Acoustic-Based Contact Detection and Geometric Reconstruction for Robotic Manipulation

Georg Wolnik
Robotics and Biology Laboratory
Technische Universität Berlin
Berlin, Germany
wolnik@campus.tu-berlin.de

***Abstract*— This work investigates acoustic-based contact detection and geometric reconstruction using an acoustic tactile finger attached to a Franka Panda robot. We develop a pipeline consisting of systematic data collection using the robot, feature engineering, classification, and geometric reconstruction. We demonstrate a proof-of-concept showing that it is possible to distinguish between different contact scenarios (contact, no contact, edges) for known data distributions and leverage the robot's state to map predictions onto a 2D coordinate system to reconstruct square patterns. Furthermore, we compare the performance between 3-class and binary classification, compare using spectrograms versus hand-crafted features as input, evaluate the generalization performance on seen objects in a new workspace and on an unseen object in a new workspace and present challenges of coupling the acoustic tactile sensor with the robot.**

I. INTRODUCTION

Contact detection is fundamental for robotic manipulation. Force sensors provide little geometric detail about contact, while visual-tactile sensors like GelSight can capture fine geometry but are complex and expensive to deploy. Vision systems struggle with occlusions and lighting conditions. Acoustic sensing offers a middle ground: a single acoustic finger with integrated speaker and microphone mounted on the robot captures rich spectral and temporal information about contact events, at low hardware cost.

Despite these advantages, using an acoustic finger with a robot arm remains largely unexplored. Prior work focused on soft pneumatic actuators [1, 2, 3], where compliant air-filled chambers create favorable acoustic properties. Combining a robot arm with an acoustic finger presents new challenges, particularly *configuration entanglement* [4]—where the robot's joint configuration affects acoustic signals independently of contact. Furthermore, prior work focused exclusively on binary contact detection, leaving open whether acoustic signals can support 2D spatial mapping with explicit edge detection, which is necessary for understanding object geometry through touch.

We investigate acoustic sensing for contact detection and geometric reconstruction using a custom acoustic tactile finger on a Franka Panda robot. The robot moves in a raster pattern across 3D-printed objects with different cutout patterns, recording acoustic signals to train machine learning models for contact classification. We then use predictions and robot state to reconstruct surface geometry, comparing different classifiers and input representations (hand-crafted features vs spectrograms) while evaluating generalization across workspaces and objects.

A. Research Questions

We investigate four questions:

RQ1: Proof of Concept. Can acoustic sensing achieve above-random accuracy for contact scenario classification on known data distribution and enable geometric reconstruction?

RQ2: Position Generalization. Can models trained at specific robot configurations generalize to new positions with the same objects?

RQ3: 3-Class vs Binary. Does including edge samples as a separate class improve performance over binary classification?

RQ4: Object Generalization. Can models trained on specific objects generalize to a new object in a new workspace?

B. Contributions

Our key contributions are:

- **3-class acoustic contact scenario classification**, achieving 70% cross-validation accuracy on known data distributions, enabling geometric reconstruction of object surfaces with square cutout patterns.
- **Generalization analysis** showing poor position generalization and object generalization, demonstrating that acoustic signatures are workspace and object specific.
- **Physics-based explanation** using eigenfrequency analysis to explain why position and object generalization fail.
- **Ablation studies** comparing 3-class vs binary classification performance, hand-crafted features vs spectrograms performance, and single vs multi-sample recordings performance.

II. RELATED WORK

A. Acoustic Sensing for Soft Robotics

Wall [1] pioneered acoustic sensing for morphological computation in soft pneumatic actuators, showing that passive acoustic signals encode both contact information and actuator state. Wall et al. [2] extended this by combining

passive monitoring with active acoustic excitation, demonstrating that actively induced signals significantly improve signal-to-noise ratio and enable contact detection as well as material classification through frequency-domain analysis.

Zöller et al. [3] developed active acoustic contact sensing for soft grippers using chirp-based excitation, showing that acoustic signatures reliably distinguish contact states even in noisy environments. However, all of these approaches rely on *soft pneumatic actuators*, where air-filled compliant chambers create favorable acoustic coupling properties. The application to *robot arms* remained unexplored, presenting new challenges: robot arm structures propagate vibrations very differently, and there is no air-based coupling to amplify contact signals.

B. Robot Configuration Entanglement

A key challenge specific to robot arms is *configuration entanglement* - the robot's joint configuration affects measured acoustic signals independently of contact, making it difficult to isolate contact information. Zhang et al. [4] systematically studied this with VibeCheck, showing that robot arm configurations create mechanical coupling that entangles joint state with contact signals. Naive training approaches fail when configurations change between training and deployment, achieving only random-chance performance on out-of-distribution configurations. They proposed configuration-aware feature engineering and multi-configuration training as partial mitigations, though their focus was vibration sensing for slip detection rather than acoustic sensing for geometric reconstruction.

Our work confirms that configuration entanglement severely affects acoustic signals: cross-workspace validation averages only 34.5%, barely above random, with two rotations performing worse than chance. This directly extends VibeCheck's findings from slip detection to 3-class contact state mapping.

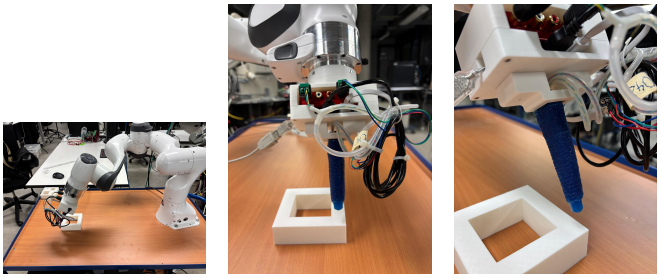
C. Research Gaps

Prior acoustic sensing work leaves three gaps this paper addresses. First, all existing work focuses on *binary contact detection* (contact vs. no-contact), never investigating explicit *edge/boundary detection* - a critical capability for understanding object geometry. Second, no systematic comparison exists between position generalization (same objects, new workspace) and object generalization (novel geometry), leaving open whether acoustic features capture workspace-invariant or object-invariant contact information. Third, there is no physics-based explanation of why generalization succeeds or fails, providing no actionable design principles.

This work introduces 3-class acoustic sensing with explicit edge modeling for robot arms, conducts systematic generalization experiments across workspace rotations and novel object holdout testing, and develops an eigenfrequency framework to explain the observed generalization failures.

TABLE I: Single vs Multi-Sample Recording Protocol Validation

Protocol	CV Acc	Std Dev	vs Random
Single (1 sample, 0ms)	65.8%	±4.0%	0.32×
Multi (5-10, 150ms)	**71.6%**	**±1.9%**	**0.43×**
Improvement	**+5.8%**	**2.1× lower**	**+11%**



(a) Overall Setup (b) Mounted finger (c) Finger closeup

Fig. 1: Experimental setup with Franka Panda and acoustic sensing end effector.

III. METHOD

A. Experimental Setup

Our setup consists of a Franka Panda robot with an acoustic tactile finger (integrated speaker and microphone) mounted on the end effector. The robot moves in a raster pattern across 3D-printed objects with different cutout patterns, recording acoustic data at each position. We label each sample with its contact state (contact, no-contact, edge) based on object geometry and robot position. The acoustic signal sweeps from 20 Hz to 20 kHz, exciting the full frequency spectrum. We record 5-10 samples per position to reduce motion artifacts.

The acoustic finger has approximately 1 cm × 0.25 cm contact area with 1 cm raster step size. We collect data from four objects across four workspaces (Table II)—different positions of objects relative to the robot base.

To reduce the effect of robot vibrations and motion artifacts on the acoustic signals, we record multiple samples (5-10) at each position. We hypothesize that robot motion artifacts introduce acoustic noise that can be more dominant than contact signals. For validation, we trained a Random Forest classifier (3-class: contact, no-contact, edge) using 5-fold cross-validation, comparing multi-sample and single-sample protocols. Multi-sample recordings achieved 71.6% accuracy (±1.9% std dev), while single-sample achieved 65.8% (±4.0% std dev). The higher accuracy and lower variance of multi-sample recordings indicate more stable predictions and support our hypothesis about motion artifacts.

B. Feature Engineering

We extract an 80-dimensional feature vector from each recording, including spectral features (centroid, bandwidth, rolloff), temporal features (zero-crossing rate, RMS energy), statistical features (mean, variance, skewness), and Mel-

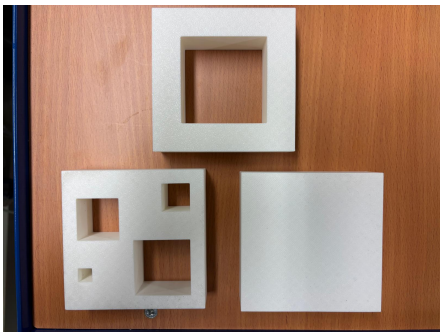


Fig. 2: The four 3D-printed test objects: A (multiple square cutouts), B (empty), C (full surface), D (large square cutout, hold-out).

TABLE II: Test Objects and Workspace Configuration

Object	Type	Workspaces
A	Multiple square cutouts	WS1, WS2, WS3
B	Empty (no object)	WS1, WS2, WS3
C	Full (no cutouts)	WS1, WS2, WS3
D	Large square cutout (hold-out)	WS4 only

frequency cepstral coefficients. We compare these hand-crafted features against spectrograms as classifier input.

C. Classification and Evaluation

We use Random Forest with 100 trees as our primary classifier after comparing five methods (Random Forest, K-NN, MLP, GPU-MLP, Ensemble). Testing all five on both feature representations, hand-crafted features outperform spectrograms across all classifiers with improvements ranging from 5-12% (Table III). This shows hand-crafted features capture more discriminative information, likely because spectrograms overfit to workspace-specific noise. We train using 5-fold stratified cross-validation to get robust performance estimates.

To test position generalization, we design three workspace rotation experiments: (1) Train on WS1+WS3, validate on WS2; (2) Train on WS2+WS3, validate on WS1; (3) Train on WS1+WS2, validate on WS3. Each rotation uses perfectly balanced 3-class labels (contact, no-contact, edge). Cross-validation accuracy measures performance on the training data, while validation accuracy measures generalization to the held-out workspace. For object generalization, we train on all data from WS1-3 and validate on WS4 with the unseen Object D, repeating the experiment across 5 random seeds to ensure reproducibility. Figure 3 shows the complete experimental design for all three rotations.

Dataset Balancing. Initial data collection, although balanced per object, produced imbalanced datasets when combining data from different objects and workspaces - no-contact samples outnumbered contact and edge samples by 2-3×. This was discovered late in the project. All experiments and reconstructions reported here were rerun after ensuring perfect class balance through stratified subsampling, making results different from those in the presentation.

TABLE III: Hand-Crafted Features vs. Spectrograms Comparison (Rotation 1: Train WS1+WS3, Validate WS2)

Classifier	Features	Spectrograms	Advantage
Random Forest	33.9%	22.9%	+11.0%
K-NN	32.7%	27.7%	+5.0%
MLP (Medium)	31.0%	23.8%	+7.2%
GPU-MLP (Medium)	33.9%	22.0%	+12.0%
Ensemble (Top3-MLP)	30.8%	22.3%	+8.4%
Win Count	**5/5**	**0/5**	—

IV. EXPERIMENTAL RESULTS

A. RQ1: Proof of Concept

We evaluate 3-class classification (contact, no-contact, edge) using 5-fold cross-validation on the balanced datasets from Workspaces 1-3. Random Forest achieves 69.9% cross-validation accuracy, which is significantly above the 33.3% random baseline ($p<0.001$). This demonstrates that acoustic sensing can successfully distinguish between different contact scenarios within known workspace configurations.

To demonstrate geometric reconstruction, we map the predictions onto 2D spatial coordinates based on the robot's position during data collection. We train a model on 80% of combined data from all workspaces (WS1+WS2+WS3) and test on the remaining 20%. The model achieves approximately 93% average accuracy across all three objects. Figure 4 shows the reconstructions for Object A (cutout patterns, 89.81% accuracy), Object B (empty workspace, 99.82% accuracy), and Object C (full contact surface, 90.17% accuracy). The reconstructions successfully reproduce the ground truth contact patterns, demonstrating that acoustic sensing can create spatial surface maps distinguishing contact, no-contact, and edge states for within-workspace scenarios.

B. RQ2: Position Generalization

To test whether models can generalize to new robot configurations with the same objects, we perform three workspace rotation experiments:

Rotation 1 (Train WS1+WS3, Test WS2): 33.9% validation accuracy (1.02× over random)

Rotation 2 (Train WS2+WS3, Test WS1): 23.3% validation accuracy (0.70× over random)

Rotation 3 (Train WS1+WS2, Test WS3): 55.7% validation accuracy (1.67× over random)

The average validation accuracy of 34.5% barely exceeds the 33.3% random baseline (1.04× normalized), representing catastrophic failure. Two rotations even perform worse than random guessing. Even the best case (Rotation 3) still shows a large drop from 69.9% cross-validation. This shows that acoustic signatures are highly workspace-specific and require workspace-specific training.

C. RQ3: 3-Class vs Binary Classification

To test whether including edge samples as a separate class improves performance, we compare 3-class classification against binary classification (contact vs no-contact, excluding all edge samples). Table IV shows the normalized performance:

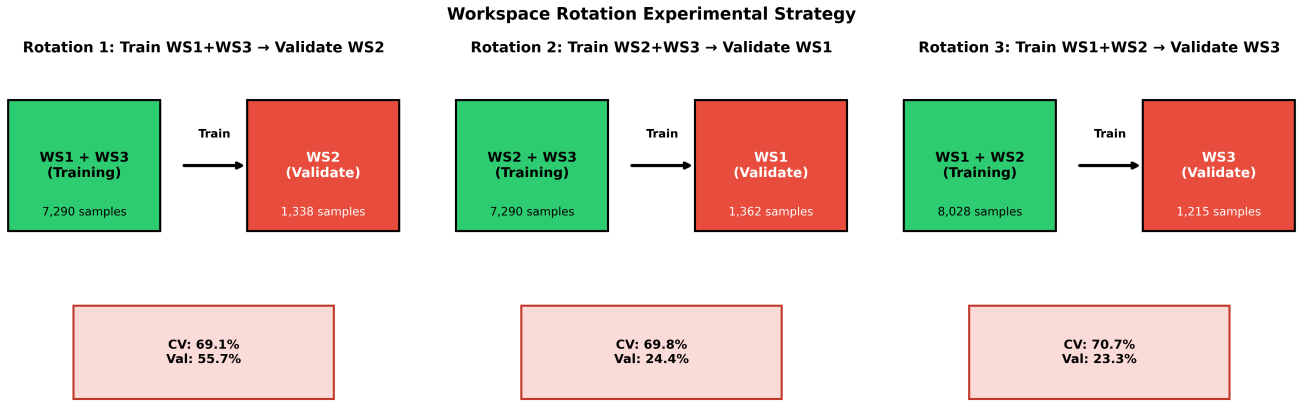


Fig. 3: Workspace rotation experimental strategy. Three rotations systematically evaluate position generalization: **Rotation 1**: Train WS1+WS3 (7,290 samples), validate WS2 (1,338 samples). **Rotation 2**: Train WS2+WS3 (7,290 samples), validate WS1 (1,362 samples). **Rotation 3**: Train WS1+WS2 (8,028 samples), validate WS3 (1,215 samples). Each rotation uses balanced 3-class splits (contact, no-contact, edge) to test whether models generalize to workspaces with different surface geometries and robot configurations.

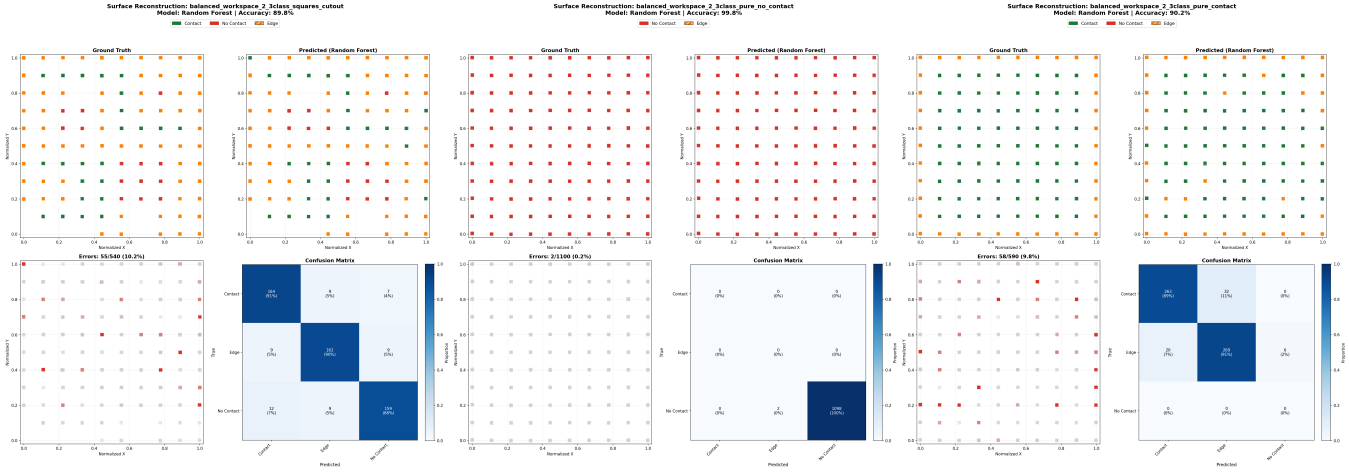


Fig. 4: Proof of concept: 3-class acoustic contact detection using 80/20 train/test split on combined workspace data (WS1+WS2+WS3). Left to right: Object A (cutout, 89.81%), Object B (empty workspace, 99.82%), Object C (full contact, 90.17%). Each panel shows ground truth and predictions with confusion matrix. Average accuracy 93.3%, well above the 33.3% random baseline. Color indicates contact state (contact=green, no-contact=red, edge=orange).

TABLE IV: 3-Class vs Binary Classification Comparison

Approach	Val Acc	Random	vs Random
Binary (exclude edge)	45.1%	50.0%	**0.90×**
3-Class (include edge)	**34.5%**	**33.3%**	**1.04×**

Although binary classification achieves higher raw accuracy (45.1%), it actually performs worse than random guessing when normalized by the baseline (0.90×). In contrast, 3-class achieves 1.04× over random. Additionally, binary classification fails in real deployment scenarios because it must misclassify all edge cases as either contact or no-contact, while 3-class explicitly handles boundary regions.

D. Generalization Comparison: Position vs Object

Table V directly compares the two generalization challenges. Position generalization (averaged across 3 rotations) achieves 34.5% validation, while object generalization (Random Forest without regularization) achieves 41.7%. Position generalization is measurably harder with a 7.2 percentage point gap.

This shows that acoustic signals contain object-specific patterns that partially transfer to new objects (41.7%), but workspace configurations produce fundamentally different signal distributions (34.5%, barely above random). Contact-class signals generalize somewhat across object geometries, but models trained in one workspace fail in another because the robot's position relative to the surface changes the acoustic signal even for identical contact.

TABLE V: Position vs Object Generalization Comparison

Challenge	Val Acc	Random	vs Random
Position (WS Rot Avg)	34.5%	33.3%	1.04×
Object (RF, no reg)	41.7%	33.3%	1.25×
Gap	**+7.2%**	—	**+0.21×**

TABLE VI: Object Generalization: Novel Object D Validation

Classifier	3-Class Val	Binary Val
Random Forest	41.7%	50.0%
GPU-MLP (No Reg)	35.7%	49.9%
GPU-MLP (HighReg)	**75.0%**	**50.0%**
K-NN	33.4%	49.9%
Ensemble	30.1%	50.0%
Random Baseline	33.3%	50.0%

E. RQ4: Object Generalization

We train on Objects A, B, C (Workspaces 1-3) and test on the unseen Object D (Workspace 4) across 5 independent random seeds. Table VI shows that performance varies drastically depending on the classifier:

Key findings: (1) Binary classification completely collapses to exactly 50% (pure random chance) across all classifiers and all 5 seeds. This shows that excluding edge samples leads to complete failure on new objects. (2) Most 3-class models barely exceed random performance (30-42%). (3) The heavily-regularized GPU-MLP with dropout (0.3) and weight decay (0.01) also achieves only random performance (33%) without confidence filtering - it only shows above-random accuracy on the small confident subset retained by filtering.

Confidence filtering analysis: We additionally tested whether the model's confidence scores correlate with prediction correctness by applying a confidence threshold (0.7) that only retains high-confidence predictions. This filters down to only 4 out of 2,280 spatial positions (0.2% coverage), on which the model happens to achieve 75% accuracy. However, given that only 4 positions are retained, this result is not statistically meaningful and does not constitute reliable object generalization. Without filtering, the same model achieves 33.03% (random chance). The key takeaway is that the model cannot reconstruct the full surface of new objects - it achieves random performance across all positions (Fig. 5).

F. Physics-Based Interpretation

We can explain the different generalization behaviors using acoustic eigenfrequency analysis. Physical objects have unique vibration patterns (eigenfrequencies) determined by their material and shape. When contact occurs, three different acoustic patterns emerge: (1) solid contact creates high-amplitude vibrations across many frequencies, (2) no-contact produces only low-amplitude background noise, (3) edge contact creates mixed signatures from partial surface contact.

Eigenfrequency framework. Objects vibrate at specific

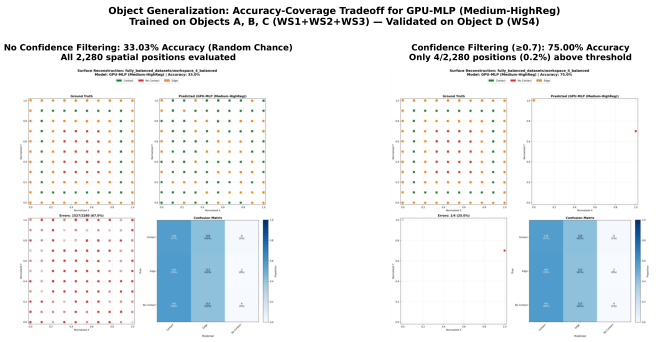


Fig. 5: Object generalization accuracy-coverage tradeoff on Object D. Left: Without confidence filtering, reconstruction achieves 33.03% (random). Right: With filtering (threshold 0.7), accuracy increases to 75% but only on 4/2,280 positions (0.2% coverage).

frequencies determined by:

$$f_n = \frac{1}{2\pi}\sqrt{\frac{k_n}{m_n}} \tag{1}$$

where f_n is the vibration frequency, k_n is the stiffness, and m_n is the mass. Contact creates object-specific acoustic signatures based on these frequencies. However, the *type of contact* (contact, no contact, edge) produces distinguishable patterns regardless of the specific frequencies: solid contact excites many frequency modes strongly, no-contact produces weak environmental noise, and edge contact creates intermediate mixed patterns.

Why regularization helps object generalization. Models without regularization memorize the specific frequency values of training objects A, B, C. When tested on Object D, which has a different cutout pattern, these memorized patterns don't transfer, achieving only 35.7% (random chance). We hypothesize that the different cutout geometry changes the effective mass distribution and structural stiffness of the object, shifting its eigenfrequencies - but this remains a hypothesis and would require dedicated measurement to confirm. Heavy regularization (dropout 0.3, weight decay 0.01) prevents this memorization by randomly dropping features during training and penalizing large weights. This forces the model to learn the *contact type patterns* - vibration amplitude, frequency distribution shape, temporal characteristics - that work across different object geometries. However, the 75% validation accuracy only applies to 0.2% of spatial positions (4 out of 2,280) where the model is confident. Full reconstruction achieves only 33.03% (random chance), showing this approach cannot reconstruct complete surfaces of new objects.

Why position generalization fails. Different workspace positions fundamentally change the mechanical coupling in the system, affecting both stiffness and mass values in the equation above. Workspace 3's severe failure (23.3%) happens because its specific position creates completely different vibration patterns than Workspaces 1 and 2. Unlike object generalization where contact patterns can generalize

across specific frequencies, position changes alter the *entire vibration path* through the robot, creating workspace-specific responses that cannot be fixed with regularization. Acoustic sensing therefore requires workspace-specific training: workspace geometry determines how vibrations propagate through the system.

V. CONCLUSION

This work presents an analysis of acoustic sensing for 3-class contact detection (contact, no-contact, edge) on robot arms.

A. Summary of Findings

We investigated four research questions:

RQ1 (Proof of Concept): 3-class classification achieves 69.9% cross-validation accuracy, significantly above the 33.3% random baseline ($p<0.001$). Binary classification on the same generalization split achieves only 45.1%, which falls below the 50% random baseline, showing that excluding edge samples hurts performance.

RQ2 (Position Generalization): The three workspace rotations show catastrophic failure with highly variable performance (23.3%, 33.9%, 55.7%, average 34.5%). This is barely above the 33.3% random baseline, and two rotations fall below it. Acoustic signatures are fundamentally workspace-specific.

RQ3 (3-Class vs Binary): 3-class classification (34.5% average on generalization splits) marginally exceeds its 33.3% random baseline, while binary (45.1%) falls below its 50% random baseline. Including edge samples as a separate class is therefore preferable both for performance and for deployment, where boundary regions must be handled explicitly.

RQ4 (Object Generalization): Results depend heavily on the classifier. Most models perform near random (30-42% for 3-class), while binary completely collapses to exactly 50% across all classifiers. The Random Forest achieves the highest 3-class result at 41.7% (above the 33.3% baseline) without confidence filtering. Confidence filtering yields only 4 out of 2,280 positions retained, which is not statistically meaningful - across the full surface all models achieve random performance (33%).

The physics-based eigenfrequency analysis explains these results: Position generalization fails because workspace configurations create completely different acoustic distributions. Object generalization partially benefits from heavy regularization because dropout and weight decay prevent overfitting to object-specific frequencies, forcing the model to learn contact patterns (solid contact, air gap, partial overlap) that are more consistent across objects.

B. Implications and Future Work

Four practical conclusions follow from our results. **(1) Workspace-specific training is mandatory** - changing the robot's position shifts the entire acoustic propagation path, producing signal distributions that do not overlap with training data. This is a physical constraint, not a tuning problem. **(2) Object generalization for full surface reconstruction remains unsolved** - confidence filtering can find a small subset of reliable positions, but these cover only 0.2% of the surface, making full geometry reconstruction infeasible. **(3) Binary classification fails on novel objects** - edge samples carry discriminative information about contact boundaries that binary training discards; 3-class with explicit edge labels is essential. **(4) Acoustic sensing is viable in controlled settings but needs complementary sensing for robust deployment** - combining it with vision and force sensing is the most practical path forward.

For future work, the most promising short-term direction is workspace-invariant feature engineering: our eigenfrequency analysis suggests that contact type produces consistent patterns across objects, but workspace position shifts the absolute frequency values. Normalizing features against workspace-specific background responses could reduce this dependence. For object generalization, regularization strategies and domain-adversarial training - forcing the model to ignore which object produced a signal - are natural next steps. Longer term, combining acoustic sensing with vision and force sensing could address both limitations at once.

C. Limitations

This study has several limitations: (1) Limited dataset size - only 4 workspaces and 3 training objects. More workspace positions and more objects with different cutout patterns would be needed to draw stronger conclusions about generalization. (2) Material homogeneity - all objects are 3D-printed from the same material. (3) Single sensor - all data was collected with one acoustic finger, so results may not generalize to different finger designs. (4) Static contact evaluation, not tested on dynamic manipulation tasks. (5) **Dataset balancing challenges:** Initial data collection produced naturally imbalanced datasets where no-contact samples outnumbered contact and edge samples by 2-3×. This reflects the reality that empty workspace regions and object cutouts naturally create more no-contact positions than solid contact surfaces. This imbalance was discovered late during cross-workspace analysis and corrected through stratified subsampling to enforce perfect 33/33/33% class balance. While this ensures the reported results reflect genuine contact discrimination rather than exploiting class frequency biases, it raises practical deployment considerations: real-world manipulation involves non-uniform contact distributions (robots spend more time hovering than touching), so models trained on balanced data may need recalibration or confidence thresholding when deployed in naturally imbalanced scenarios. Future work should investigate training with class imbalance using cost-sensitive learning to match real manipulation statistics.

REFERENCES

[1] V. Wall, "Morphological sensing for soft pneumatic actuators based on acoustics and strain," Ph.D. dissertation, Technische Universität Berlin, 2019. [Online]. Available: https://depositonce.tu-berlin.de/bitstream/11303/10158/1/Wall_2019_Morphological_Sensing.pdf

[2] V. Wall, G. Zöller, and O. Brock, "Passive and active acoustic sensing for soft pneumatic actuators," *The International Journal of Robotics Research*, vol. 41, no. 3, pp. 260–277, 2022. [Online]. Available: https://arxiv.org/pdf/2208.10299.pdf

[3] G. Zöller, V. Wall, and O. Brock, "Active acoustic contact sensing for soft pneumatic actuators," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2438–2445, 2020. [Online]. Available: https://www.static.tu.berlin/fileadmin/www/10002220/Publications/Zoeller-20-ICRA_activeacoustic.pdf

[4] K. Zhang, D.-G. Kim, E. T. Tang, H.-H. Liang, Z. He *et al.*, "VibeCheck: Using active acoustic tactile sensing for contact-rich manipulation," *arXiv preprint arXiv:2504.15535*, 2025. [Online]. Available: https://arxiv.org/pdf/2504.15535.pdf