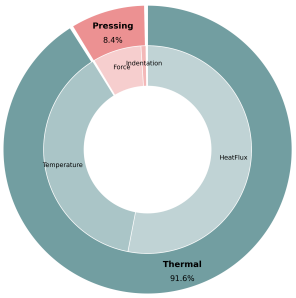
Third, participant-wise analyses show that overall modality usage trends remain similar across



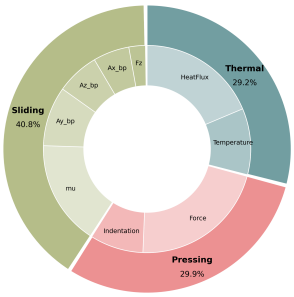
(a) Fabric



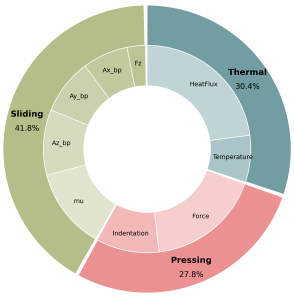
(b) Foam



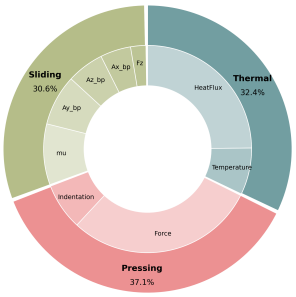
(c) Metal



(d) Paper



(e) Sandpaper



(f) Vinyl

(g) Wood

Figure 4.8: Model 3 per-class interpretability on the test set. The outer ring shows the relative attribution share of thermal, pressing, and sliding; the inner ring shows the channel-level breakdown within each modality. Larger outer-ring sectors indicate which modality contributes most strongly to each class prediction, while the inner ring should mainly be interpreted within the corresponding modality sector. Classes with a clearly dominant modality show a more specialized explanation profile, whereas mixed modality profiles indicate that the prediction depends on a broader combination of signals.

participants, with thermal contributions relatively stable and the remaining share shifting mainly between pressing and sliding. These participant-wise results, together with additional normalisation-effect checks, are reported in Appendix B. Finally, TSR regularization is used for scalar classification targets (Model 3 and the composite model) because it yields smoother and more stable time-series explanations; sample-level TSR examples and (where available) TSR vs. non-TSR comparisons are provided in Appendix H.

Supplementary performance experiments. In addition to the main multimodal results, three supplementary comparisons were conducted to better understand (i) the effect of adding participants, (ii) backbone choice, and (iii) the strength of individual modalities.

First, the material classifier was evaluated on a larger participant set. Using six participants, Model 3 remained highly stable and achieved comparable performance to the three-participant setting. This participant-coverage comparison and related normalisation effects are reported in Appendix B.

Second, a 1D-CNN backbone was compared to an LSTM backbone on the individual modality branches. The 1D-CNN performed better in all setups and is therefore used as the default backbone.

Third, unimodal baselines were evaluated for each modality (validation split). These unimodal results provide context for the multimodal performance and are included in Appendix B.

5 Discussion

This chapter discusses what the results suggest about multimodal tactile material recognition and interpretability. The main goal is not only to report performance, but also to understand which information the models rely on and whether these signals are consistent with prior work on tactile sensing and haptic perception. The detailed performance results and interpretability tables are reported in Chapter 4 and Appendix F. The interpretability method and summary choices (TSR-regularized Integrated Gradients, per-element normalization, and per-sample mean) are described in Chapter 3.

5.1. Main findings

This thesis addressed three main questions: how well multimodal interaction signals support tactile material classification, whether psychophysical adjective rating distributions can be predicted from those signals, and which parts of these signals contribute most to the final model decision. To address these questions, two modelling paths were considered: a direct multimodal classification route and an adjective-based route in which psychophysical rating distributions first served as an intermediate representation. In the final experiments, both paths relied on modality-specific 1D-CNN encoders. The trained model weights were subsequently analysed using Integrated Gradients, combined with Temporal Saliency Rescaling, to obtain saliency maps. These saliency maps were then summarized into nested donut charts at the modality, phase/bin, and channel levels.

Overall, the results showed that direct multimodal classification is the strongest approach for material classification. The adjective-based route remains valuable because it provides a more perceptually meaningful representation; however, its material classification performance was clearly weaker. This was reflected mainly in its inconsistency: some runs achieved performance close to that of the direct path, whereas others fell substantially below it.

The results further show that psychophysical adjective rating distributions can be predicted from the interaction signals, although with varying success across adjective pairs. This indicates that the measured signals encode meaningful perceptual information, even if this representation is not precise enough to outperform direct material classification.

In addition, the interpretability analysis showed a clear difference between the two modelling paths. The direct path relied on structured signal patterns across thermal, pressing, and sliding data that were consistent with meaningful interaction events and with prior literature, whereas Model 1 showed a stronger reliance on a more limited set of signals, often dominated by a single modality. This more restricted signal usage may help explain why the adjective-based route was less effective for final material classification.

More specifically, the attribution summaries suggest that the direct model formed its decisions from complementary evidence across modalities, with different materials depending on different combinations of thermal, deformation-related, and friction- or vibration-related signals. This implies that the learned representations capture meaningful aspects of tactile exploration rather than isolated or irrelevant signal fragments. Together, these findings show that multimodal tactile signals can support both accurate material classification and scientifically meaningful interpretation of model behaviour.

5.2. Material classification from multimodal interaction signals

5.2.1. Direct classification versus the adjective-mediated route

Although both paths use the same multimodal interaction signals, they differ in how the final classification decision is made: either directly from the input signals or indirectly through predicted adjective rating distributions derived from participant judgments. The direct path, implemented as Model 3, clearly outperformed the adjective-mediated route. In contrast to the 0.896 test accuracy and 0.885 macro-F1 of Model 3, the composite Model (1+2) achieved only 0.600 ± 0.115 test accuracy and 0.463 ± 0.137 macro-F1. This gap indicates that the input signals contain rich discriminative information, but that a substantial part of this information is lost when the signals are first compressed into predicted psychophysical rating distributions before final classification. Two likely explanations for this are information loss and error compounding. Predicting a full adjective distribution is already a more uncertain task than directly classifying a material, because these distributions reflect not only physical material properties but also subjectivity in human judgments. Any errors or uncertainty in the first model will then be used as an input for the second model, becoming additional noise for the final material classifier. In addition, perceptual descriptors are not uniquely tied to individual material classes: different materials can share similar perceptual profiles while still differing in signal details that remain useful for direct recognition. The adjective-based route is therefore scientifically valuable as a more human-interpretable intermediate representation, but less effective as a final classification pathway.

5.2.2. Why multimodal fusion helped

The results show that thermal, pressing, and sliding signals can all support material classification on the SENS3 dataset; however, the usefulness of these signals depends strongly on how they are represented and combined. The direct multimodal setting reached a test accuracy of 0.896 ± 0.017 and a macro-F1 of 0.885 ± 0.020. This represents a substantial improvement over the individually evaluated modality-specific backbones, which achieved validation accuracies of 0.740 ± 0.041 for the thermal encoder, 0.606 ± 0.044 for the pressing encoder, and 0.676 ± 0.073 for the sliding encoder. These single modality classification results were obtained using the maximum amount of available data, meaning the thermal and sliding backbones had 7 participants worth of data. Even with that advantage, however, none of the unimodal models matched the performance of the direct multimodal classifier. At the same time, the multimodal model had access to richer information within each individual sample, allowing each decision to be based on a more complete representation of the finger–material interaction.

5.2.3. Which materials were easier or harder to classify and why

The confusion matrices in Figures 4.1 and 4.2 provide a clearer view of how the two modelling paths differed in class-level performance. For the direct model, classification was strong across all retained classes, with Fabric and Metal classified perfectly. Paper remained the most difficult class. In contrast, the composite model showed much less stable behaviour at the class level. Only Fabric and Metal achieved relatively strong recognition, whereas the other classes were classified far less reliably. This indicates that the performance gap was not caused by a single failure class, but by a broader reduction in class separability.

A closer inspection of the off-diagonal entries shows that the composite model struggled particularly with materials that have overlapping perceptual profiles. Paper was often confused with Fabric and, to a lesser extent, with Foam and Vinyl. Sandpaper was most frequently misclassified as Fabric. Vinyl was often predicted as Wood, whereas Wood was often predicted as either Paper or Foam. These class-level patterns again show that the initial step of predicting adjective rating distributions does not preserve enough discriminative detail to separate materials that may feel partially similar along a small set of adjective dimensions. The direct model reduced most of these confusions substantially. The remaining errors were concentrated in a smaller number of class pairs, most notably Foam versus Wood and Paper versus Foam.

5.3. Prediction of psychophysical adjective distributions

5.3.1. What the distribution prediction results show

The results show that predicting psychophysical rating distributions from the interaction signals is feasible, although clearly more difficult than direct material classification. Across the five folds, Model

1 achieved an overall test KLD of 0.336 ± 0.020. At the level of individual adjective pairs, hot–cold achieved the lowest divergence at 0.294 ± 0.036, whereas hard–soft was the most difficult, with a test KLD of 0.371 ± 0.015. Rough–smooth and sticky–slippery lay in between, with test KLD values of 0.348 ± 0.021 and 0.331 ± 0.017, respectively. These values do not indicate a perfect match between the predicted and target distributions, however, they do show that the model captured meaningful structure in the human ratings rather than merely approximating a flat distribution. In addition, Kullback–Leibler divergence does not have a universal task-independent scale, since it is a relative-entropy measure and is unbounded in general. This means that a value around 0.3 can indicate relatively strong agreement in one setting and substantially weaker agreement in another. A useful intuition, introduced in Section 3.3.2, is that the overall test value of 0.336 corresponds to approximately 0.71 of the ideal geometric-mean likelihood of a perfect match. This suggests that Model 1 preserves a substantial part of the perceptual structure in the target distributions, but not so completely that the representation can be treated as near-lossless. This further supports the overall conclusion that Model 1 provides meaningful evidence that tactile signals encode human-relevant perceptual information, but does not preserve that information with sufficient precision to outperform direct signal-based classification.

5.3.2. What the class-level distributions suggest

The differences between adjective pairs are also informative. The strongest results were obtained for hot–cold. When examining the raw class-level rating distributions in Appendix A, these distributions appear relatively smooth and mostly single-peaked across classes. Combined with the fact that the thermal modality directly measures temperature and heat-flux dynamics, which are closely related to the physical process underlying warm–cold perception, it is reasonable that this adjective pair was the easiest to model. Hard–soft, in contrast, showed much greater irregularity in the class-level target distributions. Multiple material classes show broad, multi-peaked or irregular distributions. This makes the task more difficult even before the sensor input is considered. The perceptual structure is less clean and less concentrated. Another possible explanation is that compliance-related perception is harder to capture cleanly from the available signal representation, even though pressing information is included. Part of that difficulty may come from the fact that hardness perception is influenced not only by the recorded force and indentation signals, but also by subtler aspects of contact geometry and exploration that are not fully represented in the current setup. Rough–smooth and sticky–slippery occupy an intermediate position. Their raw distributions show more structure than hard–soft, but they also remain broader than a simple one-peak target. For some materials, noticeable shifts or secondary modes can be observed. This suggests that these adjective pairs still contain useful material-specific information, but not in a perfectly clean or one-dimensional way. This is consistent with the likely physical origin of these ratings: unlike hot–cold, they are not tied primarily to one especially direct sensory measurement, but instead emerge from combinations of sliding-related friction, vibration, and contact effects.

5.3.3. What Model 2 interpretability reveals about the adjective bottleneck

The most important takeaway from the Model 2 interpretability results is that the adjective-based classifier does not use the different predicted adjective ratings evenly. Instead, it relies mainly on a narrower subset of adjective pairs, with hard–soft emerging as the dominant decision axis for most classes. This shows what happens when the original multimodal signals have been compressed into psychophysical distributions: the downstream classifier is forced to organize material separation around the few perceptual dimensions that remain most discriminative in that reduced space.

This is especially interesting when comparing these adjective contribution results to the direct path. In the direct multimodal classifier, thermal information was the most important modality overall, whereas in Model 2 the corresponding perceptual descriptor, hot–cold, was the least influential adjective pair. This contrast suggests that the raw thermal signals contained class-relevant information that was not fully preserved once they were reduced to a single perceptual dimension. This may already help explain the performance gap between the direct and adjective-mediated routes. Although thermal perception was relatively easy to predict in Model 1, it seems to be less useful for separating material classes once represented in adjective space. This interpretation is also consistent with the raw participant distributions, where hot–cold showed relatively smooth and overlapping class-level patterns compared with the other retained adjective pairs. At the same time, hard–soft showed the opposite behaviour. It was the most difficult adjective pair to predict accurately in Model 1, yet it became the most influential

input for Model 2. This suggests that the adjective-based route relies most strongly on a perceptual dimension that is itself only imperfectly recovered from the sensor data. Such a mismatch is likely to increase sensitivity to upstream prediction error and may therefore contribute to the instability of the composite pipeline. The class-specific results further support this interpretation. Most classes were drawn towards a hard-soft dominated representation, which indicated that Model 2's decisions are mainly organised around one decision axis. Fabric showed a more balanced dependency, which is plausible given that fabric perception is generally multidimensional rather than dominated by one descriptor alone. Sandpaper was the clearest exception, from a human perception perspective, one might expect rough-smooth to be the most important pair, however the model relied more on sticky-slippery. This suggests that within the retained class set, tangential resistance-related differences may have separated Sandpaper more clearly than perceived roughness alone. The adjective that seems most intuitively tied to a material is not always the one that best separates that class from the others once the representation has been compressed.

5.4. Interpretation of learned multimodal signal usage

5.4.1. Overall modality use

At the broadest level, the most important interpretability finding is the clear contrast between the two modelling paths. Model 3 showed a relatively balanced multimodal explanation profile, with test-set attribution distributed across thermal, pressing, and sliding rather than collapsing predominantly onto one source. On the held-out test set, the overall split was approximately 39.9% thermal, 34.0% pressing, and 26.1% sliding. In contrast, the composite Model (1+2) was much more strongly dominated by sliding, with test-set attribution of 62.5% sliding, 24.2% thermal, and 13.4% pressing. This suggests that the stronger direct classifier relied on a richer combination of evidence, whereas the weaker adjective-mediated route depended on a narrower signal basis.

5.4.2. Phase and channel level explanations

The modality-level view becomes more convincing when it is supported by structure within each modality. For thermal signals, the main pattern across models is that the attribution is concentrated in the transient contact phases rather than in the longer phase of steady contact. In Model 1, thermal attribution was strongest around the peak and half-equilibration phases, and similar behaviour was observed for the composite model and Model 3. This is physically plausible, because early contact contains the strongest heat-transfer dynamics, whereas steady contact contains less rapidly changing information. The fact that all models consistently emphasize these transient thermal phases rather than the full contact duration suggests that they rely on meaningful thermal evidence rather than arbitrary portions of the signal.

For pressing, the attribution structure differed slightly between models. Model 1 placed most emphasis on the plateau and lift-off phases, whereas the composite model appeared to rely more on the loading and lift-off phases. Both of these models relied on a more restricted subset of the pressing signal. Model 3 showed a more even use of the different phases, suggesting a more complete use of the press–hold–release cycle.

For sliding, the main recurring result is that most attribution was concentrated in the light-contact bins, particularly at low forces across low-to-mid speed ranges. This pattern appears in Model 1, in the composite model, and in Model 3. This is a meaningful finding rather than a trivial consequence of the analysis. Since the sliding branch contains friction-related and vibration-related signals, it is plausible that useful class information emerges most clearly when contact is light enough for texture-dependent dynamics to remain visible rather than being masked by stronger normal-force effects.

The channel-level summaries add an important extra layer of specificity, because they move the interpretation from the level of broad modality labels to the level of the individual physical measurements on which the models rely. For Model 3, the thermal branch was driven more strongly by heat flux than by temperature, which is meaningful because heat flux captures the initial rate of thermal exchange during contact and therefore reflects the transient thermal response of the material more directly than the absolute skin temperature trace alone. This fits well with the earlier phase-level observation that thermal relevance is concentrated in the transient part of the interaction rather than in prolonged steady contact.

Pressing attribution was more strongly dominated by normal force than by indentation. This ob-

servation is somewhat counter-intuitive because the pressing protocol was defined around a target force of 3 N, meaning the maximum normal force should be broadly similar across materials, while the indentation reached at that force may differ depending on material compliance. One might therefore expect indentation to provide more useful information for material classification. To investigate this further, additional qualitative analysis was carried out on the raw pressing signals (Appendix I). These overlays show that indentation does vary across different materials, but that the trajectories are relatively smooth and can differ across participants and trials. The normal force signals, by contrast, retain more local temporal variation during loading, holding and release. This might explain why the learned representation relied more strongly on normal force, even though indentation still reflects underlying material-dependent deformation.

The sliding branch shows the richest internal structure. Attribution was distributed mainly across the friction coefficient and the band-pass filtered IMU channels, while normal force contributed relatively little. This is an important result, because it indicates that the model is not basing its sliding decisions primarily on the maintained load itself, but on the dynamic consequences of sliding contact. The friction coefficient captures how strongly the finger interacts tangentially with the surface, while the filtered acceleration channels reflect the vibration patterns generated during motion. Together, these channels are closely tied to texture-related and friction-related tactile signals, which are known to be informative during lateral exploration. The relatively small contribution of sliding normal force is therefore meaningful: it suggests that once contact is established, the model depends more on the material-specific friction and vibration response than on the magnitude of the applied load alone.

5.4.3. Class-specific signal usage

In Model 3, the modality balance was not fixed across classes, but shifted in a way that broadly matched the type of information expected to separate those materials during human touch. Metal is the clearest example: its explanations are strongly thermal-driven, which is highly plausible given that metallic surfaces tend to extract heat quickly from the skin and therefore produce distinctive thermal transients during early contact (Bergmann Tiest and Kappers, 2009; Ho, 2017). This also aligns with the strong classification performance for Metal, suggesting that the model benefited from a relatively clear and reliable signal source.

For fabrics, tactile perception is generally not tied to one modality only. Prior work instead describes fabric touch in terms of multiple contributing properties, such as compression softness, bending stiffness, stretching-related behaviour, and thermal comfort. Fabric materials should therefore be viewed as perceptually multidimensional rather than as depending on a single dominant signal source (Kayseri et al., 2012). The more pressing-dominant attribution pattern found for Fabric in the present experiments should therefore be interpreted cautiously. Rather than indicating that pressing is generally the main signal for textile perception, it more likely reflects the fact that the classifier was optimized for class separation within this dataset. If the fabric samples were especially distinguishable through their pressing response, then the model could legitimately prioritize that modality, even though broader textile perception in humans is often based on a richer combination of mechanical, surface, and thermal signals.

Although Foam is mechanically compliant, Model 3 does not rely mainly on pressing for this class, but instead shows a mixed contribution from thermal and sliding. This is not necessarily inconsistent with foam being “soft”: in this dataset the foam samples have a sponge-like surface with holes and irregularities, which can produce distinctive friction and vibration patterns during sliding. In addition, many foams are relatively insulating, meaning they transfer heat more slowly than rigid materials. This can make foam feel comparatively “warmer” upon contact and can also create characteristic thermal dynamics in the heat-flux and temperature signals. Together, these surface and thermal properties provide plausible alternative signals that may be more consistent across trials than the pressing features currently used, explaining why the model attributes substantial importance to sliding and thermal channels for foam.

Paper and Sandpaper showed a stronger reliance on sliding. This aligns with the idea that these materials are often distinguished by texture-related signals and vibration signatures during motion (Hollins and Risner, 2000; Hollins et al., 2002).

Wood and Vinyl showed a more mixed signal profile. These classes are naturally harder to distinguish because there is no single dominant signal. The model must combine weaker signals across modalities, which can make it more sensitive to dataset size, noise, and contact variation.

5.5. Limitations and future work

The dataset is relatively small and imbalanced across classes. This limitation likely restricts generalization and increases fold-to-fold variance, especially for the harder classes. A logical next step is to expand the dataset with more samples per class, particularly for the underrepresented classes, to increase participant coverage, and to report performance under participant-held-out evaluation.

Many contact conditions were measured but not fully controlled. In particular, sliding and friction signals depend strongly on force, speed, skin state, and local contact mechanics (Gueorguiev et al., 2017), and although force and speed are available in the SENS3 dataset, they were still participant-controlled during data collection and therefore remain variable across trials and participants. Maintaining a consistent force–speed profile during active exploration is itself difficult, meaning that part of the recorded variation likely reflects differences in execution rather than material properties alone. Future work could therefore (i) use force and speed more explicitly as conditioning variables within the model, (ii) stratify evaluation by force–speed regimes to separate material effects from interaction variability, and (iii) introduce robustness tests by perturbing force and speed within realistic bounds to quantify how stable predictions and explanations remain.

Some useful factors were not measured or synchronized with the signals. For example, participant state (skin temperature and hydration) and exploration strategy can influence tactile signals and perceived ratings (Ho, 2017). Collecting such metadata would enable stronger analyses of variance sources, and could improve modelling by allowing explicit conditioning or domain-adaptation techniques to reduce participant-specific effects.

Not all interpretability variants were available for all models. In particular, TSR-based interpretability is naturally defined for scalar targets, while Model 1 outputs distributions, making direct TSR comparisons less straightforward. A future direction is to define distribution-aware attribution targets (e.g., expected rating, distribution moments, or task-relevant contrasts between bins) and then evaluate whether the resulting explanations remain stable and faithful. In addition, faithfulness checks such as deletion/insertion (occlusion) tests could be added to quantify whether removing highly attributed segments reliably reduces the model score.

6

Conclusion

This thesis addressed three questions: (1) how well thermal, pressing, and sliding signals support tactile material classification; (2) how accurately psychophysical rating distributions can be predicted from these signals; and (3) whether attribution-based explanations identify physically meaningful evidence in interaction time series.

For material classification, the central finding is that direct end-to-end learning from multimodal interaction signals is the most effective approach on SENS3. The late-fusion multimodal classifier (Model 3) achieves 0.896 test accuracy across seven retained material classes and shows stable performance across folds. In contrast, the adjective-mediated "interpretable bottleneck" route (Model 1 → Model 2) performs substantially worse. This gap supports the conclusion that reducing high-dimensional time-series interaction evidence to a small set of perceptual distributions can discard discriminative information, and that any errors in the intermediate perceptual predictions propagate to the final classification.

For perceptual modeling, predicting full adjective rating distributions from tactile interaction signals is feasible but more ambiguous than classification. Model 1 demonstrates that distributional outputs can be learned from finger–material interactions, yet performance varies by adjective pair, reflecting both the complexity of mapping physical signals to perception and the presence of inter-participant variability. These results position perceptual distribution prediction as a promising direction, but one that likely requires more data, richer conditioning on exploration parameters, and uncertainty-aware training objectives to close the gap with material recognition.

For interpretability, the thesis shows that post-hoc attribution can be made actionable and physically grounded by aggregating raw time–channel relevance into structured summaries. Integrated Gradients (and TSR for scalar classifiers) provides sample-level attributions that, when pooled into modality-, phase/bin-, and channel-level shares, yield explanations that are consistent across folds and interpretable at the level of interaction events. The strongest classifier exhibits a balanced, class-conditional signal strategy across thermal, pressing, and sliding, rather than relying on a single modality. Within thermal data, attributions concentrate in early-to-mid transient phases, consistent with heat-transfer dynamics being informative. Within pressing data, evidence is distributed across loading, plateau, and lift-off, indicating that the model exploits the full contact dynamics. Within sliding data, relevance peaks in low-force regimes across multiple speed ranges, suggesting a reliance on light-contact friction/vibration signatures. Per-class analyses further reinforce plausibility: Metal is predominantly thermal-driven and yields high performance, whereas mixed-modality classes (e.g., Wood/Vinyl/Sandpaper) show more distributed evidence profiles and correspondingly higher confusion potential.

Several limitations remain. Dataset size and class imbalance constrain generalization, and sliding signals are sensitive to force, speed, and finger condition. Not all interpretability variants are equally comparable across output types (e.g., scalar classification versus distribution prediction), and some methodological choices (such as target selection for distribution outputs) affect explanation form. Future work should expand participant coverage with fully matched multimodal trials, explicitly condition models on interaction parameters (force/speed/skin state), and evaluate robustness under controlled perturbations and domain shifts. Methodologically, combining uncertainty-aware perceptual objectives with phase- or concept-aware modeling, while retaining faithful post-hoc evaluation, could further strengthen the link between human perception and model evidence.

In summary, this thesis demonstrates that multimodal finger–material interaction signals enable strong material recognition, and that structured attribution aggregation can produce explanations that map model evidence onto meaningful interaction phases and signal sources. This combination of performance and interpretability is a step toward reliable tactile intelligence for haptic interfaces and embodied robotic systems.

Bibliography

Balasubramanian, J. K., Kodak, B. L., & Vardar, Y. (2024, November). Sens3: Multisensory database of finger-surface interactions and corresponding sensations. In *Haptics: Understanding touch; technology and systems; applications and interaction* (pp. 262–277). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70058-3_21

Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A. T., & Bizarro, P. (2021). TimeSHAP: Explaining Recurrent Models through Sequence Perturbations [arXiv:2012.00073 [cs]], 2565–2573. https://doi.org/10.1145/3447548.3467166

Bergmann Tiest, W. M. (2010). Tactual perception of material properties [Perception and Action: Part I]. *Vision Research*, *50*(24), 2775–2782. https://doi.org/https://doi.org/10.1016/j.visres.2010.10.005

Bergmann Tiest, W. M., & Kappers, A. M. L. (2009). Tactile perception of thermal diffusivity. *Attention, Perception, & Psychophysics*, *71*(3), 481–489. https://doi.org/10.3758/APP.71.3.481

Boureau, Y.-L., Ponce, J., & Lecun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, 111–118.

Butterworth, S. (1930). On the Theory of Filter Amplifiers. *Experimental Wireless & the Wireless Engineer*, *7*, 536–541.

Cacciari, I., & Ranfagni, A. (2024). Hands-on fundamentals of 1d convolutional neural networks—a tutorial for beginner users. *Applied Sciences*, *14*, 8500. https://doi.org/10.3390/app14188500

Cao, G., Jiang, J., Bollegala, D., Li, M., & Luo, S. (2023, June). Multimodal Zero-Shot Learning for Tactile Texture Recognition [arXiv:2306.12705 [cs]]. https://doi.org/10.48550/arXiv.2306.12705

Devillard, A., Ramasamy, A., Faux, D., Hayward, V., & Burdet, E. (2023). Concurrent Haptic, Audio, and Visual Data Set During Bare Finger Interaction with Textured Surfaces. *2023 IEEE World Haptics Conference (WHC)*, 101–106. https://doi.org/10.1109/WHC56415.2023.10224372

Felicetti, L., Sutter, C., Chatelet, E., Latour, A., Mouchnino, L., & Massi, F. (2023). Tactile discrimination of real and simulated isotropic textures by friction-induced vibrations. *Tribology International*, *184*, 108443. https://doi.org/https://doi.org/10.1016/j.triboint.2023.108443

Fishel, J., & Loeb, G. (2012). Bayesian exploration for intelligent identification of textures. *Frontiers in neurorobotics*, *6*, 4. https://doi.org/10.3389/fnbot.2012.00004

Gibbs, J. K., Gillies, M., & Pan, X. (2022). A comparison of the effects of haptic and visual feedback on presence in virtual reality. *International Journal of Human-Computer Studies*, *157*, 102717. https://doi.org/https://doi.org/10.1016/j.ijhcs.2021.102717

Gueorguiev, D., Vezzoli, E., Mouraux, A., Lemaire-Semail, B., & Thonnard, J.-L. (2017). The tactile perception of transient changes in friction. *Journal of the Royal Society Interface*, *14*(137), 20170641. https://doi.org/10.1098/rsif.2017.0641

Hassan, W., Joolee, J. B., & Jeon, S. (2023). Establishing haptic texture attribute space and predicting haptic attributes from image features using 1D-CNN. *Scientific Reports*, *13*(1), 11684. https://doi.org/10.1038/s41598-023-38929-6

Ho, H.-N. (2017). Material recognition based on thermal cues: Mechanisms and applications. *Temperature*, *5*(1), 36–55. https://doi.org/10.1080/23328940.2017.1372042

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hollins, M., Bensmaïa, S. J., & Roy, E. A. (2002). Vibrotaction and texture perception. *Behavioural Brain Research*, *135*(1-2), 51–56. https://doi.org/10.1016/S0166-4328(02)00154-7

Hollins, M., & Risner, S. R. (2000). Evidence for the duplex theory of tactile texture perception. *Perception & Psychophysics*, *62*(4), 695–705. https://doi.org/10.3758/BF03206916

Ige, A. O., & Sibiya, M. (2024). State-of-the-art in 1d convolutional neural networks: A survey. *IEEE Access*, *12*, 144082–144105. https://doi.org/10.1109/ACCESS.2024.3433513

Ioffe, S., & Szegedy, C. (2015, July). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 448–456, Vol. 37). PMLR. https://proceedings.mlr.press/v37/ioffe15.html

Ismail, A. A., Gunady, M., Bravo, H. C., & Feizi, S. (2020, October). Benchmarking Deep Learning Interpretability in Time Series Predictions [arXiv:2010.13924 [cs]]. https://doi.org/10.48550/arXiv.2010.13924

Ismail, A. A., Gunady, M., Corrada Bravo, H., & Feizi, S. (2020). Benchmarking deep learning interpretability in time series predictions. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 6441–6452, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/47a3893cc405396a5c30d913Paper.pdf

Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)*, *349*, 255–60. https://doi.org/10.1126/science.aaa8415

Kayseri, G. Ö., Özdil, N., & Süpüren Mengüç, G. (2012). Sensorial comfort of textile materials. In H.-Y. Jeon (Ed.), *Woven fabrics* (pp. 235–266). InTechOpen. https://doi.org/10.5772/37596

LeCun, Y., Yere, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–44. https://doi.org/10.1038/nature14539

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. https://arxiv.org/abs/1312.4400

Luo, S., Yuan, W., Adelson, E., Cohn, A. G., & Fuentes, R. (2018, March). ViTac: Feature Sharing between Vision and Tactile Sensing for Cloth Texture Recognition [arXiv:1802.07490 [cs]]. https://doi.org/10.48550/arXiv.1802.07490

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9). https://doi.org/10.3390/info15090517

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, *55*(13s), 1–42. https://doi.org/10.1145/3583558

Pham, A.-D., Kuestenmacher, A., & Ploeger, P. G. (2022, August). TSEM: Temporally Weighted Spatiotemporal Explainable Neural Network for Multivariate Time Series [arXiv:2205.13012 [cs]]. https://doi.org/10.48550/arXiv.2205.13012

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. https://arxiv.org/abs/2207.13243

Richardson, B. A., & Kuchenbecker, K. J. (2020). Learning to Predict Perceptual Distributions of Haptic Adjectives [Publisher: Frontiers Media SA]. *Frontiers in Neurorobotics*, *13*. https://doi.org/10.3389/fnbot.2019.00116

scikit-learn developers. (2026). *Stratifiedkfold* [Accessed: 2026-04-09]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

Shultz, C., Colgate, E., & Peshkin, M. (2018). The application of tactile, audible, and ultrasonic forces to human fingertips using broadband electroadhesion. *IEEE Transactions on Haptics*, *PP*, 1–1. https://doi.org/10.1109/TOH.2018.2793867

Siddiqui, S. A., Mercier, D., Dengel, A., & Ahmed, S. (2020, April). TSInsight: A local-global attribution framework for interpretability in time-series data [arXiv:2004.02958 [cs]]. https://doi.org/10.48550/arXiv.2004.02958

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929–1958.

Sundararajan, M., Taly, A., & Yan, Q. (2017, June). Axiomatic attribution for deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 3319–3328, Vol. 70). PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html

Wang, J., Zhang, R., & Li, Q. (2025). TF-LIME : Interpretation Method for Time-Series Models Based on Time–Frequency Features [Publisher: MDPI AG]. *Sensors*, *25*(9), 2845. https://doi.org/10.3390/s25092845

Zhang, X., Li, S., Yang, J., Bai, Q., Wang, Y., Shen, M., Pu, R., & Song, Q. (2021). Target Classification Method of Tactile Perception Data with Deep Learning. *Entropy*, *23*(11), 1537. https://doi.org/10.3390/e23111537

A

Psychophysical target distributions

This appendix shows the empirical class-level rating distributions for all psychophysical adjective pairs available in the SENS3 rating data. The figures are split into the adjective pairs retained for the final modelling setup and the adjective pairs that were excluded after exploratory inspection of the target distributions.

A.1. Selected adjective pairs

A.2. Excluded adjective pairs

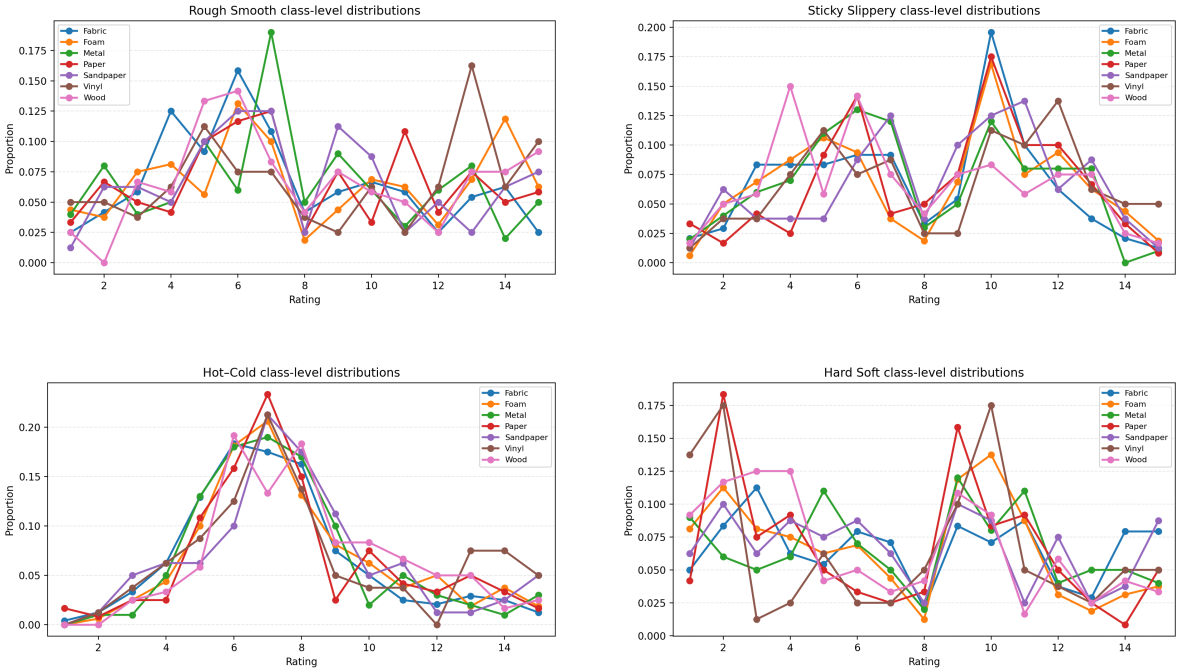


Figure A.1: Empirical class-level rating distributions for the adjective pairs retained in the final modelling setup: *rough–smooth*, *sticky–slippery*, *hot–cold*, and *hard–soft*.

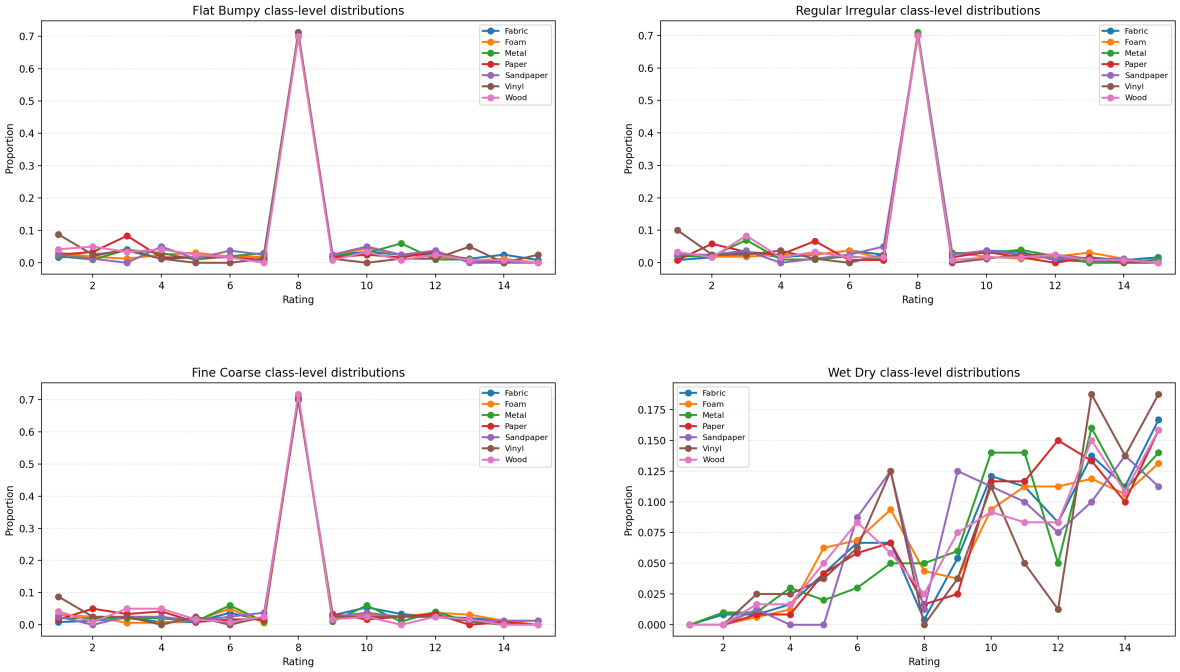


Figure A.2: Empirical class-level rating distributions for the adjective pairs excluded from the final modelling setup: *flat–bumpy*, *regular–irregular*, *fine–coarse*, and *wet–dry*.

B

Participant-wise performance and normalisation effects

To assess how well the thermal classifier generalises across individuals, a series of experiments was run in which the model was trained on subsets of participants and evaluated separately for each participant in the validation set.

Figure B.1 shows participant-specific validation accuracy as a function of the number of participants included in the training subset. Each coloured curve corresponds to a held-out participant (P1–P7). Accuracy generally increases when moving from one to two training participants and then fluctuates mildly as additional participants are added. Participants P2 and P4 reach high accuracies (often above 0.8) even with relatively small training subsets, whereas P3 and P5 are more challenging, with accuracies that remain lower and more sensitive to the exact subset composition. This suggests that some participants exhibit more idiosyncratic interaction patterns, requiring more diverse training data to capture reliably.

A complementary view is given in Figure B.2, which summarises, for each participant, the mean validation accuracy and standard deviation across all subset configurations. Participant 4 achieves the highest mean accuracy (around 0.87), followed by Participants 1 and 2 (around 0.75–0.78), while Participants 3 and 5 show lower mean accuracies (around 0.60–0.63) and larger variability. Overall, the spread across participants is moderate: even the most difficult participants remain well above chance, but there is a clear gap between the easiest (P4) and hardest (P3, P5) individuals.

To quantify the aggregate effect of adding more participants to the training data, the mean validation accuracy was averaged over all held-out participants for each subset size. Figure B.3 shows this *scaling curve* for the globally normalised model. Accuracy rises sharply from one to two training participants and then oscillates around 0.7–0.75 as more participants are added, with relatively wide confidence bands for small subset sizes. This indicates diminishing returns beyond about four participants, but still some instability, particularly when the training subset is small.

Figure B.4 reports the same analysis for the per-participant z-score normalisation. The overall shape of the scaling curve is similar, but two improvements are visible: (i) average accuracy is slightly higher for most subset sizes, particularly when six or seven participants are included, and (ii) the standard deviation band is narrower, indicating more stable performance across different subset compositions. This supports the main motivation for per-participant normalisation: by removing systematic offsets in baseline skin temperature and contact behaviour, the model becomes less sensitive to which specific participants are used for training and generalises more consistently to new individuals.

Taken together, these analyses show that (i) cross-participant generalisation is feasible but not uniform across individuals, and (ii) per-participant normalisation modestly improves both the level and the stability of validation accuracy, especially when training on larger, more heterogeneous participant sets.

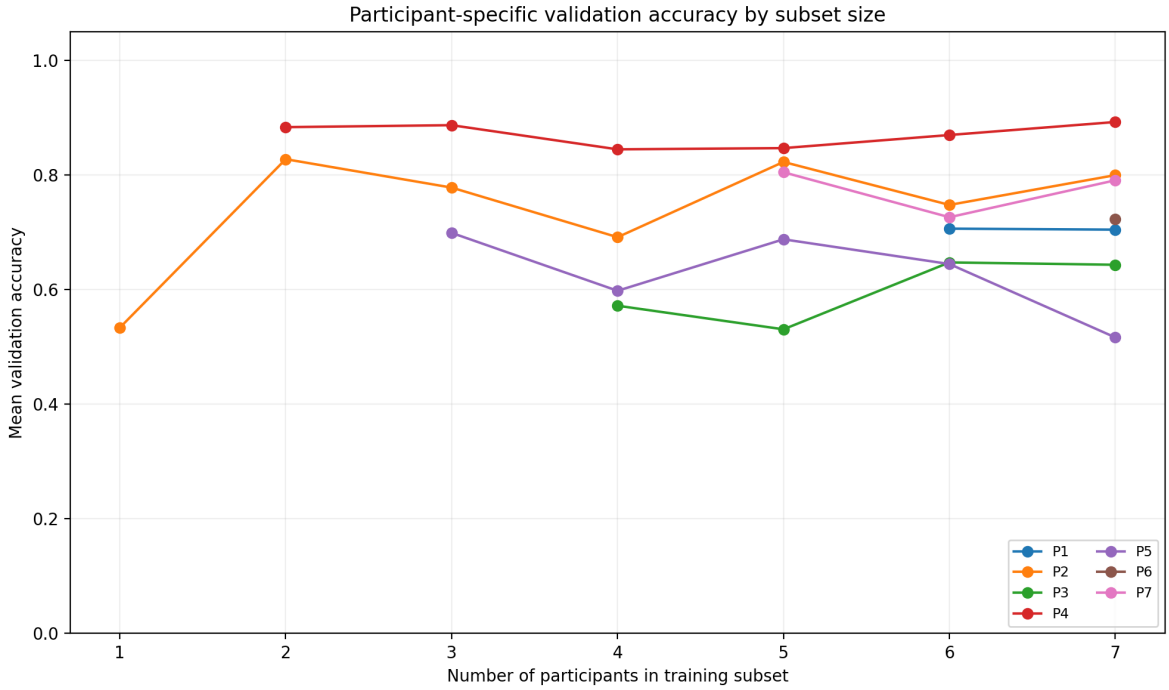


Figure B.1: Participant-specific validation accuracy as a function of the number of participants included in the training subset. Each curve corresponds to a different held-out participant.

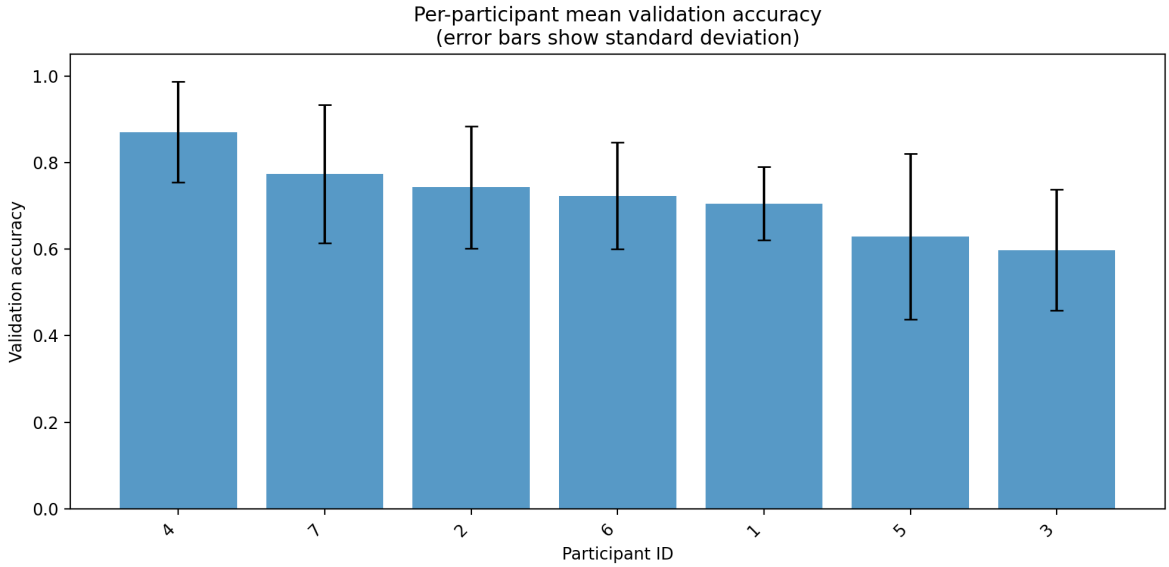


Figure B.2: Per-participant mean validation accuracy with standard deviation across all training-subset configurations.

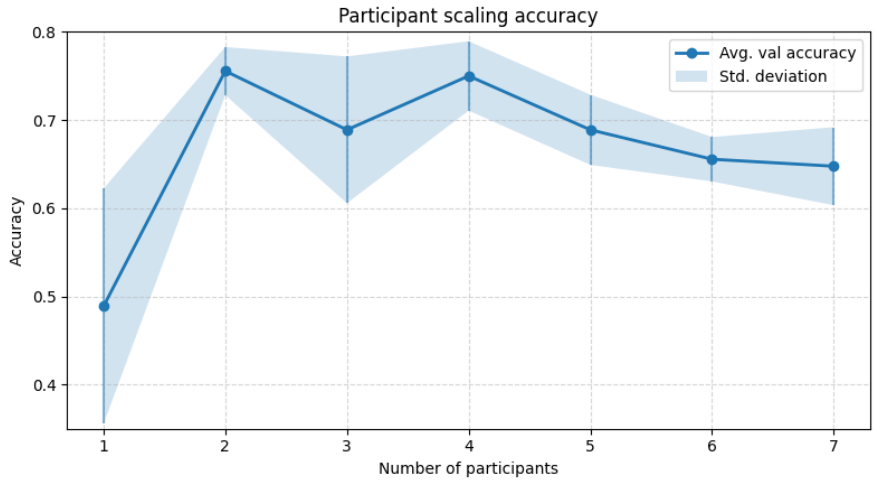


Figure B.3: Average validation accuracy across participants as a function of the number of participants in the training set, using global normalisation. The shaded region denotes one standard deviation.

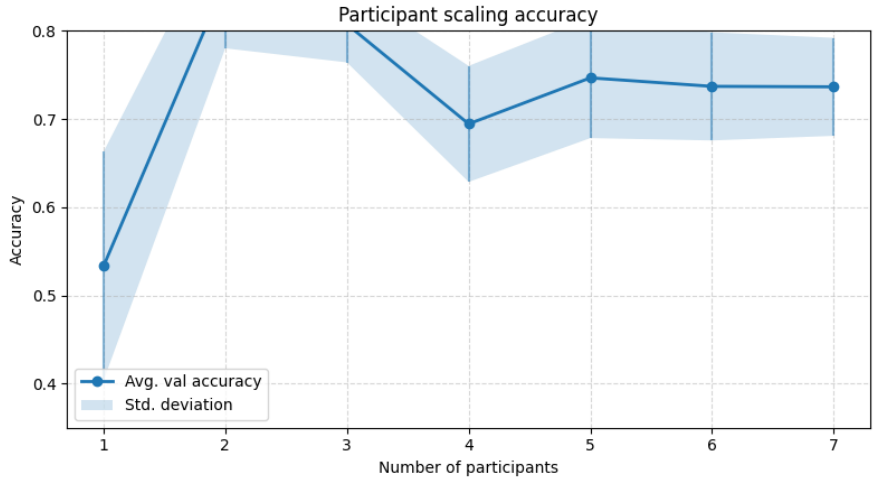


Figure B.4: Average validation accuracy across participants as a function of the number of participants in the training set, using per-participant normalisation. The shaded region denotes one standard deviation.

C

Pipeline details

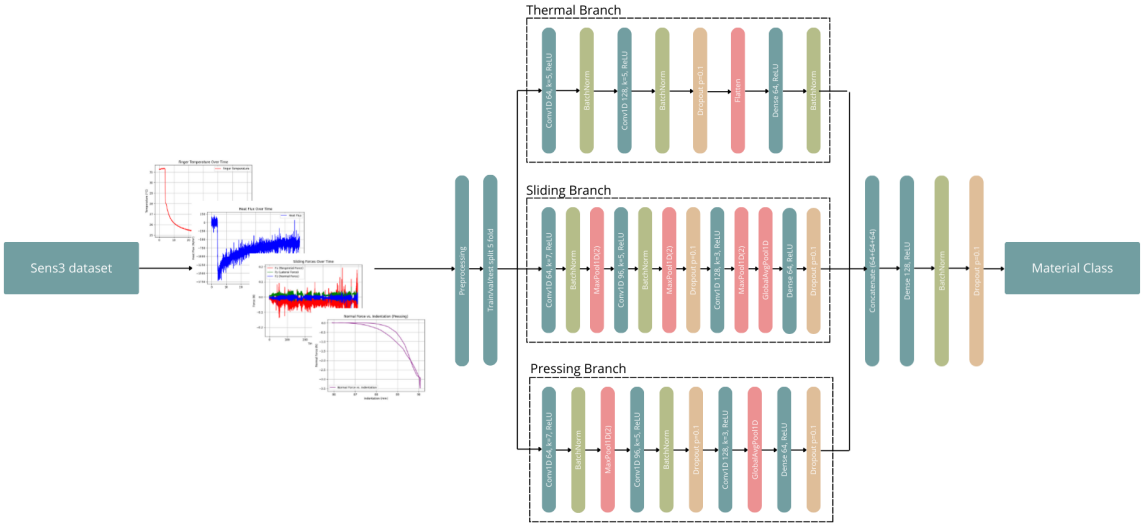


Figure C.1: Detailed pipeline for Model 3, showing the internal processing blocks (per-modality encoders, fusion operation, and classification head).

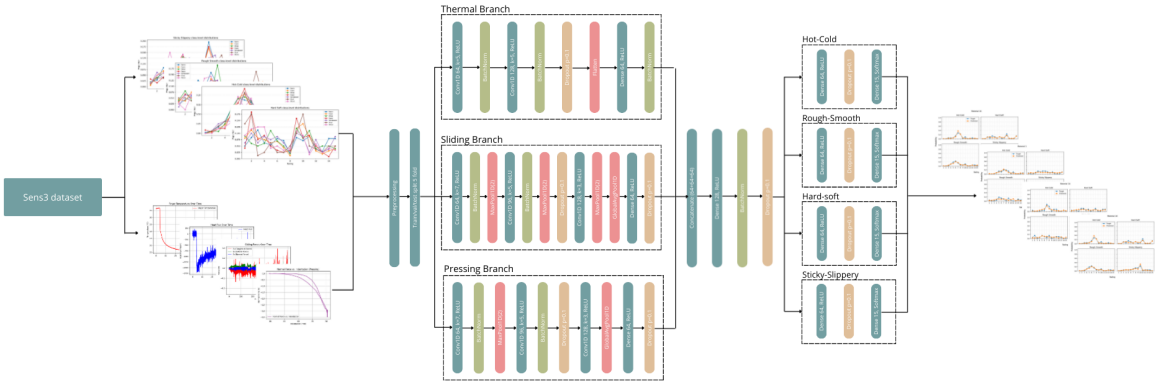


Figure C.2: Detailed pipeline for Model 1, showing the internal processing blocks (per-modality encoders, shared representation, and adjective-specific distribution heads).

D

Hyperparameter and Architecture Tuning

D.1. Best configurations by validation accuracy

Table D.1 summarises the best-performing configurations from the hyperparameter sweep, ranked by mean validation accuracy over k folds.

Table D.1: Top hyperparameter configurations ranked by mean validation accuracy over k folds.

Rank	lr	batch	dropout	patience	arch	k	mean val acc $\pm$ std
1	5×10^{-4}	8	0.10	8	base	5	0.6952 ± 0.0156
2	5×10^{-4}	8	0.10	12	base	5	0.6794 ± 0.0273
3	5×10^{-4}	4	0.10	12	base	5	0.6571 ± 0.0078
4	1×10^{-3}	4	0.05	4	base	5	0.6317 ± 0.0367
5	5×10^{-4}	4	0.10	4	base	3	0.5968 ± 0.0224

Mean best epoch for the top configuration: 21.33; for the second: 39.00; for the third: 25.33.

D.2. Observed patterns

D.2.1. Architecture.

Across the sweep, the *base* CNN architecture (two convolutional blocks followed by Flatten) consistently outperformed both a deeper variant (with an additional convolutional layer) and a *pool* variant using MaxPooling + Global Average Pooling (GAP). The pooling-based models not only achieved lower mean validation accuracy but also exhibited much higher variability (standard deviations on the order of 0.22–0.24 in some runs), suggesting instability or underfitting due to over-aggressive temporal compression.

D.2.2. Kernel size.

Within the base architecture, a kernel size of $k = 5$ performed better than $k = 3$, indicating that slightly longer temporal context helps the model capture the relevant dynamics in the thermal traces.

D.2.3. Batch size.

Among the better-performing settings, a batch size of 8 systematically outperformed 4. This likely reflects a more favourable gradient signal-to-noise trade-off and interacts well with the chosen early-stopping patience.

D.2.4. Learning rate.

A learning rate of 5×10^{-4} yielded the best overall performance and stable convergence. Higher (10^{-3}) and lower (10^{-4}) learning rates underperformed in terms of mean validation accuracy and/or robustness across folds.

D.2.5. Early-stopping patience.

Patience values of 8 and 12 both produced competitive models, but the top-ranked configuration used patience $= 8$ (mean validation accuracy ≈ 0.70) compared to ≈ 0.68 for patience $= 12$ with otherwise identical settings. Longer patience sometimes led the optimiser to "chase" late-epoch noise and mild overfitting, whereas the best runs tended to converge within roughly 20–25 epochs.

D.2.6. Dropout.

Within this sweep, a dropout rate of 0.1 worked best. Higher dropout did not improve performance, suggesting that current performance is limited more by architecture and data (amount/variability) than by insufficient regularisation.

.

E

Final hyperparameters and training settings

This appendix lists the final hyperparameters used for each model. Unless stated otherwise, results are reported over 5 folds.

E.1. Model 1: Perceptual distribution prediction

Scripts: `model1_3participants.py`

Table E.1: Final hyperparameters for Model 1 (distribution prediction).

Setting	Value
Batch size	8
Epochs	60
Loss	KLD
Learning rates	`lr_thermal` $= 5 \cdot 10^{-4}$, `lr_sliding` $= 10^{-3}$, `lr_pressing` $= 10^{-3}$, `lr_fusion` $= 10^{-3}$
Dropout	`dropout_thermal` $= 0.10$, `dropout_sliding` $= 0.10$, `dropout_pressing` $= 0.10$, `fusio`
Normalization	`norm_thermal` = `per_participant`, `norm_sliding` = `global`, `norm_pressing` = `per_`
Cross-validation	$k = 5$ folds
Early stopping	enabled, patience = 10
ReduceLROnPlateau	enabled
ReduceLR factor	0.5 (thermal, sliding, pressing, fusion)
ReduceLR patience	5 (thermal, sliding, pressing, fusion)
ReduceLR min lr	10^{-6} (thermal, sliding, pressing, fusion)
Plotting (diagnostics)	`plot_all_validation=True`, `plot_chunk_size=48`

Additional setting for `model1_3participants.py`: a fixed test split of `test_size` $= 0.2$ is held out *before* cross-validation.

E.2. Model 2: Material classification from perceptual distributions

Script: `model2.py`

E.3. Model 3: Direct multimodal material classification

Script: `model3_3participants.py`

Data split: a fixed global train+val/test split of `test_size` $= 0.2$ is applied before cross-validation.

Table E.2: Final hyperparameters for Model 2 (MLP classifier on adjective distributions).

Setting	Value
Hidden layers	`hidden1` $= 32$, `hidden2` $= 0$, `hidden3` $= 0$
Activation	ReLU
Batch normalization	enabled
Dropout	0.30
L2 regularization	10^{-4}
Label smoothing	0.20
Learning rate	$9.422992779213991 \cdot 10^{-4}$
Epochs	180
Batch size	8
Validation split	0.20
Early stopping	enabled, monitor=`val_loss`, patience=12, restore best weights
ReduceLROnPlateau	enabled, monitor=`val_loss`
ReduceLR factor	0.20
ReduceLR patience	5
ReduceLR min lr	10^{-6}
ReduceLR min delta	10^{-4}

Table E.3: Final hyperparameters for Model 3 (direct multimodal classifier).

Setting	Value
Batch size	8
Epochs	60
Learning rates	`lr_thermal` $= 5 \cdot 10^{-4}$, `lr_sliding` $= 10^{-3}$, `lr_pressing` $= 10^{-3}$, `lr_fusion` $= 1$
Dropout	`dropout_thermal` $= 0.15$, `dropout_sliding` $= 0.15$, `dropout_pressing` $= 0.15$,
Normalization	`norm_thermal` $=$ `per_participant`, `norm_sliding` $=$ `global`, `norm_pressing` $=$
Cross-validation	$k = 5$ folds
Early stopping	disabled
ReduceLROnPlateau	enabled
ReduceLR factor	0.5 (thermal, sliding, pressing, fusion)
ReduceLR patience	4 (thermal, sliding, pressing, fusion)
ReduceLR min lr	10^{-6} (thermal, sliding, pressing, fusion)

F

Full Interpretability Results

This appendix provides the full interpretability result tables used in Chapter 4. All attribution percentages in this appendix are reported using the primary interpretability setting: TSR-regularized Integrated Gradients with per-element normalization and per-sample mean aggregation (see Chapter 3).

F.1. Overall modality attribution (outer ring)

F.2. Per-class modality attribution (outer ring)

F.2.1. Model 3

F.2.2. Composite (1+2)

F.3. Within-modality attribution: phases and sliding bins (inner ring)

F.3.1. Thermal phases

F.3.2. Pressing phases

F.3.3. Sliding top-3 bins

F.4. Within-modality attribution: channel breakdown (inner ring)

Table F.1: Overall modality attribution (outer ring) under the primary interpretability setting. Values are percentage shares (sum to 100% per row).

Model	Split	Thermal (%)	Pressing (%)	Sliding (%)
Model 1 (Hot–Cold)	Val	45.9	18.4	35.7
Model 1 (Hot–Cold)	Test	45.2	18.9	35.8
Model 1 (Hard–Soft)	Val	45.9	18.9	35.2
Model 1 (Hard–Soft)	Test	46.8	20.2	33.0
Model 1 (Rough–Smooth)	Val	48.6	18.1	33.4
Model 1 (Rough–Smooth)	Test	50.1	19.1	30.8
Model 1 (Sticky–Slippery)	Val	41.7	21.9	36.4
Model 1 (Sticky–Slippery)	Test	45.7	21.1	33.2
Composite (1+2)	Val	21.5	13.3	65.2
Composite (1+2)	Test	24.2	13.4	62.5
Model 3	Val	40.2	34.6	25.2
Model 3	Test	39.9	34.0	26.1

Table F.2: Model 3 modality attribution per class (outer ring). Values are percentage shares (sum to 100% per row).

Class	Val Thermal	Val Pressing	Val Sliding	Test Thermal	Test Pressing	Test Sliding
Fabric	29.4	54.6	16.1	26.1	60.6	13.3
Foam	38.4	18.3	43.3	43.1	19.0	37.9
Metal	86.4	9.8	3.8	91.6	8.4	0.0
Paper	26.9	49.5	23.6	29.2	29.9	40.8
Sandpaper	40.7	25.2	34.2	30.4	27.8	41.8
Vinyl	44.5	28.5	27.0	32.4	37.1	30.6
Wood	34.6	32.0	33.4	40.5	30.9	28.6

Table F.3: Composite Model (1+2) modality attribution per class (outer ring). Values are percentage shares (sum to 100% per row).

Class	Val Thermal	Val Pressing	Val Sliding	Test Thermal	Test Pressing	Test Sliding
Fabric	18.9	16.4	64.7	23.8	16.5	59.8
Foam	18.9	10.9	70.1	24.7	12.1	63.2
Metal	12.5	16.7	70.9	15.9	20.6	63.5
Paper	18.3	11.5	70.2	14.8	10.2	75.0
Sandpaper	39.9	13.8	46.3	25.7	3.3	70.9
Vinyl	25.0	7.1	67.8	31.2	8.1	60.7
Wood	26.0	12.5	61.5	37.9	14.0	48.1

Table F.4: Thermal phase attribution (inner ring) under the primary interpretability setting. Values are shares within thermal (sum to 100% per row).

Model	Split	Half-eq (%)	Peak (%)	Steady (%)	No-contact (%)
Model 1 (Hot–Cold)	Val	31.3	36.8	14.3	17.7
Model 1 (Hot–Cold)	Test	30.2	36.7	12.5	20.7
Model 1 (Hard–Soft)	Val	31.1	37.8	13.9	17.3
Model 1 (Hard–Soft)	Test	32.0	36.3	12.4	19.2
Model 1 (Rough–Smooth)	Val	30.2	36.5	14.2	19.1
Model 1 (Rough–Smooth)	Test	29.7	37.1	12.5	20.7
Model 1 (Sticky–Slippery)	Val	30.5	37.5	12.1	19.9
Model 1 (Sticky–Slippery)	Test	29.5	36.8	12.8	20.9
Composite (1+2)	Val	38.9	26.4	26.0	8.7
Composite (1+2)	Test	40.6	30.4	21.0	8.0
Model 3	Val	46.7	28.4	18.3	6.5
Model 3	Test	43.2	26.1	19.8	11.0

Table F.5: Pressing phase attribution (inner ring) under the primary interpretability setting. Values are shares within pressing (sum to 100% per row).

Model	Split	Loading (%)	Plateau (%)	Lift-off (%)
Model 1 (Hot–Cold)	Val	28.1	43.1	28.8
Model 1 (Hot–Cold)	Test	27.7	42.0	30.3
Model 1 (Hard–Soft)	Val	29.2	41.3	29.5
Model 1 (Hard–Soft)	Test	27.7	40.5	31.8
Model 1 (Rough–Smooth)	Val	27.9	44.9	27.1
Model 1 (Rough–Smooth)	Test	27.0	44.3	28.6
Model 1 (Sticky–Slippery)	Val	28.6	42.5	28.9
Model 1 (Sticky–Slippery)	Test	27.9	42.2	29.9
Composite (1+2)	Val	39.0	12.9	48.1
Composite (1+2)	Test	35.9	8.5	55.7
Model 3	Val	36.2	33.0	30.8
Model 3	Test	35.3	32.8	31.9

Table F.6: Top-3 sliding force–speed bins (inner ring within sliding). Values are shares within the sliding-bin mask distribution (top-3 only).

Model	Split	Top-3 bins (share)
Model 1 (Hot–Cold)	Val	F 0.2–0.4 N, v 99.0–132.0 mm/s (18.3%); F 0.2–0.4 N, v 33.0–66.0 mm/s (15.8%); F 0.2–0.4 N, v 66.0–99.0 mm/s (14.5%).
Model 1 (Hot–Cold)	Test	F 0.2–0.4 N, v 33.0–66.0 mm/s (29.9%); F 0.2–0.4 N, v 66.0–99.0 mm/s (20.7%); F 0.2–0.4 N, v 99.0–132.0 mm/s (10.2%).
Model 1 (Hard–Soft)	Val	F 0.2–0.4 N, v 99.0–132.0 mm/s (18.6%); F 0.2–0.4 N, v 33.0–66.0 mm/s (15.6%); F 0.2–0.4 N, v 66.0–99.0 mm/s (14.8%).
Model 1 (Hard–Soft)	Test	F 0.2–0.4 N, v 33.0–66.0 mm/s (29.4%); F 0.2–0.4 N, v 66.0–99.0 mm/s (20.7%); F 0.2–0.4 N, v 99.0–132.0 mm/s (10.3%).
Model 1 (Rough–Smooth)	Val	F 0.2–0.4 N, v 99.0–132.0 mm/s (18.1%); F 0.2–0.4 N, v 33.0–66.0 mm/s (15.9%); F 0.2–0.4 N, v 66.0–99.0 mm/s (14.6%).
Model 1 (Rough–Smooth)	Test	F 0.2–0.4 N, v 33.0–66.0 mm/s (29.8%); F 0.2–0.4 N, v 66.0–99.0 mm/s (20.3%); F 0.2–0.4 N, v 99.0–132.0 mm/s (10.3%).
Model 1 (Sticky–Slippery)	Val	F 0.2–0.4 N, v 99.0–132.0 mm/s (17.8%); F 0.2–0.4 N, v 33.0–66.0 mm/s (16.0%); F 0.2–0.4 N, v 66.0–99.0 mm/s (14.8%).
Model 1 (Sticky–Slippery)	Test	F 0.2–0.4 N, v 33.0–66.0 mm/s (29.7%); F 0.2–0.4 N, v 66.0–99.0 mm/s (20.5%); F 0.2–0.4 N, v 99.0–132.0 mm/s (10.2%).
Composite (1+2)	Val	F 0.2–0.4 N, v 33.0–66.0 mm/s (16.4%); F 0.2–0.4 N, v 66.0–99.0 mm/s (15.5%); F 0.2–0.4 N, v 99.0–132.0 mm/s (15.1%).
Composite (1+2)	Test	F 0.2–0.4 N, v 33.0–66.0 mm/s (26.8%); F 0.2–0.4 N, v 66.0–99.0 mm/s (21.2%); F 0.2–0.4 N, v 99.0–132.0 mm/s (14.1%).
Model 3	Val	F 0.2–0.4 N, v 99.0–132.0 mm/s (21.9%); F 0.2–0.4 N, v 33.0–66.0 mm/s (18.1%); F 0.2–0.4 N, v 66.0–99.0 mm/s (14.7%).
Model 3	Test	F 0.2–0.4 N, v 33.0–66.0 mm/s (28.7%); F 0.2–0.4 N, v 66.0–99.0 mm/s (20.0%); F 0.2–0.4 N, v 99.0–132.0 mm/s (12.1%).

Table F.7: Channel-level attribution shares within each modality (inner ring: channels). Values are shares within a modality (sum to 100% within thermal / pressing / sliding).

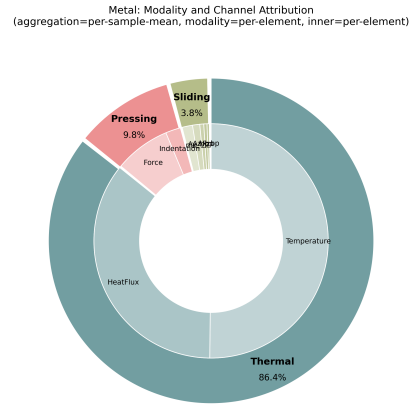
Model / Split	Modality	Channel shares (%)
Composite (1+2) / Val	Thermal	HeatFlux 53.7, Temperature 46.3
Composite (1+2) / Val	Pressing	feature_1 73.4, feature_0 26.6
Composite (1+2) / Val	Sliding	μ 23.9, Ay_bp 22.5, Ax_bp 20.7, Az_bp 20.0, Fz 12.9
Composite (1+2) / Test	Thermal	HeatFlux 52.7, Temperature 47.3
Composite (1+2) / Test	Pressing	feature_1 76.6, feature_0 23.4
Composite (1+2) / Test	Sliding	μ 23.0, Ay_bp 22.7, Az_bp 20.8, Ax_bp 20.3, Fz 13.2
Model 3 / Val	Thermal	HeatFlux 66.9, Temperature 33.1
Model 3 / Val	Pressing	feature_1 73.3, feature_0 26.7
Model 3 / Val	Sliding	μ 31.0, Az_bp 20.4, Ay_bp 20.2, Ax_bp 17.4, Fz 10.9
Model 3 / Test	Thermal	HeatFlux 64.6, Temperature 35.4
Model 3 / Test	Pressing	feature_1 74.1, feature_0 25.9
Model 3 / Test	Sliding	μ 32.5, Ay_bp 21.8, Az_bp 20.2, Ax_bp 16.5, Fz 9.1

G

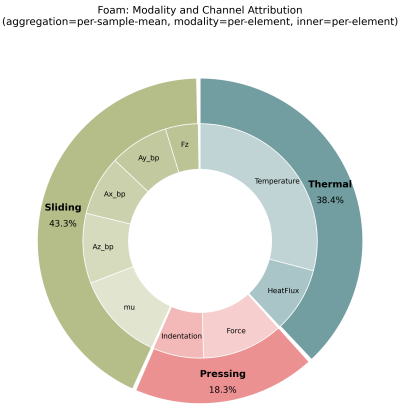
Full Interpretability Figure Sets

This appendix provides the complete set of per-class nested donut charts that complement the representative examples shown in Chapter 4. For each model and split, all seven retained material classes are shown. Each figure summarizes attribution shares at the channel level (outer ring: modality; inner ring: within-modality structure), using the `per_sample_mean` aggregation and `per_element` normalization.

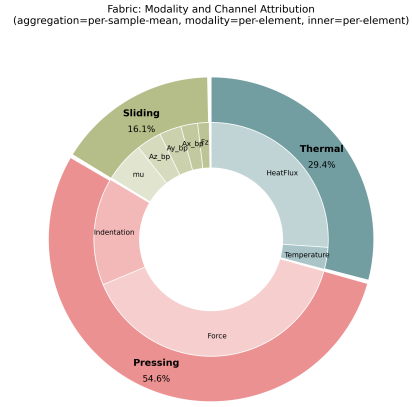
G.1. Model 3: Validation set (all classes)



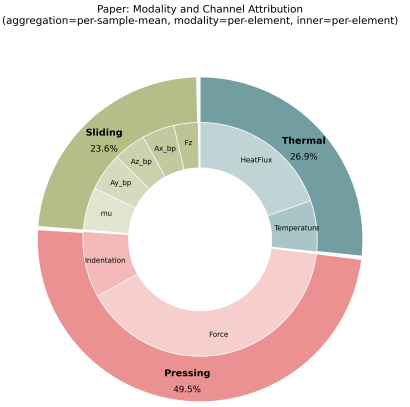
(a) Metal



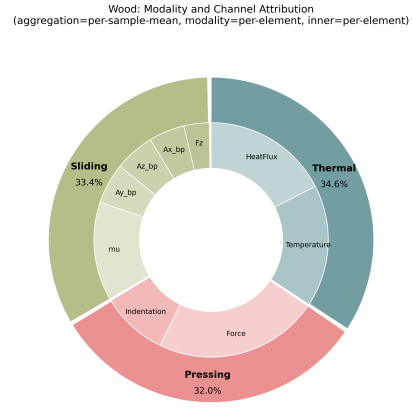
(b) Foam



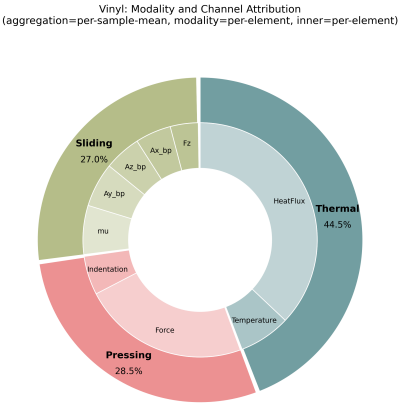
(c) Fabric



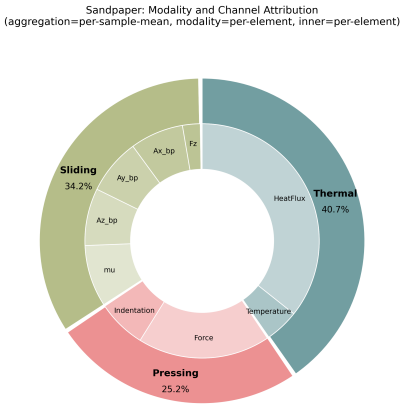
(d) Paper



(e) Wood



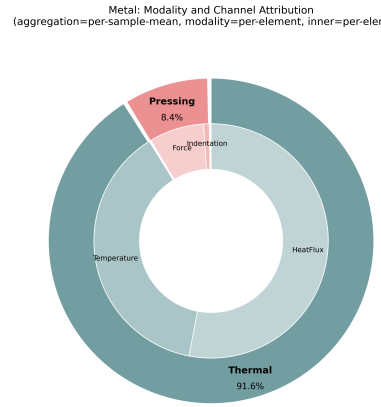
(f) Vinyl



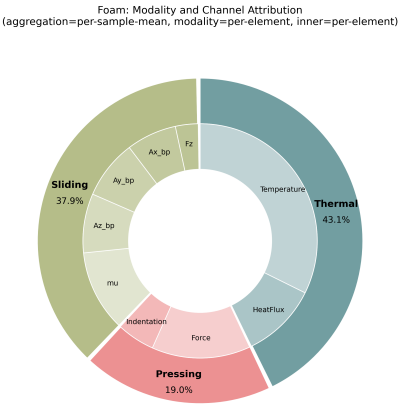
(g) Sandpaper

Figure G.1: Model 3 validation set: nested donut charts for all retained material classes.

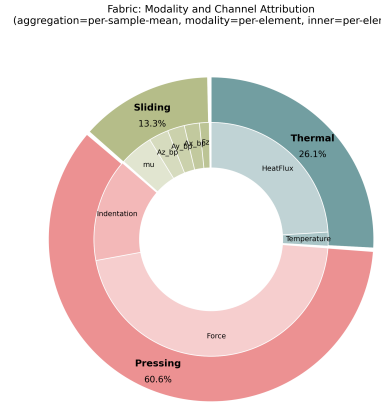
G.2. Model 3: Test set (all classes)



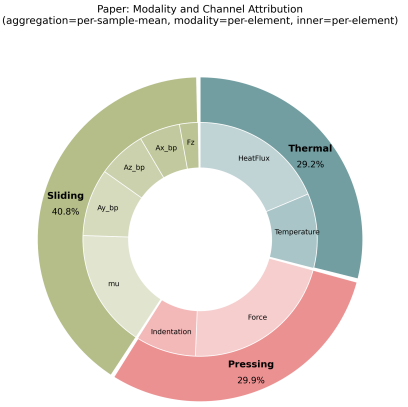
(a) Metal



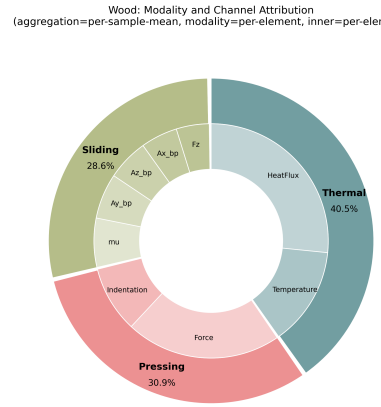
(b) Foam



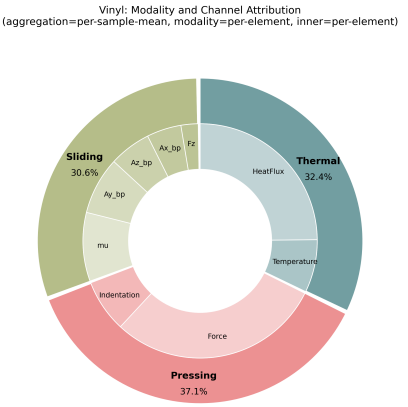
(c) Fabric



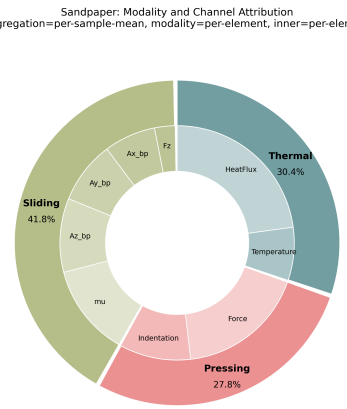
(d) Paper



(e) Wood



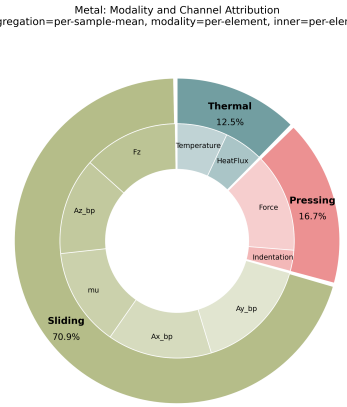
(f) Vinyl



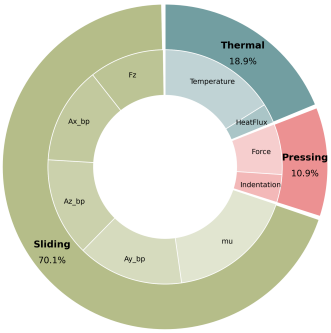
(g) Sandpaper

Figure G.2: Model 3 test set: nested donut charts for all retained material classes.

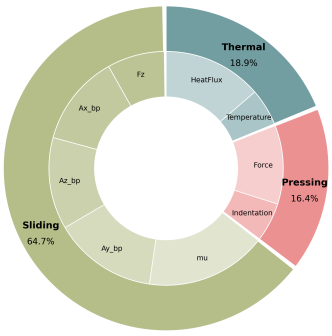
G.3. Composite Model (1+2): Validation set (all classes)



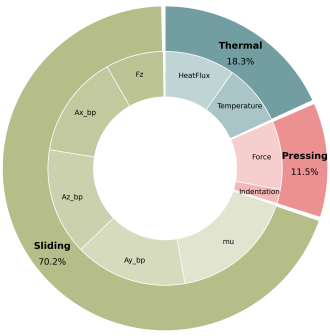
(a) Metal



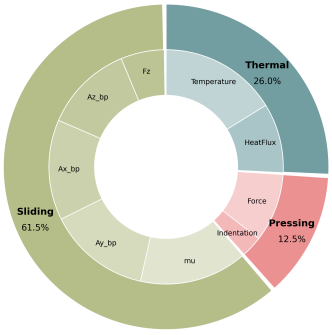
(b) Foam



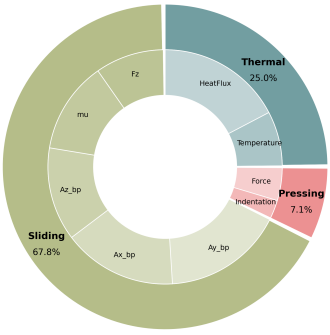
(c) Fabric



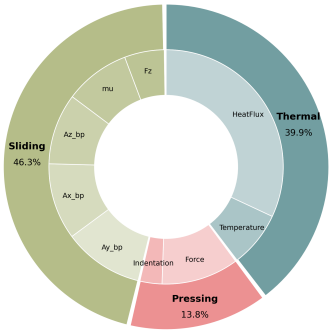
(d) Paper



(e) Wood



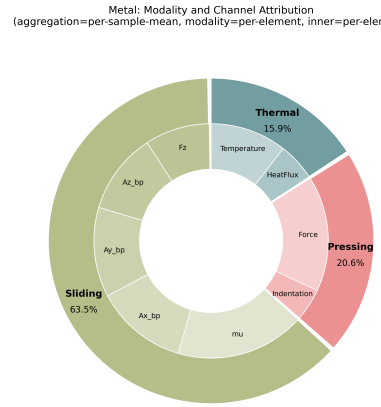
(f) Vinyl



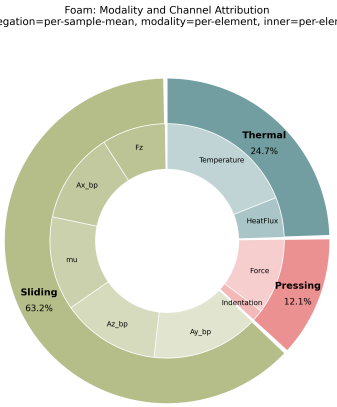
(g) Sandpaper

Figure G.3: Composite model (1+2) validation set: nested donut charts for all retained material classes.

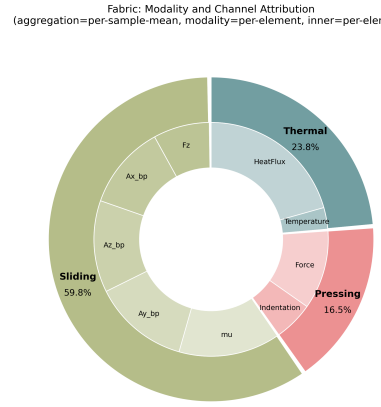
G.4. Composite Model (1+2): Test set (all classes)



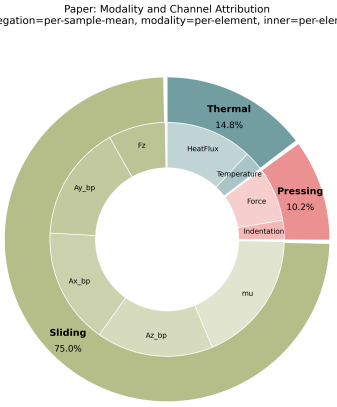
(a) Metal



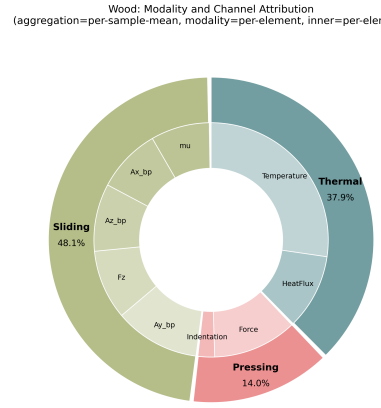
(b) Foam



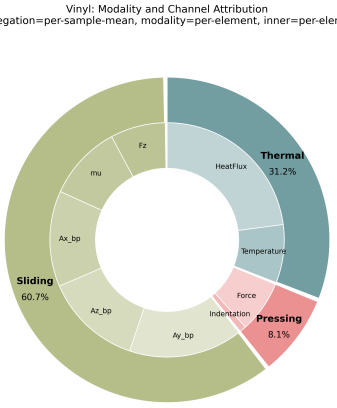
(c) Fabric



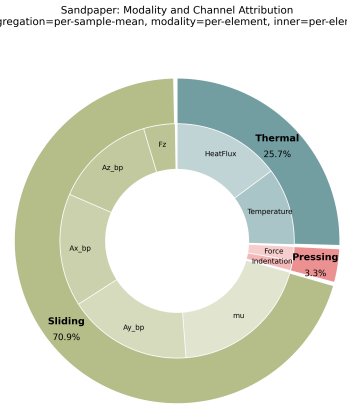
(d) Paper



(e) Wood



(f) Vinyl



(g) Sandpaper

Figure G.4: Composite model (1+2) test set: nested donut charts for all retained material classes.

H

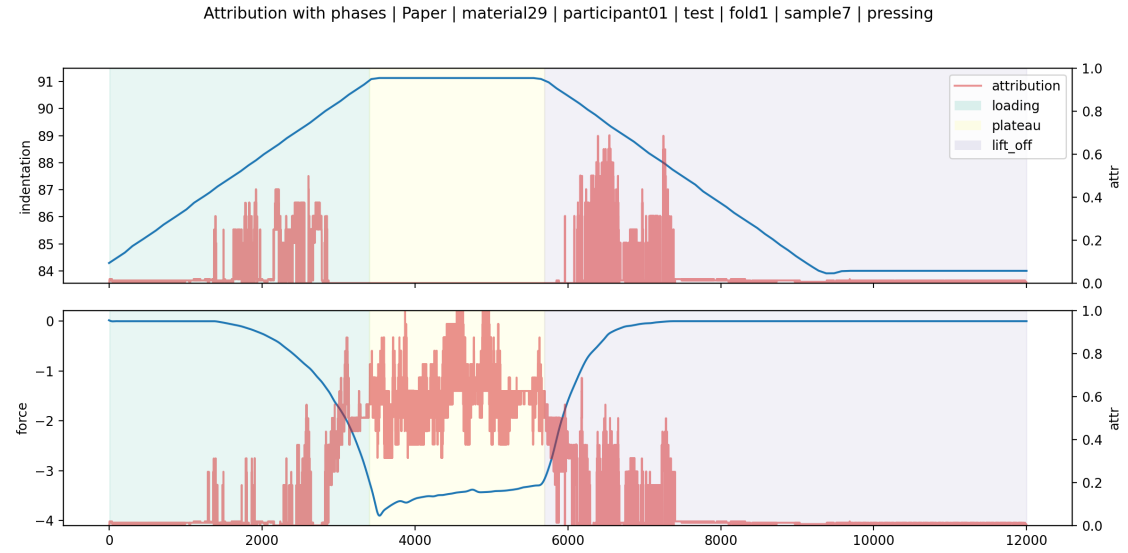
Attribution Sensitivity Analyses

H.1. Sample-level attribution maps (TSR phase overview)

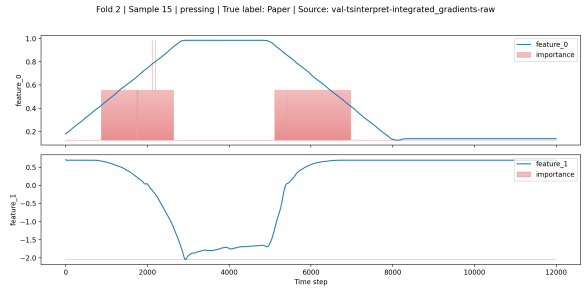
Figure H.1 shows a representative sample-level attribution visualization using Integrated Gradients with Temporal Saliency Rescaling (TSR). The plots illustrate how attribution mass aligns with interaction structure across pressing, sliding, and thermal modalities for the same sample.

H.2. TSR vs. raw Integrated Gradients (single-sample example)

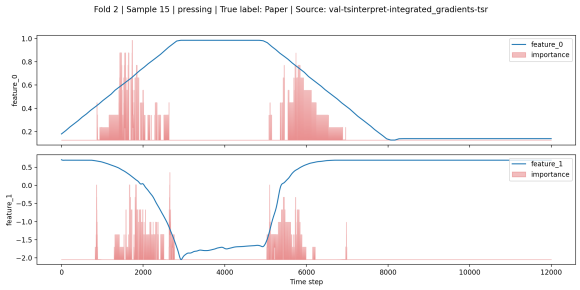
Figure H.2 shows an example attribution map for the same sample (Fold 2, Sample 15; true label: Paper), comparing raw Integrated Gradients to Integrated Gradients with Temporal Saliency Rescaling (TSR) for each modality.



(a) Pressing: TSR phase overview

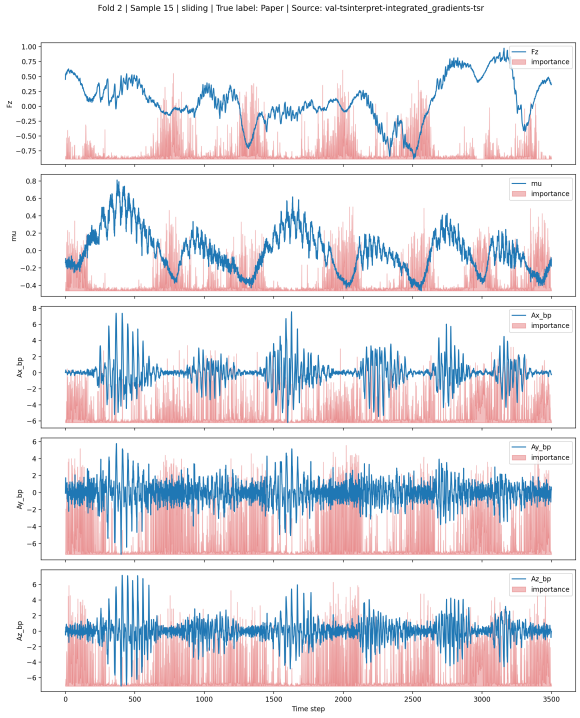


(a) Pressing: IG (raw)

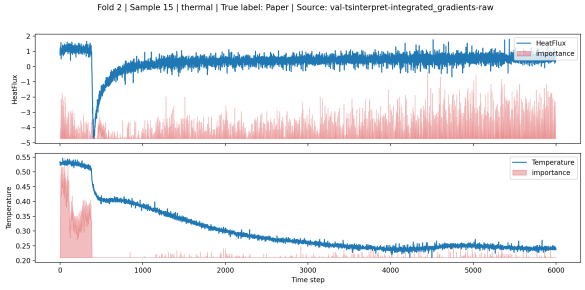


(b) Pressing: IG + TSR

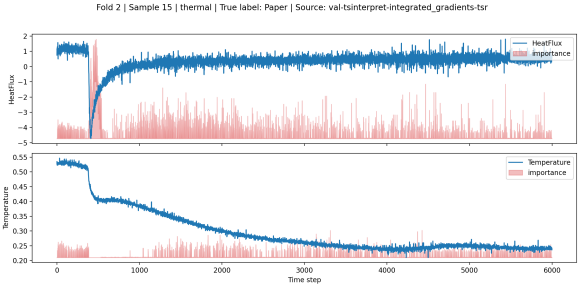
(c) Sliding: IG (raw)



(d) Sliding: IG + TSR



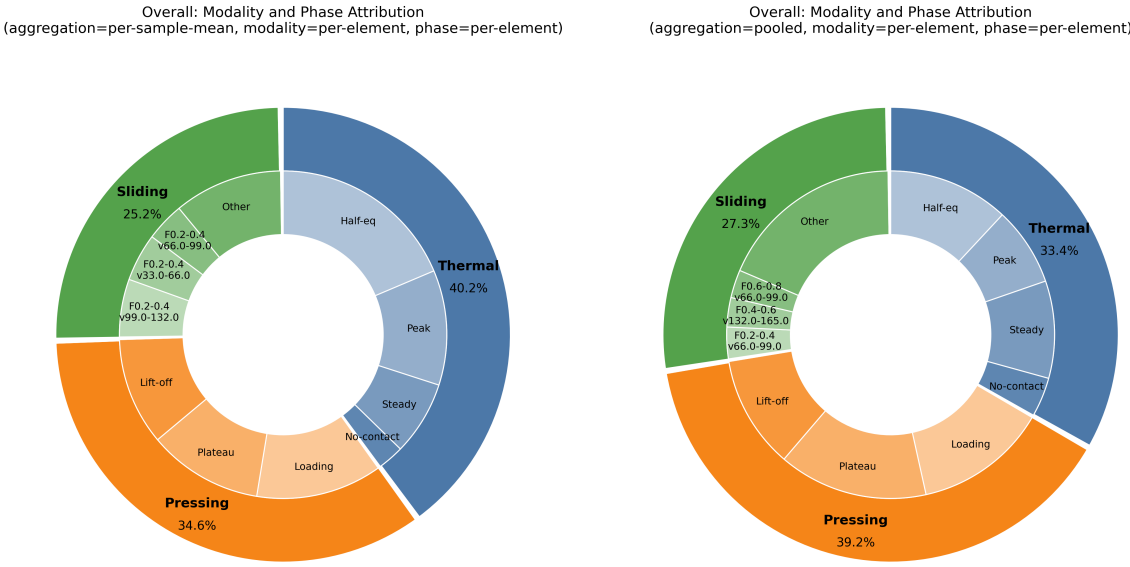
(e) Thermal: IG (raw)



(f) Thermal: IG + TSR

Figure H.2: Raw Integrated Gradients versus Integrated Gradients with Temporal Saliency Rescaling (TSR) for a single sample (Fold 2, Sample 15; true label: Paper). Rows correspond to modalities (pressing, sliding, thermal).

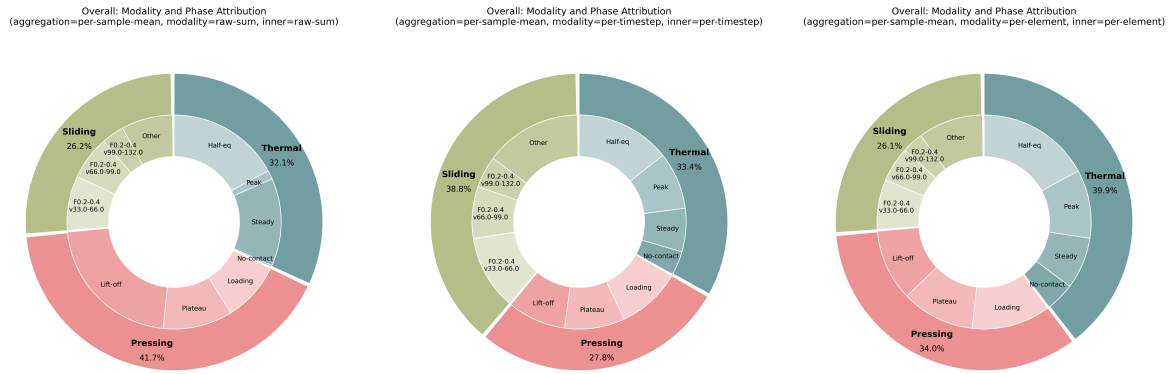
H.3. Effect of aggregation strategy (test set): per-sample mean vs. pooled



(a) Per-sample mean aggregation (b) Pooled aggregation

Figure H.3: Overall test-set modality and phase attribution under two aggregation strategies (per-sample mean vs. pooled), holding normalization fixed to per-element.

H.4. Effect of normalization strategy (per-sample mean): raw sum vs. per-timestep vs. per-element



(a) Raw sum (channel view) (b) Per-timestep (channel view) (c) Per-element (phase view)

Figure H.4: Overall attribution summaries under three normalization strategies (raw sum, per-timestep, per-element), using per-sample-mean aggregation. The first two panels show the channel breakdown; the per-element example is shown using the corresponding phase summary.

Representative raw pressing signal overlays

This appendix provides qualitative overlays of the raw pressing signals to better understand the relative attribution assigned to normal force and indentation in the main results. Under the pressing protocol, the target force was fixed at 3 N, so the maximum normal force should be broadly comparable across materials, whereas the indentation reached at that force may differ depending on material response. Because the main interpretability results nevertheless showed stronger attribution to normal force than to indentation, the raw pressing traces were inspected directly.

Figures I.1–I.9 show representative overlays at three levels: one representative sample per retained class, within-class overlays for each retained material for participant 7, and one cross-participant comparison for a single material. Overall, the indentation trajectories show clear material-dependent differences, but are relatively smooth and can vary in absolute level across trials and participants. The force traces, in contrast, retain richer temporal variation during loading, holding, and release, including differences in slope, recovery timing, and residual fluctuations. These observations support the interpretation that normal force provided a more temporally detailed and consistently exploitable cue for the model in the present dataset.

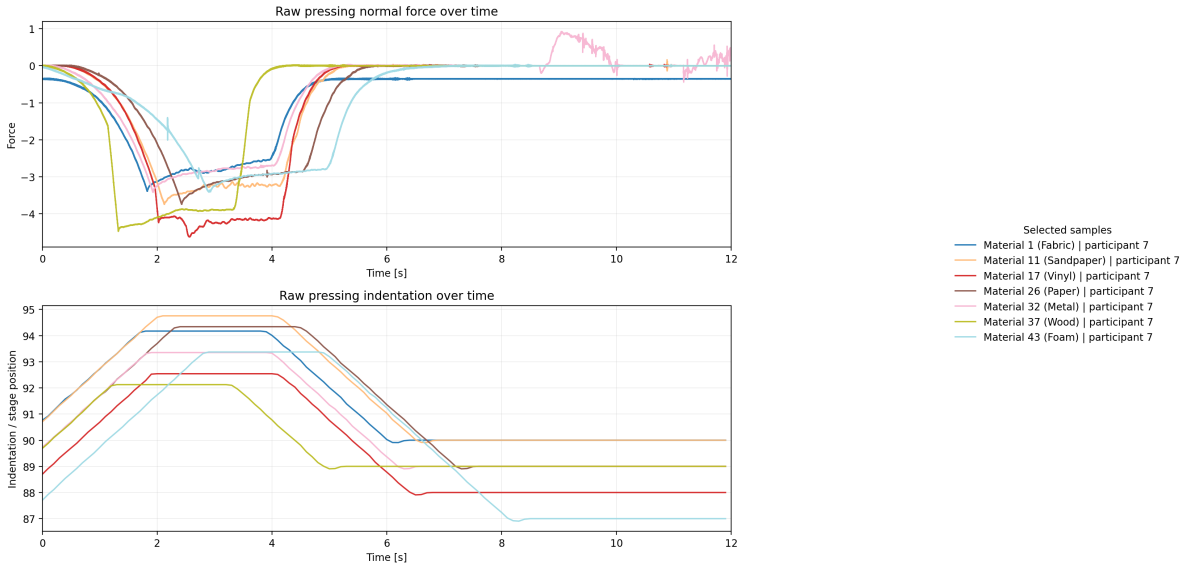


Figure I.1: Representative raw pressing overlays for one selected sample per retained material class for participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time. The figure highlights that indentation differs across classes, while the force traces retain more local temporal variation during loading, holding, and release.

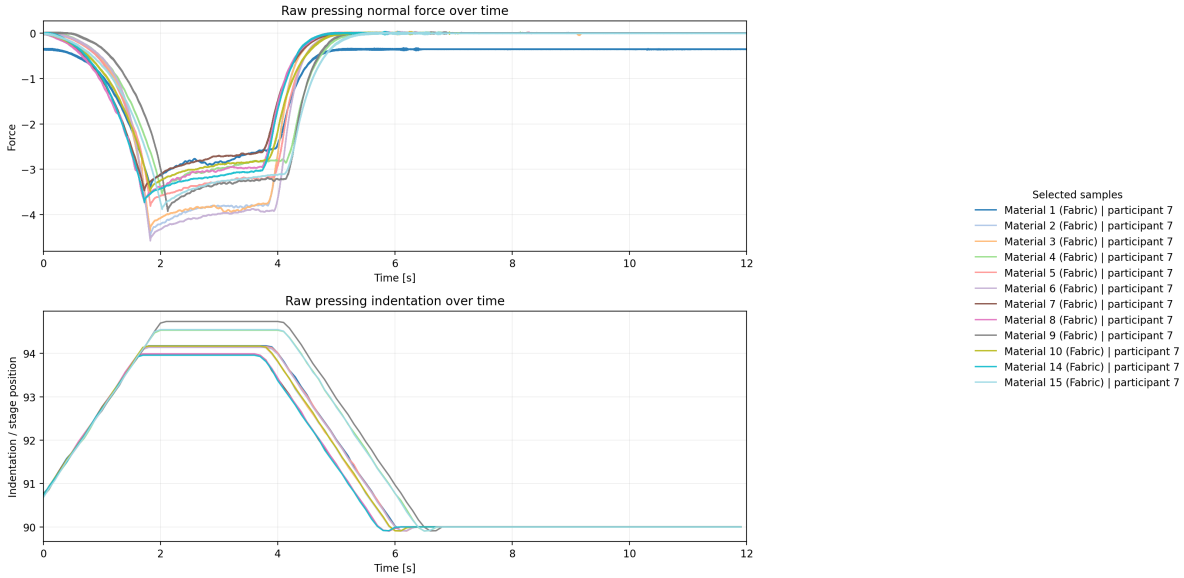


Figure I.2: Raw pressing overlays for Fabric samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

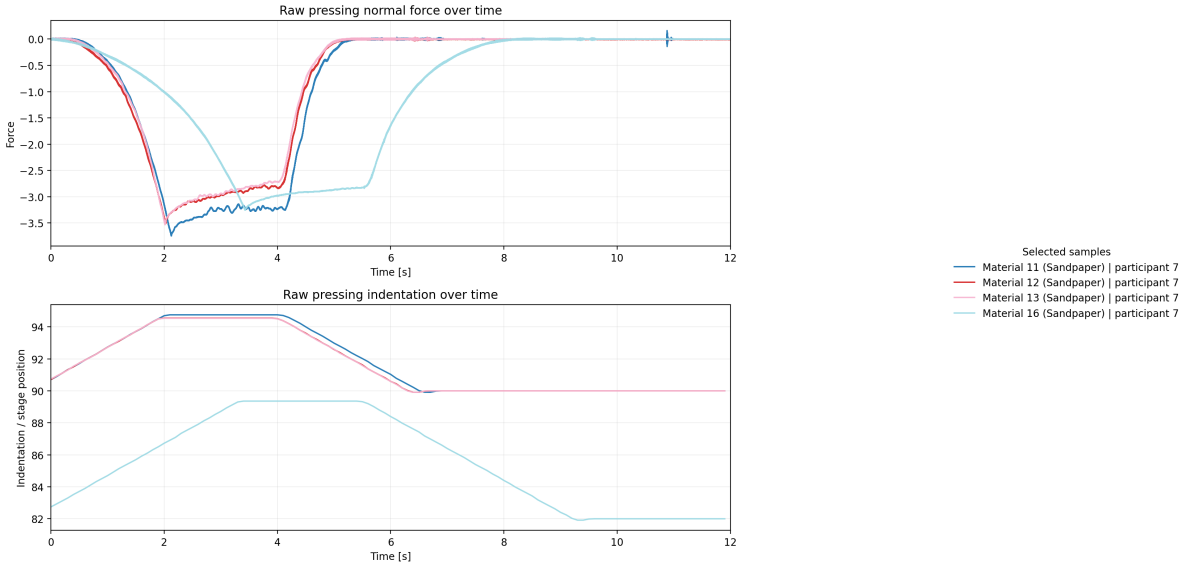


Figure I.3: Raw pressing overlays for Sandpaper samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

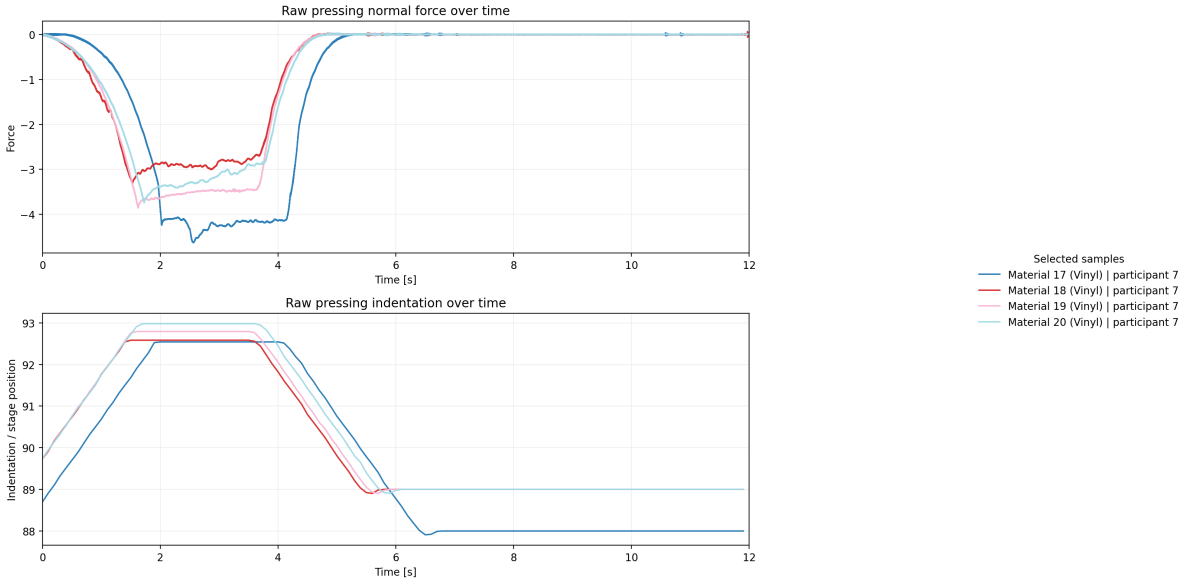


Figure I.4: Raw pressing overlays for Vinyl samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

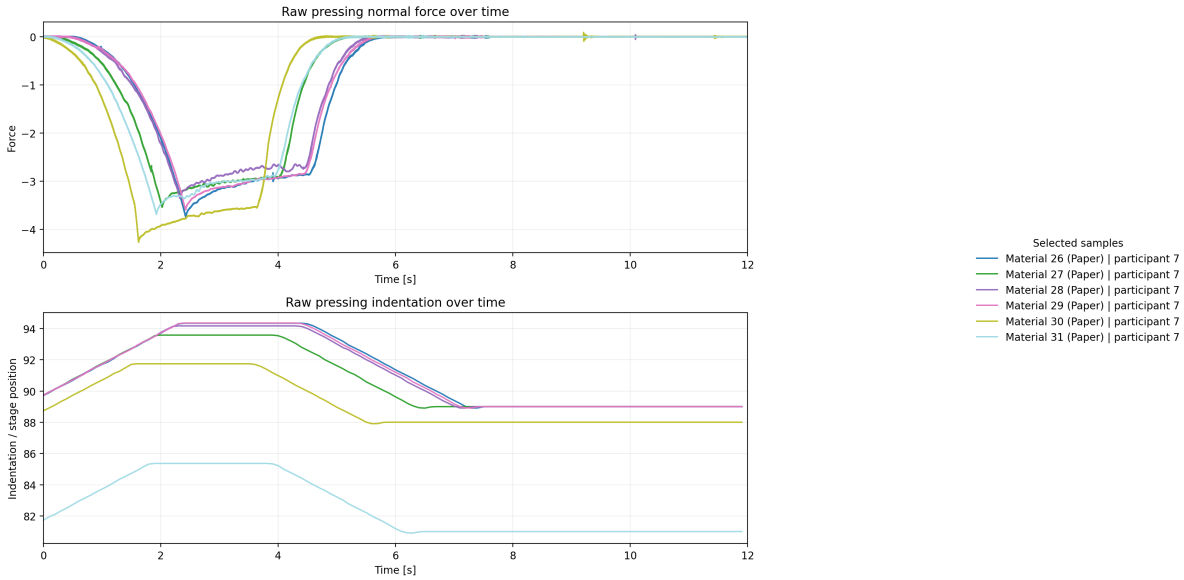


Figure I.5: Raw pressing overlays for Paper samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

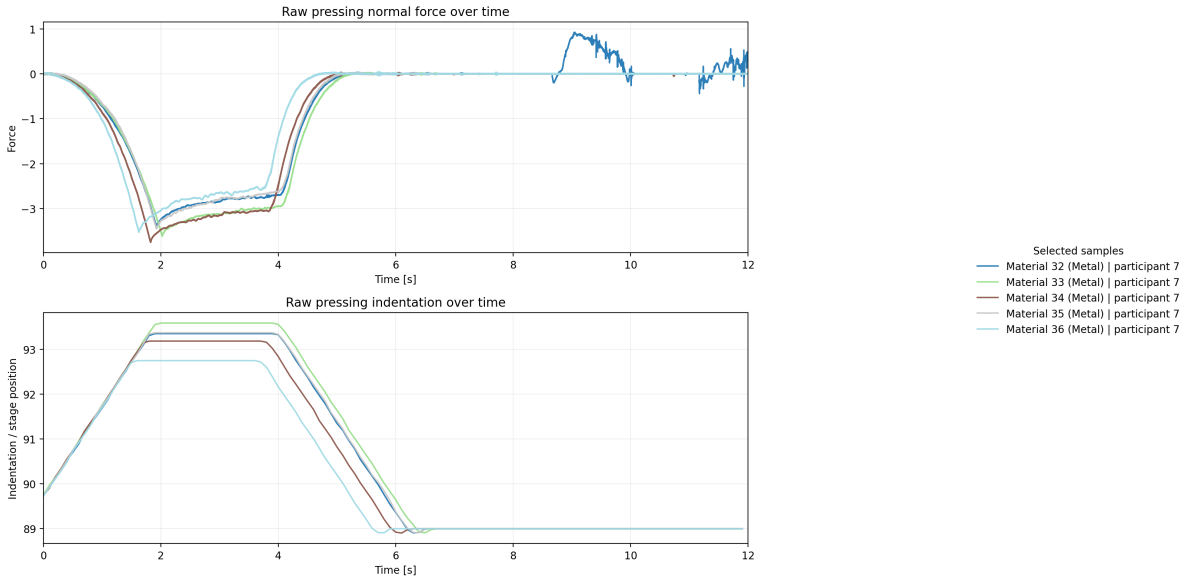


Figure I.6: Raw pressing overlays for Metal samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

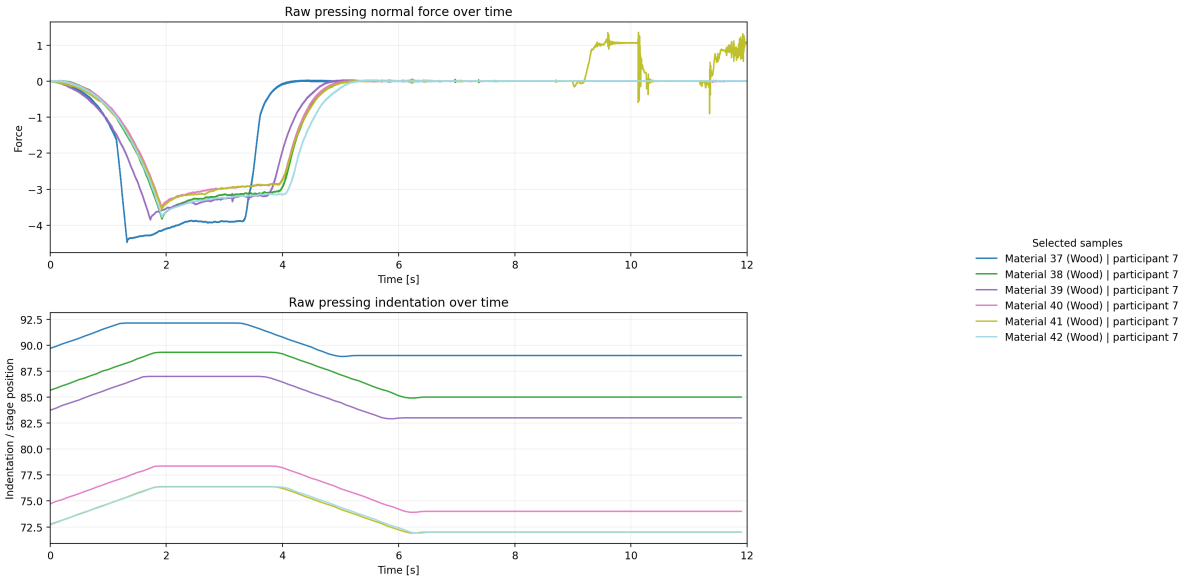


Figure I.7: Raw pressing overlays for Wood samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

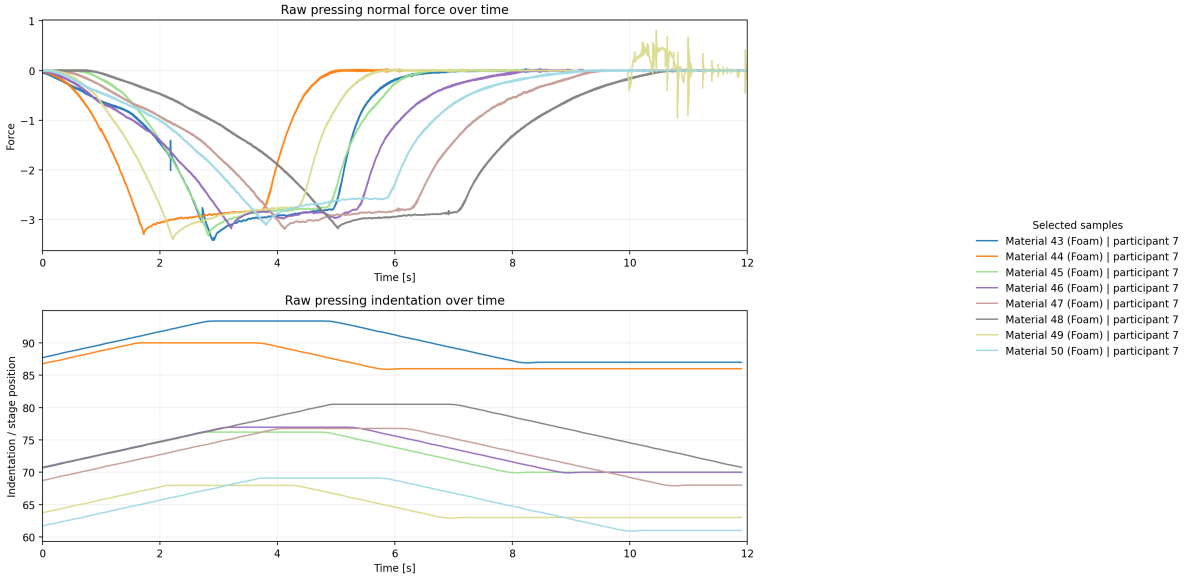


Figure I.8: Raw pressing overlays for Foam samples from participant 7. The top panel shows normal force over time and the bottom panel shows indentation over time.

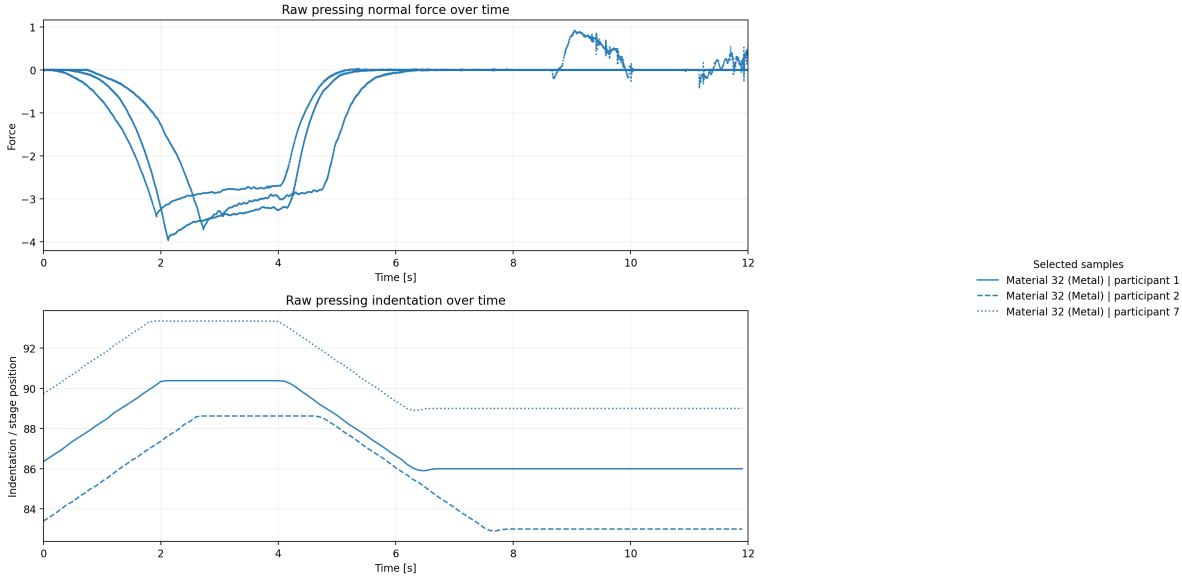


Figure I.9: Cross-participant raw pressing overlay for Material 32 (Metal). The top panel shows normal force over time and the bottom panel shows indentation over time for different participants. The figure illustrates that indentation can shift in absolute level across participants, even for the same material.