Guided by Touch: Teaching Robots How to Feel

Catharina Maschka
Department of Computer Science, University College London
e-mail: catharina.maschka.24@ucl.ac.uk

July 11th, 2025

***Keywords*:** Imitation Learning; Tactile Sensing; Robotics; Dexterous Manipulation

1 Introduction

Consider a task as mundane as determining whether the USB connector is inserted correctly. This seemingly simple task can be executed owing to our extremely complex perceptive abilities, allowing us to flip the USB connector based on the impedance force feedback obtained from jamming it into the socket the wrong way around. Nevertheless, these tactile tasks remain challenging to reproduce in robots [1]. For this reason, current robotic technology lacks the dexterity and adaptability required for nuanced tasks in dynamic environments.

However, combining soft robotic technology with imitation learning opens the possibility of emulating human manipulation skills. Integrating innovations in machine learning with the latest sensor technology may transform sectors such as surgical robotics, crop harvesting, and extraterrestrial operations, enabling complex manipulation tasks to be executed independently of human control.

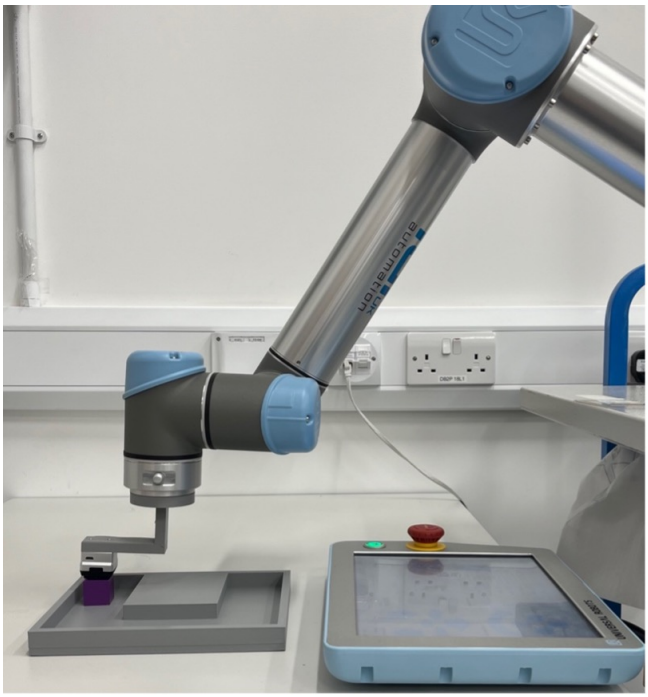

Figure 1: Task Setup

This paper demonstrates how tactile sensing via GelSight sensors can train a Universal Robot UR5 (a 6-degree-of-freedom robotic arm, depicted in Figure 1) to navigate a cube through a maze using imitation learning. This is achieved without

any global vision input or a visual map of the maze to aid navigation, which has not been previously done.

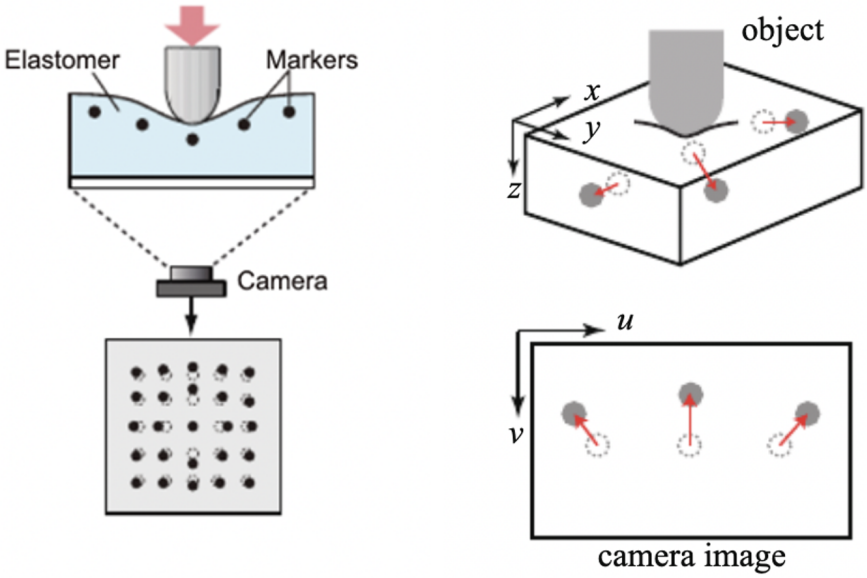

Figure 2: Diagram of GelSight Sensor Adapted from [2]

The GelSight sensor offers a novel approach to tactile sensing that does not rely on force detection. Instead, the GelSight sensor detects the geometry of the contact surface via a deformable elastomer. An embedded camera visualises the deformation of the soft reflective membrane, illustrated in Figure 2, by illuminating it with LEDs [2]. Furthermore, markers within the soft elastomer help indicate the degree of applied local forces and shear, allowing high spatial-resolution tactile frames to capture directional information, as shown in Table 1 below. From Figure 3, it is evident that changes in the

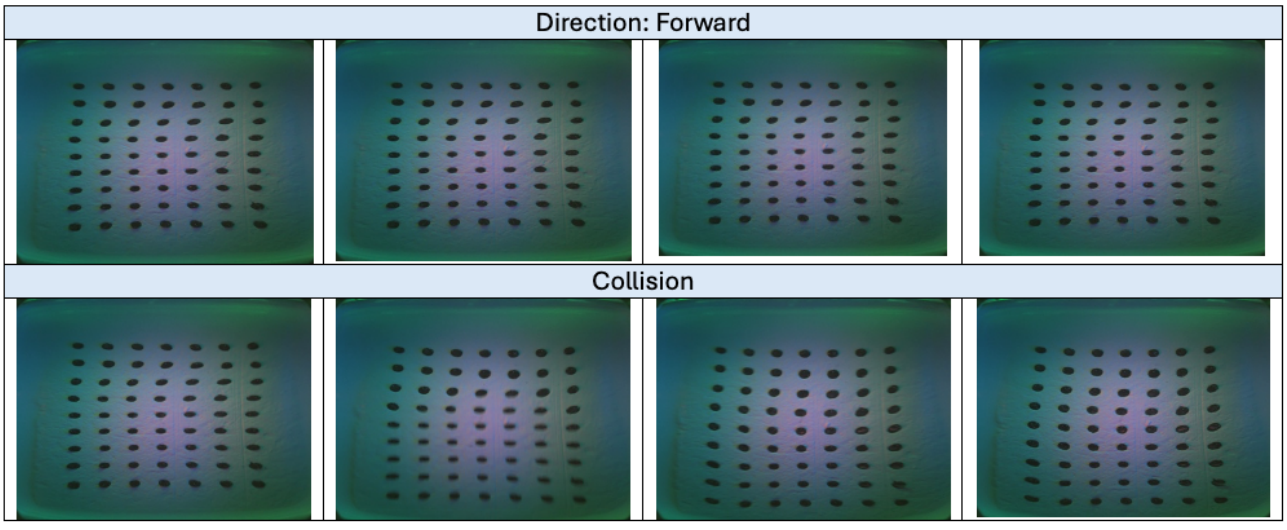

Figure 3: Sequential tactile frames captured by GelSight sensor

tactile images are extremely subtle and often imperceptible to the human eye. Therefore, machine learning techniques such as imitation learning are essential for extracting meaningful patterns from the data and achieving human-like tactile manipulation.

In essence, imitation learning utilises observations from expert demonstrations (e.g., human) to generate robot actions [3]. The aim is to employ data from human demonstrations to teach the UR5 navigation strategy, rather than memorising a fixed trajectory. This will enable the model to generalise to any simple maze design.

Long Short-Term Memory (LSTM) networks are designed to capture long-term temporal relationships, making them a natural fit for predicting sequential tactile interactions, like the ones depicted in Figure 3 [4]. A bidirectional LSTM (BiLSTM) model further improves time series predictions by processing sequential data in a forward and backward direction, enhancing its contextual understanding of the inputs [5].

1.1 Related Work

Robotic manipulation has long been reliant on vision. However, recent advancements, especially with high-resolution tactile sensors like GelSight, promise to change the reliance on visual data by offering a more nuanced, force-sensitive interface. For instance, [6] present an imitation learning framework focusing on robotic match lighting. It was concluded that including tactile information, in addition to visual inputs, boosted performance by 40%, highlighting the value of tactile-based data.

Early work by [7] provides a prime example of touch-conditioned planning (without vision), where predictive tactile models directly control complex, contact-rich manipulations, such as controlling balls, joysticks, and dice. The paper defines tasks through goal-oriented tactile images and utilises model-predictive control (MPC) to achieve this ideal tactile state at each time step. While this goal-conditioned MPC approach is effective at achieving immediate tactile objectives, it lacks the capacity for long-horizon exploration or adaptation in unknown environments. In contrast, our paper implements a LSTM model, which excels at capturing order-sensitive, temporal dependencies in sequential data to support both long-horizon planning and adaptive responses.

Action Chunking with Transformers (ACT), introduced by [8], is an imitation learning algorithm that predicts sequences of actions as unified chunks rather than individual steps, reducing task horizon and improving long-horizon planning [9]. Using transformer self-attention, ACT captures non-Markovian dependencies, making it well suited for sequential tasks such as tactile manipulation [10], [3]. Unlike step-wise models, ACT mitigates compounding errors and demonstrates strong performance (80–90% success) on real-world manipulation tasks with minimal visual demonstration time [3]. However, ACT is computationally costly as the input sequence increases [11].

For this reason, a BiLSTM network will serve as the basis for this project. However, to address the limitations of single predictive models, [12], along with [13] and [14], propose a hybrid model, combining a Temporal Convolutional Network (TCN) with a BiLSTM. Results showed improved prediction accuracy for noisy time-series tasks compared to standalone models as well as enhanced computational efficiency [12], [13]. Therefore, a hybrid model is adopted.

2 Materials and Method

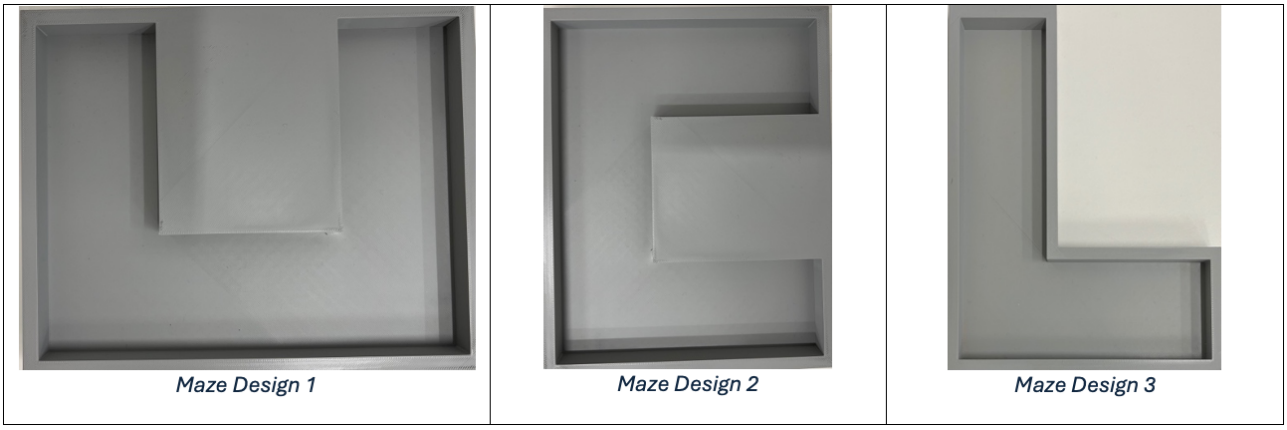

Figure 4: Various 3D printed maze designs used in data collection

As outlined in Figure 4, various mazes are designed to maximise the variety of tactile data obtained. Including sharp edges rather than rounded ones is intentional, ensuring that the tactile image generated during cube-wall interaction differs significantly from free-moving images.

In addition to other essential components, such as the GelSight sensor and UR5, end-effector design is crucial for maintaining surface contact between the GelSight sensor and the textured cube surface while navigating the maze. Refer to Figure 5 below for the experimental setup and end-effector design used during data collection. It is essential to note that the 3D-printed end effector is slightly compliant, thereby protecting the fragile GelSight sensor from harsh impacts.

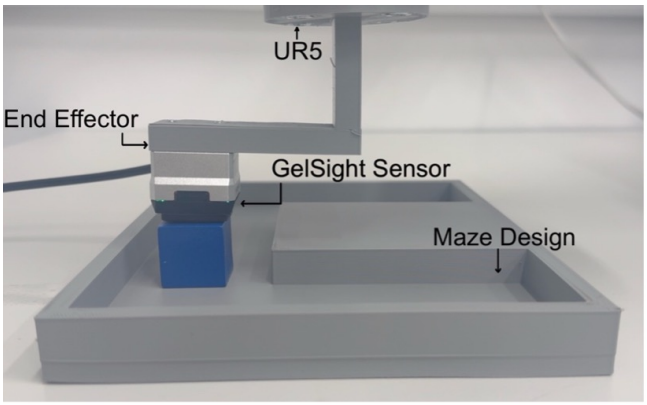

Figure 5: Experimental setup

2.1 Tactile Demonstration

The method aims to produce exploration-driven demonstrations of each maze design. As seen in Figure 5, a cube is placed within the maze, and the UR5 is lowered so that the GelSight sensor is centred on the cube, and the tactile image clearly displays the cube's surface. This allows the UR5 to manoeuvre the cube through the maze without losing contact. As the robot naturally explores the environment during demonstrations, the cube's direction (forward, backwards, right, left, or none) is updated manually and stored alongside additional data, including the end-effector coordinates, joint angles, and

change in position, which is sampled at 10 Hz by the UR5 system. Every GelSight frame is paired with a timestamped direction label and corresponding robot state, creating a synchronised, multimodal dataset. The label "none" describes no movement at the beginning of data collection and serves as a baseline tactile image.

Since wall collisions serve as implicit cues that guide the next step during demonstrations, the entire interaction period is also manually labelled as "collision" and stored within the dataset. This process is repeated across 40 demonstration runs for each of the three maze designs.

2.2 Controls

Certain factors must be kept consistent during data collection to ensure that the explorative navigations are independent of visual cues from the demonstrator. For instance, a similar distance must be moved in repeatable time intervals. This helps eliminate subconscious adjustments (e.g., slowing down near walls) that may be guided by vision. Furthermore, demonstrations require a standardised recoil distance to ensure that the trained model can establish a generalizable response upon collision with a wall.

3 Model Architecture

3.1 Classification Model

Before training a complex predictive model with the collected data, it is crucial to validate that meaningful information can be extracted from the tactile images. To achieve this, a simple image classification model was designed, as outlined in Figure 6. The preprocessing pipeline transforms individual frames in an episode into temporal sequences by grouping them according to their corresponding actions. Consecutive frames sharing the same action label (e.g. forward, backwards, left, right, or none) are grouped. A fixed chunk size of 10 consecutive frames, equivalent to 1 second of tactile data (sampled at 10 Hz), is used, as it captures the tactile dynamics of a single complete motion (e.g., one forward push or post-collision recovery).

An early fusion approach is employed by concatenating 10 tactile images along the channel axis. This ensures that the model sees spatial and temporal information simultaneously from the outset [15]. Although this approach is not ideal, as explored by [16], the ease of implementation and computational efficiency make it suitable for obtaining preliminary results [17].

A ResNet-18 backbone (pre-trained on the ImageNet dataset) is used, with five output classes for directional classification and two output classes for collision classification. Regarding critical training parameters, a batch size of 16 was chosen due to memory constraints and to prevent overfitting (the dataset shrinks significantly after grouping by action) [18]. Furthermore, the ADAM optimiser is utilised with a base learning rate of 0.0005 to preserve the pre-trained ImageNet features, as suggested by [19]. A higher-than-default weight decay of 0.001 is implemented to improve L2 regularisation against overfitting on limited tactile data. Additionally, 30 epochs were chosen, but early stopping was implemented to mitigate overfitting and enhance training efficiency [20].

3.2 Predictive Model

The hybrid deep learning model integrates the previously trained classification model for visual feature extraction with a hierarchical temporal stack comprising BiLSTM and Temporalc Convolution (TC), as depicted in Figure 7. The final output is aggregated using global average pooling (GAP) and classified using a multilayer perceptron (MLP) with dropout (0.3) for regularization to avoid overfitting [21]. Likewise, by reducing dimensionality, GAP prevents overfitting [22]. A

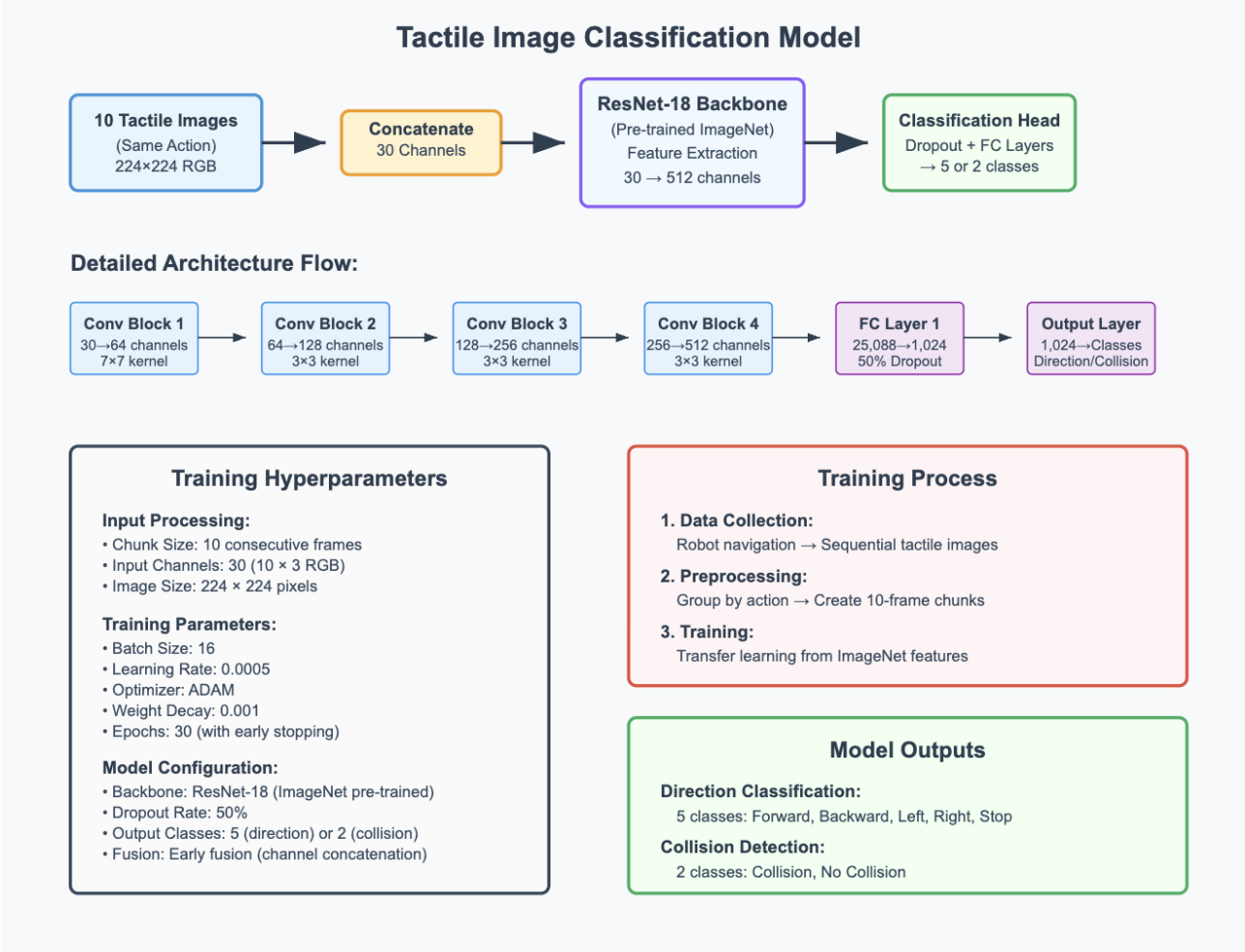

Figure 6: A diagram of the tactile image classification model architecture

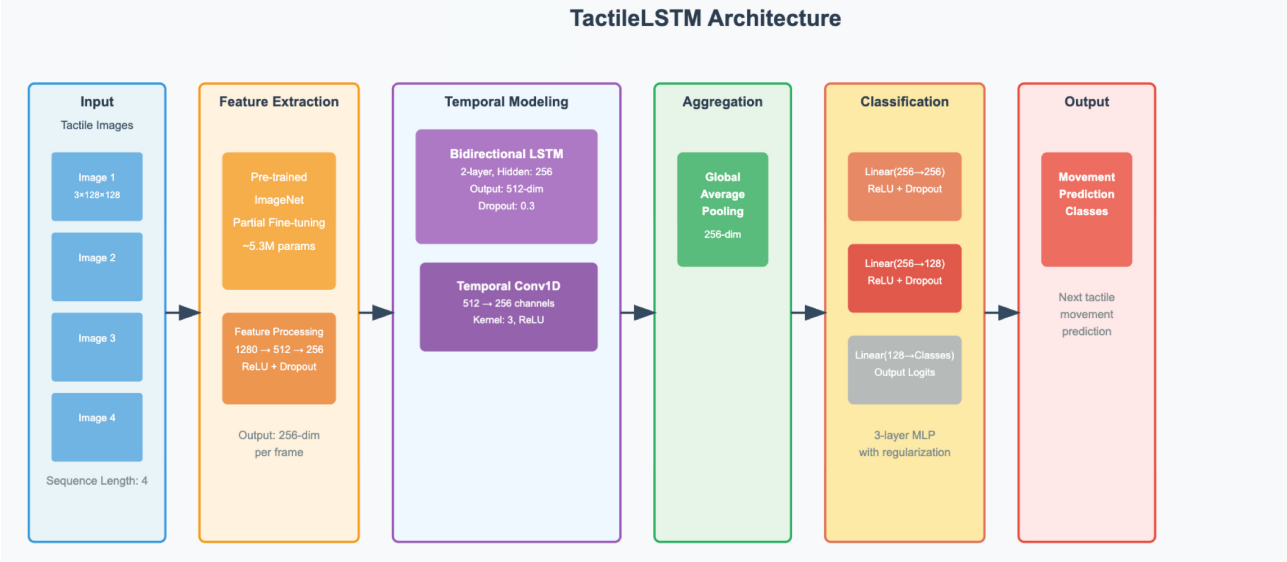

Figure 7: A diagram of the TCN-BiLSTM predictive model architecture

3-layer MLP with ReLU activation accelerates convergence, therefore serving as a computationally cheap and efficient classifier [23]. Together, GAP and ReLU ensure that the architecture remains lightweight yet robust.

4 Results

4.1 Preliminary Results

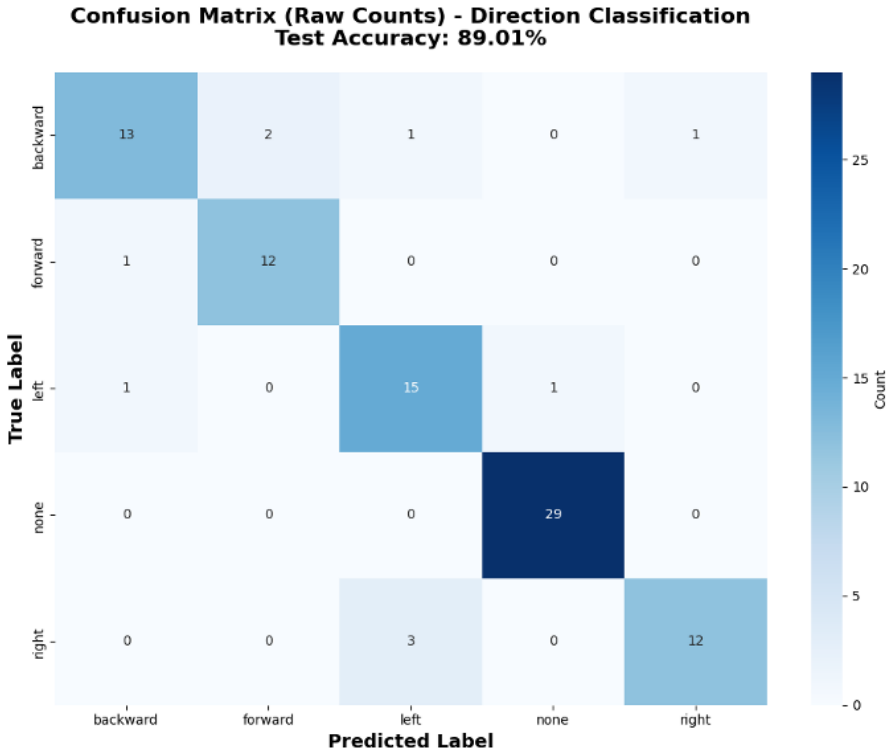

Figure 8: Confusion Matrix Direction Classification Test Results

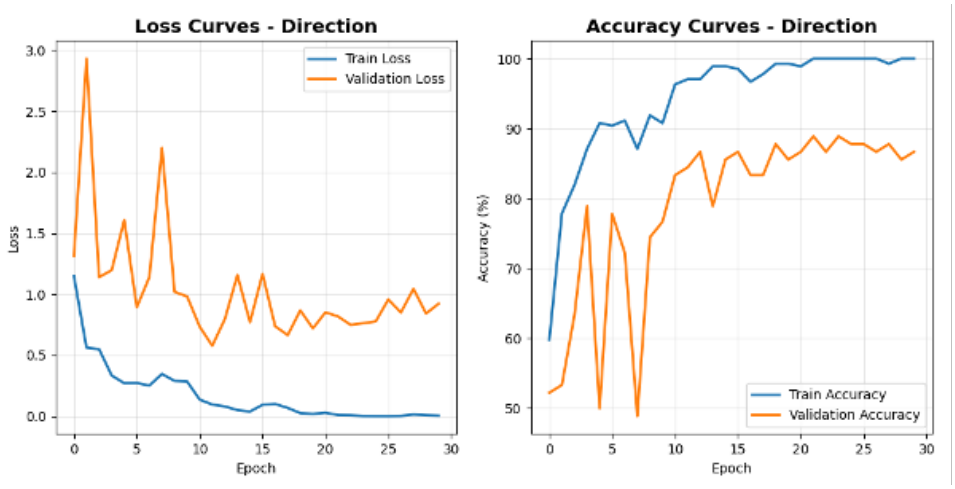

Figure 9: Preliminary results in direction classification: Training Loss (Figure 9a) and Accuracy (Figure 9b)

Preliminary results indicate that tactile data chunks can be accurately classified to a large extent according to both movement direction and collision type. From Figure 8, the direction model achieves 89% test accuracy, with an exceptional ability (100% accuracy) to identify no motion ("none"). Furthermore, classifying forward motion was the most accurate, followed by leftward motion (with accuracies of 92% and 88%, respectively).

Nevertheless, Figure 9a shows erratic fluctuations in validation loss, suggesting unstable learning. [24] highlights that such behaviour arises from suboptimal hyperparameter selection, noisy data, or inadequate preprocessing. Furthermore, the slight increase in the validation loss, as the training loss continues to decrease, may be indicative of overfitting [25]. Although employing a pre-trained network saves training time and requires less data, it may not generalise well to specific tasks, such as feature extraction from tactile images [26]. This is illustrated by the notable generalisation gap formed between the Train and Test Accuracy in Figure 9b. The lack of convergence implies that the model struggles to extract consistent tactile features for direction classification.

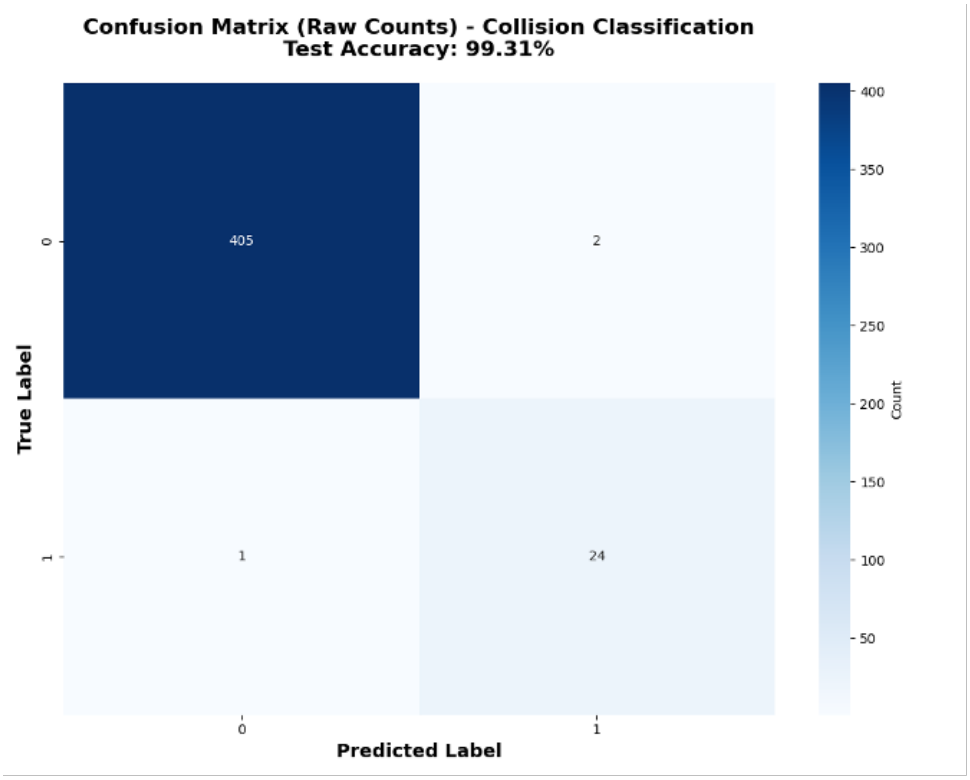

Figure 10: Confusion Matrix Direction Classification Test Results

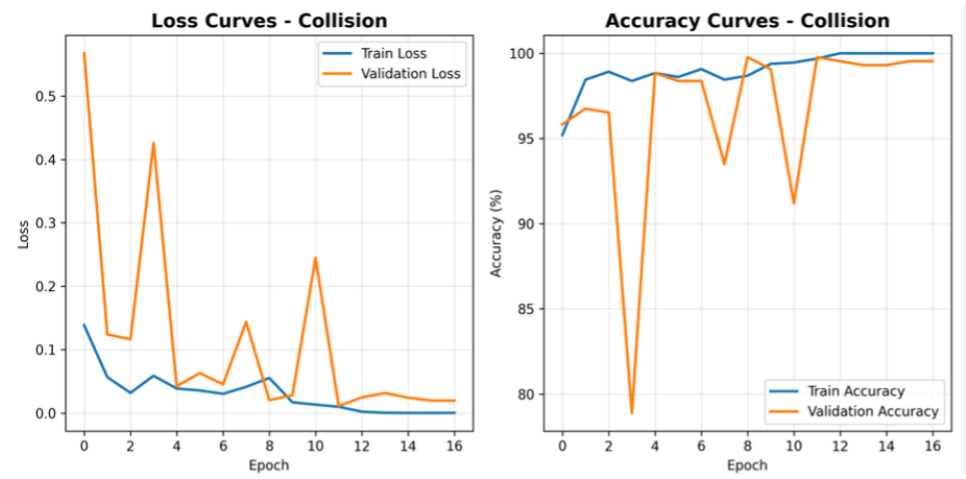

Figure 11: Preliminary results in collision classification: Training Loss (Figure 11a) and Accuracy (Figure 11b)

On the other hand, as shown in Figure 10, the collision model with 100% training accuracy achieves a test accuracy of 99.31%. The near-perfect collision classification is likely due to the distinct and consistent tactile contact patterns that occur. The slightly superior accuracy in identifying no collision ("0") versus collision ("1") (100% and 96% accuracy, respectively) may be due to the significantly smaller number of collision samples.

Both the training and validation loss curves in Figure 11a converge towards 0. In conjunction with the high train and test accuracy, this indicates strong convergence and generalises well on unseen tactile data. The fluctuations in the validation loss curves may be linked to outlier samples, as manual collision labelling is subjective, which can introduce significant human error. Despite the obvious limitations in the model design, the test accuracy indicates that the demonstration data is meaningful and learnable to a certain extent, paving the way for an effective predictive model.

4.2 Final Results

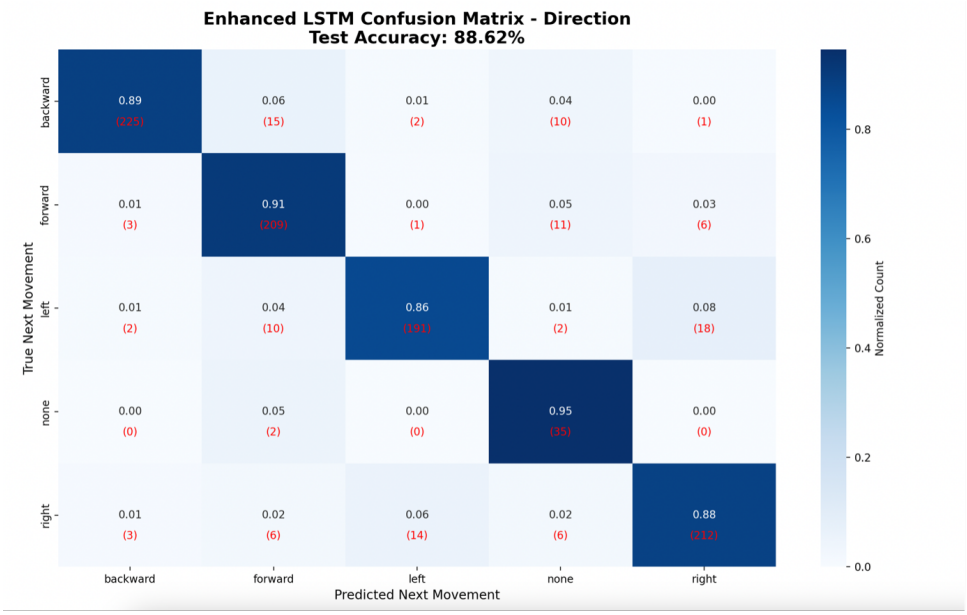

Figure 12: LSTM confusion matrix for predicted next movement

The results in Figure 12 indicate that, in theory, tactile manipulation skills have been acquired with high accuracy (88.62%). As expected, the "none" class – indicating no movement– is predicted with the most accuracy due to the simplicity of the tactile frame. This performance represents the model's current upper bound, highlighting room for improvement. The high accuracy in the classification of forward motion in Figure 8 aligns with the model's highest directional prediction accuracy (91%). This correlation between classification and prediction is consistent for backward and rightward motion. Interestingly, despite similarly strong classification results for leftward motion (Figure 8), it yielded the lowest directional prediction accuracy at 86%, indicating potential asymmetry in temporal modeling.

5 Discussion

5.1 Limitations and Improvements

The quality of manual demonstration data impacts all aspects of the project. The manually guided demonstration limits the quality of the data due to subconscious visual inputs and consistency in the demonstrations. Likewise, the accuracy of

the manual labels is subject to human error, contributing to noisy labels, which may reduce the model's ability to identify precise boundary events. Furthermore, collision labelling is particularly subjective and especially limited by human reaction time. A more reliable alternative is to implement an automatic collision detection system based on predefined velocity thresholds.

Unintentional biases in the maze geometry also influence the robot's performance. Narrower maze designs require significantly more precision and leave less margin for post-collision correction. This incorporates a task-specific bias that may prevent the model from generalising to slightly different maze designs. Furthermore, complex mazes – ones without sharp corners, multiple paths, and dead ends – may require more sophisticated maze-solving strategies that are not taught in these demonstrations [27].

On the algorithmic side, although LSTM is a computationally efficient model, it's stepwise architecture leads to compounding errors [3]. As mentioned previously, ACT , despite being more computationally costly, may be more suitable for complex manipulation tasks as it is not susceptible to compounding errors, and it excels long-horizon planning.

Furthermore, despite the high resolution of GelSight tactile frames, the shape limits its versatility. The project could expand the complexity of the dexterous manipulation task if the GelSight sensor were replaced with a 3D tactile finger like the one explored in [28]. The 3D nature would create a more holistic dataset, and may even allow the use of other, more complex shapes, like spheres, to be used in the maze navigation. This is because learning to regain contact with a ball if the sensor slips during maze navigation would require more dynamic tactile data, which can be achieved with a new sensor design. This additional layer of complexity would be one step closer to achieving human level dexterity skills.

6 Conclusion

In summary, while the LSTM-based model demonstrated strong theoretical performance in tactile classification and prediction, its real-world deployment on the UR5 revealed limited generalization, as the robot was only able to consistently execute rightward motion. This highlights a clear gap between offline accuracy and embodied performance, emphasizing the need for improved data quality, more robust modeling, and richer tactile sensing to achieve reliable human-like dexterity in practice.

6.1 Reflection

From the outset, I was exposed to what it truly means to conduct research at the forefront of scientific knowledge. We were in a constant state of problem-solving, as we continuously encountered issues that required us to revise our methodology. It took two weeks just to establish a robust methodology for a significantly simplified version of the original task. Although we had originally envisioned maneuver a ball through a maze – a task that is even difficult to execute as a human without visual input – we eventually settled for navigating a cube instead. This simplification prevents contact loss with the object, which would otherwise complicate the task astronomically. Perfecting the demonstration data collection also required extensive trial and error. Maneuvering the robotic arm by hand was difficult, as we were too clumsy, leading to poor-quality data (not "expert-level" data needed in imitation learning). After trying various teleoperation techniques, we settled on using the built-in robotic remote-control interface. Furthermore, the training data could also not be influenced by any subconscious visual input from the demonstrator, requiring additional methodological adaptations. Once data were collected, three different algorithms were tested; optimizing for computational efficiency and accuracy, my data analysis skills were honed extensively. The failure of the real-world implementation of the model was the most important, yet heartbreaking experience that I didn't know I needed.

References

[1] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, pp. 2762–2762, Nov. 2017. DOI: 10.3390/s17122762. (visited on 06/05/2025).

[2] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors*, vol. 19, pp. 3933–3933, Sep. 2019. DOI: 10.3390/s19183933. (visited on 06/05/2025).

[3] H. Xue, J. Ren, W. Chen, *et al.*, *Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation*, Cornell University, Apr. 2025. [Online]. Available: https://arxiv.org/abs/2503.02881# (visited on 06/07/2025).

[4] F. Pastor, J. García-González, J. M. Gandarias, *et al.*, "Bayesian and neural inference on lstm-based object recognition from tactile and kinesthetic information," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 231–238, 2021. DOI: 10.1109/LRA.2020.3038377.

[5] R. Hamad, *What is lstm? introduction to long short-term memory*, Medium, Dec. 2023. [Online]. Available: https://medium.com/@rebeen.jaff/what-is-lstm-introduction-to-long-short-term-memory-66bd3855b9ce (visited on 06/24/2025).

[6] N. Funk, C. Chen, T. Schneider, and J. Peters, "On the importance of tactile sensing for imitation learning: A case study on robotic match lighting," *ResearchGate*, Apr. 2025. DOI: 10.48550/arXiv.2504.13618. (visited on 06/08/2025).

[7] S. Tian, F. Ebert, D. Jayaraman, *et al.*, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 818–824. DOI: 10.1109/ICRA.2019.8794219.

[8] T. Zhao, V. Kumar, S. Levine, and C. Finn, *Learning fine-grained bimanual manipulation with low-cost hardware*, Apr. 2023. [Online]. Available: https://arxiv.org/pdf/2304.13705.

[9] A. Lee, I. Chung, L.-Y. Chen, and I. Soltani, *Interact: Inter-dependency aware action chunking with hierarchical attention transformers for bimanual manipulation*, Arxiv.org, 2018. [Online]. Available: https://arxiv.org/html/2409.07914v1 (visited on 06/27/2025).

[10] T. Parr, G. Pezzulo, and K. Friston, "Beyond markov: Transformers, memory, and attention," *Cognitive Neuroscience*, 2025. DOI: 10.1080\/17588928.2025.2484485. (visited on 07/01/2025).

[11] J. Xie, P. Cheng, X. Liang, Y. Dai, and N. Du, "Chunk, align, select: A simple long-sequence processing method for transformers," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13 500–13 519. DOI: 10.18653/v1/2024.acl-long.729.

[12] J. Bai, W. Zhu, S. Liu, C. Ye, P. Zheng, and X. Wang, "A temporal convolutional network–bidirectional long short-term memory (tcn-bilstm) prediction model for temporal faults in industrial equipment," *Applied Sciences*, vol. 15, p. 1702, Feb. 2025. DOI: 10.3390/app15041702.

[13] S. Baker, W. Xiang, and I. Atkinson, "A computationally efficient cnn-lstm neural network for estimation of blood pressure from features of electrocardiogram and photoplethysmogram waveforms," *Knowledge-Based Systems*, vol. 250, p. 109 151, May 2022. DOI: 10.1016/j.knosys.2022.109151.

[14] Y. Zhang, Z. Kan, Y. Tse, Y. Yang, and M. Wang, *Fingervision tactile sensor design and slip detection using convolutional lstm network*, Oct. 2018. DOI: 10.48550/arXiv.1810.02653.

[15] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, *Early vs late fusion in multimodal convolutional neural networks*, IEEE Xplore, Jul. 2020. DOI: 10.23919/FUSION45008.2020.9190246. (visited on 06/21/2023).

[16] Y. Hu, M. Lu, and X. Lu, "Spatial-temporal fusion convolutional neural network for simulated driving behavior recognition," *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, vol. 17, Nov. 2018. DOI: 10.1109/icarcv.2018.8581201. (visited on 06/25/2025).

[17] G. Barnum, S. Talukder, and Y. Yue, *On the benefits of early fusion in multimodal representation learning*, arXiv.org, 2020. [Online]. Available: https://arxiv.org/abs/2011.07191?utm_source=chatgpt.com (visited on 06/25/2025).

[18] D. Singh, *Mini-batch size in deep learning: A balancing act for fast convergence and strong generalization*, Medium, Jan. 2025. [Online]. Available: https://medium.com/ai-enthusiast/mini-batch-size-in-deep-learning-a-balancing-act-for-fast-convergence-and-strong-generalization-5222a795ce88 (visited on 06/25/2025).

[19] Y. Shao, J. Yang, W. Zhou, *et al.*, "An improvement of adam based on a cyclic exponential decay learning rate and gradient norm constraints," *Electronics*, vol. 13, p. 1778, May 2024. DOI: 10.3390/electronics13091778.

[20] U. of Toronto, *Preventing overfitting*, Toronto.edu, 2025. [Online]. Available: https://www.cs.toronto.edu/~lczhang/360/lec/w05/overfit.html (visited on 06/25/2025).

[21] P. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment analysis using word2vec and long short-term memory (lstm) for indonesian hotel reviews," *Procedia Computer Science*, vol. 179, pp. 728–735, Jan. 2021. DOI: 10.1016/j.procs.2021.01.061.

[22] S. P. Bijoy, *Global average pooling and gradient tape: Simplifying deep learning*, Medium, Jan. 2025.

[23] Ashutosh, *The power of rectified linear unit (relu) activation function in multilayer perceptrons (mlps)*, Medium, May 2023. [Online]. Available: https://ashutoshkriiest.medium.com/the-power-of-rectified-linear-unit-relu-activation-function-in-multilayer-perceptrons-mlps-2ce6032e9f1d (visited on 06/24/2025).

[24] P. Hallaj, *How to tame noisy training loss in deep learning: Strategies and tips*, Medium, Jan. 2024. [Online]. Available: https://medium.com/@pouyahallaj/how-to-tame-noisy-training-loss-in-deep-learning-strategies-and-tips-c68213c6b6b9 (visited on 06/24/2025).

[25] P. Mohajerani and H. Sotoudeh, *Essentials of ai techniques: With a focus on medicine and healthcare*, ResearchGate, Feb. 2020. [Online]. Available: https://www.researchgate.net/publication/349898718_Essentials_of_AI_Techniques_With_a_focus_on_medicine_and_healthcare (visited on 07/02/2025).

[26] X. Tan, T. Li, R. Chen, F. Liu, and L. Zhang, *Challenges of using pre-trained models: The practitioners' perspective*, arXiv.org, 2024. [Online]. Available: https://arxiv.org/abs/2404.14710 (visited on 06/25/2025).

[27] R. Kumar, P. Jitoko, S. Kumar, *et al.*, "Maze solving robot with automated obstacle avoidance," *Procedia Computer Science*, vol. 105, pp. 57–61, 2017. DOI: 10.1016/j.procs.2017.01.192. (visited on 06/30/2025).

[28] R. Kõiva, M. Zenker, C. Schürmann, R. Haschke, and H. J. Ritter, "A highly sensitive 3d-shaped tactile sensor," in *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 2013, pp. 1084–1089. DOI: 10.1109/AIM.2013.6584238.