Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Master's Thesis

Multimodal Deep Anomaly Detection for Robot Pouring Task

Adithya Narasimhaiah Sathish

Submitted to Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
in partial fullfilment of the requirements for the degree
of Master of Science in Autonomous Systems

Supervised by

Prof. Dr. Paul G. Plöger
Prof. Dr. Nico Hochgeschwender
Santosh Thoduka

December 2023

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work.

Date	Adithya Narasimhaiah Sathish

Abstract

Pouring involves the transfer of materials or liquids from one container to another and holds significant importance for human experts across various domains. This activity or task is labor-intensive, repetitive, and, in certain cases, poses potential dangers. Automation of pouring tasks has been facilitated with the assistance of robots, promising efficiency and safety in diverse applications.

Robots employed in pouring tasks can exhibit anomalous behavior due to uncertainties. These behaviors not only heighten the human operator workload but also pose dangers in safety-critical pouring applications. Therefore, anomaly detection is important for robots to attain pouring abilities that are on par with those of human experts.

This research work aims to utilize data-driven or machine learning-based anomaly detection methods to identify pouring anomalies. The anomaly detection methods are trained in a semi-supervised manner using a custom multimodal pouring dataset containing both normal and anomalous pouring executions. The work also includes a comparative study of anomaly detection methods and an ablation study to identify crucial sensor modalities for pouring anomaly detection.

Acknowledgements

"Gratitude can transform common days into thanksgivings, turn routine jobs into joy, and change ordinary opportunities into blessings." - William Arthur Ward

First and foremost, I would like to express my gratitude to Germany, its citizens, and taxpayers for providing me access to free education. I extend my sincere thanks to my supervisors, Prof. Dr. Paul G. Plöger, Prof. Dr. Nico Hochgeschwender, and Santosh Thoduka, for their invaluable feedback and guidance. Special thanks to my friend Proneet for his unwavering support throughout my master's studies. I dedicate this thesis to my parents, Sri. Sathish Narasimhaiah and Smt. Lalitha Sathish, my sister Putty, and my grandparents, Sri. Narasimhaiah Shankaranarayana and Smt. Padmavathi Narasimhaiah.

Contents

List of Figures

List of Tables

1

Introduction

"Where shall I begin, please your Majesty? he asked.
Begin at the beginning, the King said gravely, and go on till you come to the end: then stop"

"Alice's Adventures in Wonderland" by Lewis Carroll

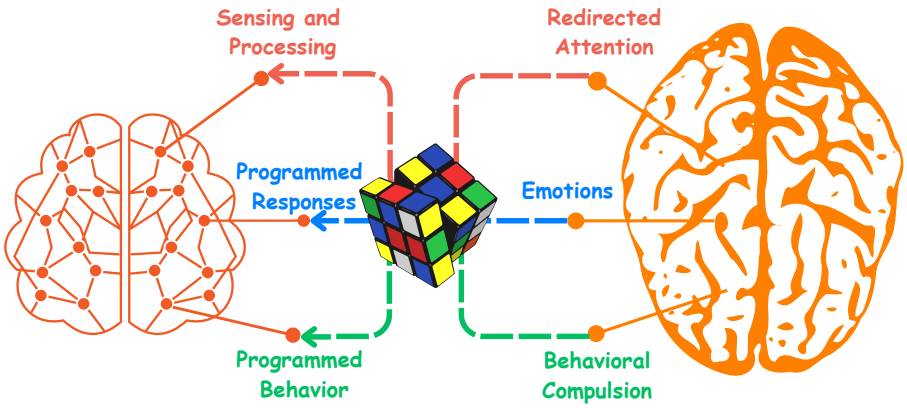


Figure 1.1: Animals and human beings have natural biological tendencies to identify anomalies, thanks to Mother Nature and the process of evolution. For example, animals can respond to changes in smell and go in search of food. Early humans relied on their five sensory organs to sense environmental anomalies as an indication of potential dangers. Similar to animal and/or human instincts, as outlined by clinical psychologist Jordan Peterson in his book Maps of Meaning, technology has embraced the concept of anomaly detection to develop new age devices and robots. The figure demonstrates the mapping of human intelligence to machine intelligence challenging and timely process [1] and is created using www.canva.com.

Anonymous: Who are you, Easter?
Easter: I am Easter, alive because of one man who
was supposed to be dead but resurrected on the third day.
Anonymous: That's quite anomalous, isn't it?

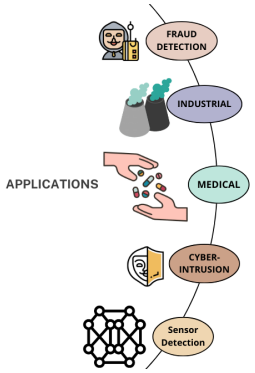


Figure 1.2: Wide applications of anomaly detection. Figure created using www.canva.com

This thought-provoking question-and-answer dialogue sets the stage for a deeper discussion of anomalies and anomaly detection. Events, observations, or system behaviors that deviate from the well-established norm are known as anomalies, and the process of identifying anomalies is called anomaly detection [6]. Anomaly detection is widely used in a variety of fields, from medical sciences to military applications to robotics. For example, in the medical sciences, anomaly detection helps in identifying diseases such as chest lesions [7]. In military applications, it aids in monitoring restricted areas [8]. In finance, it assists in identifying abnormal bank transactions [9]. Currently, researchers are actively distilling the 250-year-old science of anomaly detection in robotics, aiming to enhance the cognitive capabilities of robots [1]. For example, [10] focused on teaching robots to recognize anomalies related to book placements, and [11] focused on teaching robots about anomalies related to table-top manipulations.

1.1 Motivation

Pouring is a physical activity wherein materials (solids or liquids) are transferred from one container to another in a controlled manner. A bartender skillfully pouring your favorite drink from a bottle into a mocktail glass or a foundry setup transferring molten metal from a bucket to molds are some examples of pouring activities, as illustrated in Figure 1.3.

Pouring activities play a crucial role across various industries, presenting workers with physical challenges and safety concerns. In foundries, human workers engage in physically demanding work by repeatedly lifting heavy containers to pour molten metal, posing a direct threat to their safety due to the molten nature of metals. Pouring molten metals into a container for alloy preparation necessitates precise pouring, but human experts may encounter difficulties in consistently achieving the necessary accuracy due to various factors. The varied scope of pouring activities indicates that they generally involve labor-intensive, repetitive tasks, demand high accuracy, and exhibit hazardous characteristics.

Traditionally, robots have been utilized to speed up and automate tasks that are monotonous, time-consuming, laborious, and potentially dangerous for human experts. The *Adam* robot, developed by *RichTech Robotics*, is employed as a coffee bartender, freeing human workers from the repetitive nature of pouring coffee. Similarly, the Kawasaki robot is utilized in foundry operations, where it excels at the hazardous task of pouring hot molten aluminum into molds. The use of robots not only ensures efficiency but also enhances safety by reducing human exposure to risks, showcasing the valuable role of robots in mitigating challenges posed by any activity.

Figure 1.3: *The left image demonstrates the dangerous nature of pouring molten metal in a foundry setup and the image is sourced from https://www.wingspans.com/story/calvin-hubbard. The right image demonstrates the pouring of a cocktail from a bottle to a glass in a restaurant and the image is sourced from https://www.tastingtable.com/1175386/where-does-the-bartending-term-well-drink-come-from/*

.

Figure 1.4: *Adam robot is employed as a coffee maker. The image of Adam robot is taken from https://www.richtechrobotics.com/adam*

Like human experts and machinery, robots also experience anomalies due to numerous sources of uncertainty [2]. Uncertainties may arise from human or power intervention during the execution of a task or from physical damage to the on-board sensory system. As a consequence, robots exhibit anomalous behavior and produce undesired results, and innocent robots are not even aware of anomalies.

We require a diagnostic procedure to detect and identify anomalies [2], which consists of three steps illustrated in Figure 1.5. In this procedure, the first step, *Anomaly Detection*, identifies anomalies, while the middle step, *Anomaly Identification* [2], classifies them. In the final stage of *Recovery*, the robot takes

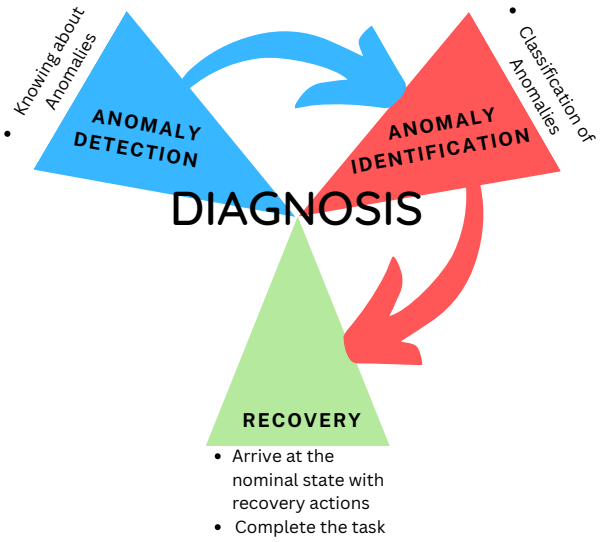


Figure 1.5: Three step diagnostic procedure [2]

actions to return to the normal state and complete the task [2]. The scope of this research is focused on the first step of the diagnostic procedure, i.e., *Anomaly Detection*, specifically for the robot pouring task in an indoor environment.

1.2 Challenges and Difficulties

In the following section, we briefly highlight the complexities, challenges, and difficulties related to pouring anomaly detection.

- **Ch1 Complex Anomalies:** Anomalies are typically considered rare, unknown, and heterogeneous [12]. Pouring anomalies are rare in the sense that they seldom occur in an instrumented setup. Analogous to unhappy families in Leo Tolstoy's quote, pouring anomalies are unique and very different from each other. Heterogeneous anomalies are complex and can only be captured using multiple sensor modalities [12]. Very few methods focus on using multiple modalities for anomaly detection [12].
- **Ch2 Lack of Datasets:** Real-world anomaly detection datasets are scarce [12], with the majority focusing on computer vision, often incorporating image or video modalities. Furthermore, there are a limited number of datasets specifically for robot manipulation anomaly detection. [10] introduced a multimodal dataset focusing on book placement activities, collected using the Toyota Human Support Robot. Similarly, [11] presented a dataset to identify table-top manipulation anomalies. While the book placement dataset is accessible to the general public, the dataset that [11] presented is no longer available to the research community. It is recommended that datasets recognize and incorporate various types of anomalies for comprehensive coverage and should not be limited to point anomalies [12].

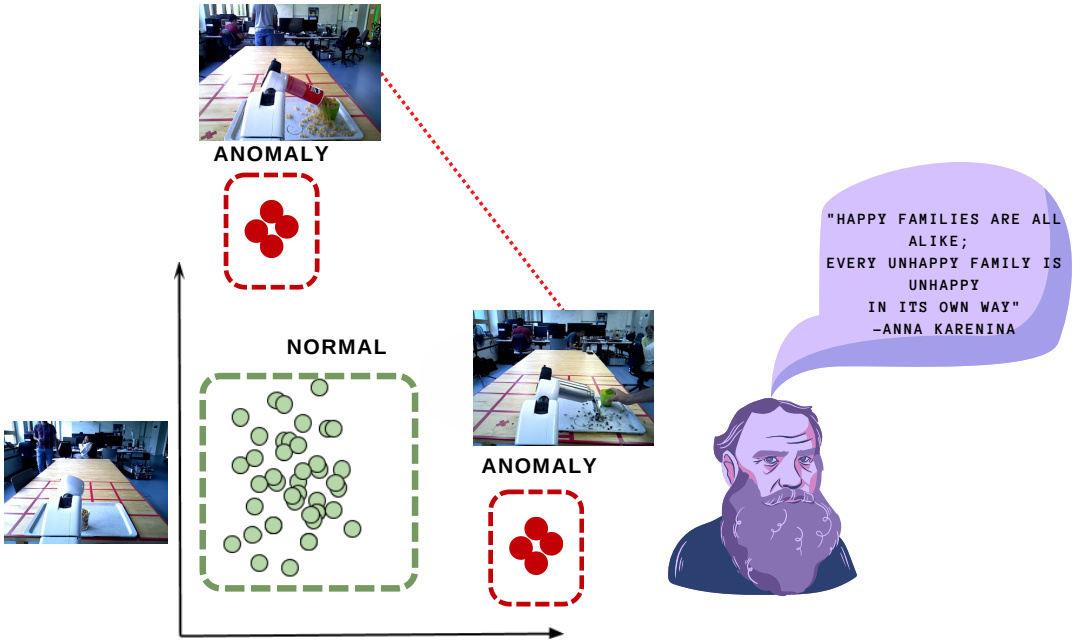


Figure 1.6: *Heterogeneous nature of pouring anomalies. The image of Leo Tolstoy is taken from www.canva.com*

- **Ch3 Sample and Computation Efficient:** Classical machine learning methods employ hand-crafted features and rely on a smaller number of samples to learn representations, but they exhibit a high false positive rate [12]. In contrast, deep learning methods are statistically superior but depend on a large sample set [12]. The challenge of designing methods that can learn representations from a very small sample set remains unresolved [12].

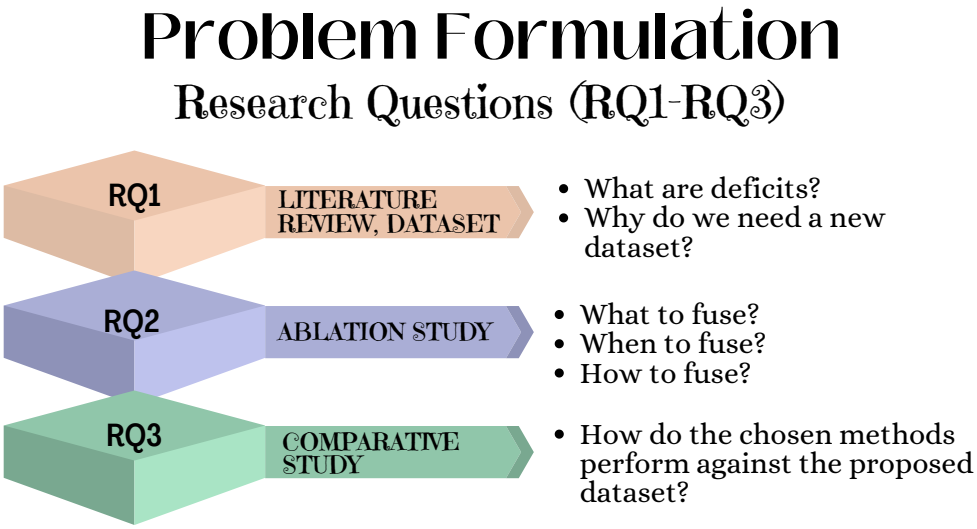
1.3 Problem Formulation

> *"The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill."*
>
> **Albert Einstein**

The current research proposes to create a dataset for the robot pouring task for granular media. We highlight the limitations of an existing collection of robot pouring datasets and present a new multimodal pouring dataset obtained from the on-board sensors of a service robot. We formulate anomaly detection as a machine learning problem in which methods are trained in a semi-supervised fashion using the proposed multimodal dataset.

As part of the research, the following research questions were answered:

- **RQ1** What are the deficits in the existing collection of robotic pouring datasets, and why do we need a new dataset?
- **RQ2** Which sensor fusion scheme works best for multimodal deep anomaly detection?
  - RQ2.1 Which sensor modalities should be fused for anomaly detection?
  - RQ2.2 When to fuse different types of sensor modalities for anomaly detection?
  - RQ2.3 How to fuse different types of sensor modalities for deep anomaly detection?
- **RQ3** How do the chosen anomaly detection methods perform on the proposed dataset?



2

Related Works

"If I have seen further than others, it is by standing upon the shoulders of giants."

Isaac Newton

In our literature review, we explore different aspects of anomaly detection methods, multimodal sensor fusion, and multimodal pouring datasets. We point out the limitations of the current dataset collection to justify a proposal for a new dataset.

2.1 Anomaly Type

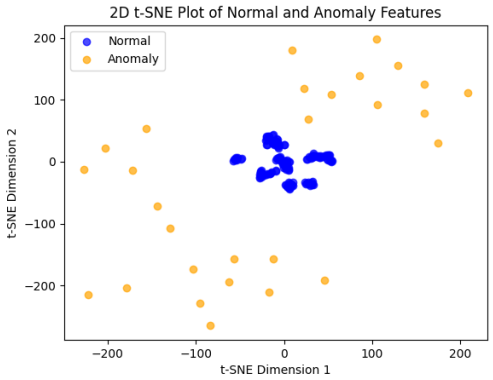


Figure 2.1: Blue points represent normal pouring samples. Orange points represent pouring anomalies such as collision of containers

Anomalies are categorized into three types: Point anomalies, Conditional anomalies, and Group anomalies. In the context of this thesis, the relevant anomaly type is point anomaly, and a brief explanation is provided below:

Point Anomalies: A point anomaly is the simplest type of anomaly where an individual data point exhibits an anomaly relative to the rest of the data [9]. Orange points in Figure 2.4 are point anomalies because they differ from normal data instances represented by blue points. Spam accounts on social media platforms, the occurrence of pathological abnormalities such as *chest lesions*[a] [7] and pouring anomalies are examples of Point anomalies.

[a] Lesion is an abnormal change or damage in the tissue due to a disease [7]

2.2 Classification of Methods

Anomaly detection methods fall into two categories: model-based and model-free methods [11]. Model-based techniques, such as the renowned Kalman filter, are employed for anomaly detection [13]. Model-based methods come with the major limitation that deriving equations for state transition and observation models is mathematically intensive and requires excellent mathematical and domain-specific knowledge [14]. Model-free methods address the limitations of model-based methods and are gaining popularity in recent times.

Model-free methods rely on pattern recognition techniques for anomaly detection [11]. Methods in this paradigm can be further classified into classical machine learning and deep learning methods. Below, we discuss prominent works in both classical and deep learning-based anomaly detection methods.

- **Classical Machine Learning:** [15] introduced an anomaly detection method in which they trained a Hidden Markov Model using haptic, audio, and kinematics data to identify anomalies related to indoor pushing activities, such as operating a switch or closing a microwave oven, as well as assisting physically challenged participants in important daily activities like feeding. The researchers trained the proposed method in a semi-supervised manner using only normal executions to identify unseen anomalous executions in the test dataset.

  The proposed method outperformed two other baseline methods and demonstrated superior performance in a multimodal configuration. The researchers employed a finite number of handcrafted features to represent high-dimensional raw sensor modalities. The feature engineering of hand-crafted features demands domain-specific knowledge, and the low-dimensional nature of handcrafted features may fail to capture complex anomalies [16].

- **Deep Learning:** [16] addressed the limitations of [15] by proposing an LSTM-based variational auto-encoder to identify anomalies. The autoencoder is trained on normal samples, minimizing reconstruction errors during training. During testing, it reconstructs normal samples with low reconstruction error and anomaly samples with high reconstruction error. The proposed method utilizes LSTM networks to model the temporal dependencies in sensor data. The paper solely focused on providing feeding tasks to participants with physical disabilities and proposed a large dataset containing normal and anomalous samples of feeding activity. The method exhibited superior performance compared to classical and deep learning methods and also showcased enhanced performance with an increase in the number of sensor modalities.

  In a similar work, [11] proposed a fully supervised method to identify anomalies related to tabletop manipulations, including pouring activity. Fully supervised anomaly detection methods rely heavily on the availability of large and balanced labeled datasets [12]. However, acquiring large datasets is a costly and labor-intensive process, particularly due to the tedious nature of data annotation [17]. Compounding this challenge is the inherent rarity of anomalies, making it difficult to create a balanced dataset [12]. Models trained on imbalanced datasets may incorrectly report anomalies [12]. Therefore, semi-supervised methods are preferred over fully-supervised methods [12].

Deep learning methods come with a lot of advantages in the context of anomaly detection. First, they can seamlessly integrate multiple sensor modalities to identify complex anomalies; they can process different types of data, ranging from graphs and can learn complex observation models over high-dimensional data. On the other hand, they have shortcomings. Training deep learning methods is time-intensive and computationally expensive, and their performance depends on the availability of a large dataset.

Both classical and deep learning methods have their advantages and disadvantages. Classical methods exhibit less dependence on large datasets, while deep learning methods provide a hierarchy of features. Leveraging this concept, [18] proposed a method to identify industrial anomalies using high-dimensional multimodal data, including images and point clouds. Domain-specific transformer-based models were employed to extract high-dimensional features to train two one-class support vector machines. One of the trained models classifies the input as normal and anomaly samples, and the other model localizes anomalies in their respective 2D images. We draw inspiration from this paper to use domain-specific deep learning-based models to extract feature vectors and employ a one-class support vector machine to identify anomalies.

Anomaly detection methods are classified as uni-modal and multi-modal based on the variety of sensors used. The popularity of multi-modal methods stems from their ability to overcome limitations inherent in uni-modal methods in several ways [12]. Within a multi-modal configuration, each sensor modality captures a unique aspect of anomalies, contributing to a comprehensive understanding of anomalies [12]. Using the pouring activity as an example, it is depicted in Figure 2.2 that relying solely on the audio modality is inadequate for distinguishing between normal and anomalous samples. Nonetheless, the fusion of audio, video, and depth modalities effectively differentiates between normal and anomaly samples, as depicted in Figure 2.4. All the prominent works mentioned so far have showcased the superior performance of a multimodal configuration over a uni-modal one through ablation studies.

In the context of multimodal sensor fusion, three crucial research questions must be addressed, and the answers to each question are as follows:

- **What to Fuse?** The selection of sensor modalities is contingent on the specific problem or application. For example, the target intrusion system proposed by [8] employs infrared images for monitoring enemies in restricted military areas. Radiographs from X-rays are utilized to detect chest anomalies such as lesions [7]. In the case of robot manipulation activities, [16] proposed to use five different sensors, namely a camera, microphone, current and force sensor, and joint encoder signals, to detect anomalies related to feeding tasks. [11] proposes to fuse audio, depth and video sensor modalities to detect pouring anomlaies. In the context of pouring activity, human experts seamlessly integrate visual observation, auditory cues, and tactile feedback from hand grip information to achieve precise results [19]. This underscores the necessity of incorporating a spectrum of sensor modalities, including vision, depth, audio, and tactile sensors, for robots to attain robust and accurate pouring capabilities.

- **When to Fuse?** Deep neural networks organize features hierarchically, providing a diverse range

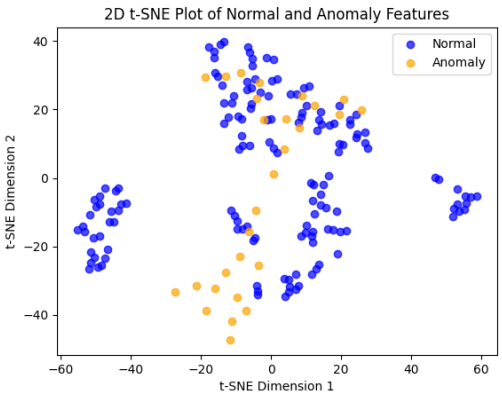


Figure 2.2: *Audio features representing normal and anomaly pouring samples*

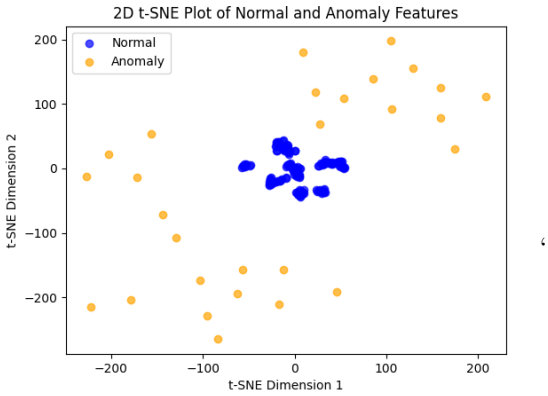


Figure 2.3: *Feature space derived from the fusion of audio, video, and depth modalities represents normal and anomaly pouring samples*

of options to combine sensing modalities at early, middle, or late stages [20]. Based on this, [21] presented three sensor fusion schemes, and a brief explanation of each is provided below:

- **Early/Data Fusion:** The early fusion method is suitable for sensor modalities with the same sampling rate. [22] proposed a method that employs the early fusion scheme to integrate multiple 1-D time series data for identifying anomalies, especially in applications related to autonomous vehicles and industrial machinery. The study fused three-channel GPS and three-channel LIDAR data to detect six different types of point anomalies related to autonomous driving.

  With regards to pouring activity, only the depth and RGB data streams from one sensor have an equal sampling rate, and both modalities can pass through a single domain-specific network for feature extraction. The early fusion method is not suitable for pouring activity, as the activity is captured using audio and haptic data along with RGB-D data. In cases where there is significant noise in audio and haptic data, or when audio and haptic data are not available, the early fusion method is suitable for feature extraction from the RGB-D modality, providing the added benefit of reduced memory requirements and minimized information loss [22].

- **Late Fusion:** The late fusion scheme is a popular choice for fusing heterogeneous modalities. For instance, [2] has proposed a method to detect and classify robot manipulation anomalies by employing late fusion to integrate audio, vision, and haptic data. Notably, all the prominent multimodal anomaly detection methods mentioned earlier, with the exception of [11], utilize the late sensor fusion scheme.

  Unlike early fusion, late fusion fuses can seamlessly integrate new sensor types without impacting existing domain networks. The modular and flexible nature of the late fusion scheme is the reason why late fusion is more suitable for audio, video, and depth modalities than early fusion.

  - **Middle Fusion:** [2] has proposed a middle-fusion scheme for detecting robot manipulation anomalies. This sensor fusion scheme employs early fusion for the video and depth modalities and late fusion for audio and vision features. Middle fusion combines both early and late fusion methods, allowing the integration of heterogeneous sensor modalities, similar to late fusion. Moreover, it reduces memory requirements by employing early fusion of sensor modalities with equal sampling rates. The suitability of this sensor fusion scheme over others is contingent on the availability of pretrained models.

Early, middle, and late fusion methods represent primitive sensor fusion schemes with inherent advantages and disadvantages [20]. The suitability of sensor fusion schemes relies on the nature of sensor modalities and network architecture and can be determined through experimentation [20].

- **How to Fuse?** Prominent mathematical operations for feature fusion include Addition/Average, Concatenation, Ensembles, and Mixtures of Experts. Among these, Concatenation and Mixture of Experts are commonly used in late fusion, and we will specifically discuss these two operations.

  - **Concatenation:** Concatenation operations involve stacking features from fully connected layers along one dimension. It is a common and straightforward operation extensively used in the multimodal anomaly detection literature. Due to the ease of implementation, concatenation is utilized in all the prominent studies mentioned earlier. Concatenation assigns equal weight to all modalities without considering sensor modality noise levels. Mixture of Experts is employed to address this limitation [23].
  - **Mixture of Experts:** Introduced by [23], Mixture of Experts computes the weighted sum of features as follows:

$$\text{Output} = \sum_{i=1}^{n} w_i \cdot \text{Expert}_i(\mathbf{X}_i), \quad \sum_{i=1}^{n} w_i = 1$$

    where w_i represents the weight assigned to each modality, and $\text{Expert}_i(\mathbf{X}_i)$ is the output of the domain-specific network (expert) for the i-th modality. The gating network assigns weights to each sensor modality, and the resulting weighted sum of features serves as the input to the prediction network. To the best of our knowledge, there are no studies on multimodal anomaly detection methods that employ this operation for the fusion of heterogeneous modalities.

Concatenation and Mixture of Experts [23] play pivotal roles in multimodal anomaly detection, particularly with heterogeneous sensor modalities such as audio, video, depth, and haptic data.

2.3 Pouring Datasets: Review

- **Dataset #1:** [3] proposed is a multimodal dataset for liquid weight estimation, classification of liquids and containers, and overflow detection. All samples, excluding overflow instances, can be labeled as normal executions, whereas overflow samples can be labeled as anomalies. It is important to note that the dataset was not specifically collected for anomaly detection. The manual segregation of samples into normal and anomaly categories is a time-consuming process, and the test data would

only contain one type of pouring anomaly. Consequently, while this dataset may not be considered complete, it can still be utilized to assess the generalization of methods trained on different pouring datasets.

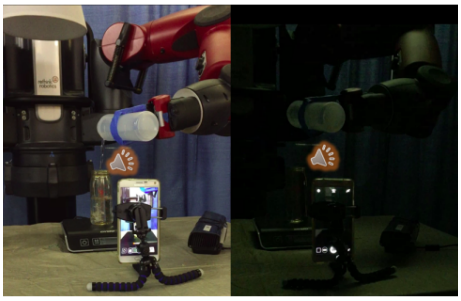


Figure 2.4: Smartphone is used for recording video modality [3]

The dataset exclusively includes audio and video modalities using a smartphone and a Kinect sensor. The audio modality was recorded using a microphone array, and the video modality was recorded using a smartphone from the observer's point of view. The placement of the microphone renders the setup instrumented and contradicts human perception. A single Kinect sensor could have been used for recording all three modalities—audio, depth, and video—instead of using a smartphone. The use of a smartphone is only recommended when the robot is not equipped with a suitable sensor for recording video.

Additionally, the video frames were cropped to hide the movement of the robotic arm performing the pouring activity. Occlusion is a common disturbance and should be considered in the dataset to test if the method can still identify anomalies using occluded video frames and other sensor modalities. The audio modality contains background noise from the setup and lacks natural background noises such as human interaction and room acoustics.

- **Dataset #2:** [19] proposed is a multimodal pouring dataset for estimating pouring rate and pouring height. The pouring execution was recorded using stereo, video, depth, audio, and the weight of the pouring container. The dataset was collected in a quiet and highly instrumented environment, featuring microphones placed next to the receiver and a camera positioned directly above it.

  The dataset considered granular pouring over liquid pouring with different physical attributes. It also contains different pairs of containers with various colors and transparency levels. A human expert carried out the pouring activity. Although the dataset was not originally intended for anomaly detection, all the samples can be labeled as normal and utilized as a training dataset for semi-supervised methods.

- **Dataset #3:** [11] proposed a multi-modal dataset that encompasses executions of five distinct tabletop manipulations, including pouring. The number of samples for each activity is quite limited, with no more than 25 normal samples in each category. While there is an adequate number of anomalous samples for every activity, the size of the training data with normal samples for a given activity is very limited. Consequently, the dataset as a whole is better suited as a test dataset for methods trained on different pouring datasets.

  It is important to note that the dataset only provides a top-camera view of activities (RGB + Depth). The side and top camera views are equally important for a comprehensive understanding of

pouring activity, closely resembling human perception. Additionally, we assert that manipulation activities bear a close resemblance to toy demonstrations, thus rendering the dataset suitable only for proof-of-concept.

The mentioned datasets are not without limitations, but they offer useful information to propose a new multimodal pouring dataset for anomaly detection with as few limitations as possible.

2.4 Pouring Datasets: Limitations

In the following section, we present the limitations of the existing collection of pouring datasets.

Generalization: Anomaly detection methods for pouring activity should generalize to different types of containers and pouring substances [4]. Containers come in diverse shapes, sizes, transparency levels, and colors, while pouring substances can span a spectrum from granular materials to liquids to semi-solids. The dataset proposed by [11] for anomaly detection contains pouring samples with a fixed pair of container and substance. Methods trained on this dataset may not generalize to unseen containers and substances. Therefore, the dataset should encompass pouring samples involving both transparent and opaque containers, along with granular substances with varied densities, shapes, sizes, and colors.

Quiet Environment: Human experts closely observe the change in pitch while a container is being filled, making the audio modality an important clue. In the existing collection of datasets, the audio modality was recorded when the robot was operated in a calm and noise-free environment. Service robots are often deployed in busy and noisy areas, such as kitchens. To incorporate authentic ambient sounds, it is necessary to capture the audio modality using robot sensors deployed in densely populated environments.

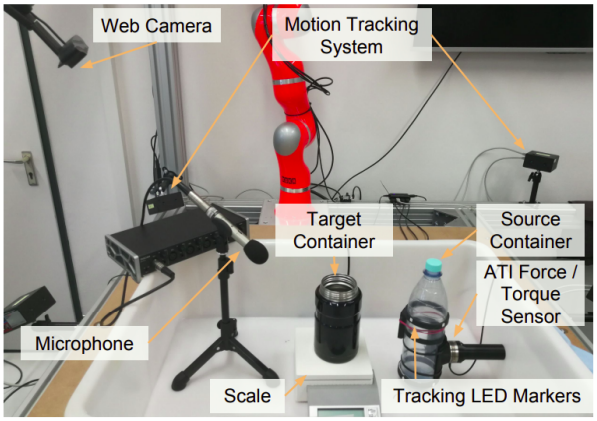


Figure 2.5: Highly instrumented setup proposed for recording multi-modal pouring dataset [4]

Highly Instrumented Environment: The on-board sensors of robots enable them to perceive their surroundings similarly to how humans perceive their environment. In the existing collection of datasets, microphones are positioned adjacent to the receiving containers, as shown in Figure 2.5. The setup is highly instrumented, contradicts human perception, and may not be suitable for all public applications.

Demonstrator: Haptic sensor data plays a vital role in capturing the interaction between the pouring container and the substance being poured [24]. This type of data serves as a crucial information source for the robot, especially in scenarios with significant noise in vision and audio data. It can also alert the robot to situations where the pouring container might fall from its grip. However, multimodal pouring datasets with human demonstrations often lack haptic data, making them suitable only for benchmark case studies involving video and audio data. Additionally, the question of whether machine learning models trained on human demonstrations can generalize effectively to robot demonstrations remains an open research challenge.

3

Methodology

This chapter outlines the proposed multimodal pouring dataset and the methodology for implementing pouring anomaly detection. The methodology addresses feature extraction from individual sensor modalities, the choice of feature extractors for each sensor modality, provides concise descriptions of trainable modules, discusses sampling techniques, and outlines the metrics employed for evaluations.

3.1 Data Collection

We introduce a novel multi-modal **g**ranular **po**uring **d**ataset, **GPOD**, captured using the Toyota Human Support Robot, HSR [25]. The HSR robot, shown in Figure 3.1, is equipped with a variety of sensors, including an RGB-D camera, microphone, robotic arm for interacting with objects in the surrounding environment, LIDAR sensors, and more. We positioned an external HyperX microphone right next to the camera on the robotic platform to mimic human perception. The HyperX microphone has different polar patterns, and the cardioid polar pattern was selected to predominantly capture the sound in front of the robot.

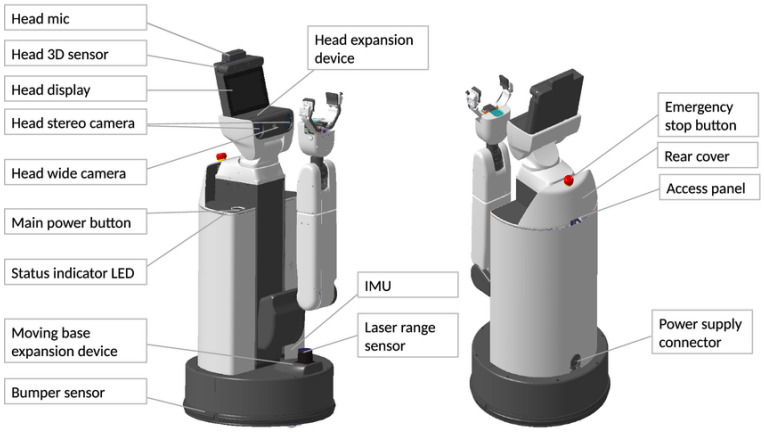


Figure 3.1: Toyota Human Support Robot. Image taken from [5]

The proposed dataset aims to overcome limitations identified in the previous chapter regarding pouring datasets. Various granular pouring datasets, including corn, rice, and coffee beans, were considered. Corn was chosen for the training data, while coffee beans and rice were selected for the test data.

Different pairs of pouring and receiving containers were explored, featuring varying attributes such as shape, size, color, and transparency. The pouring substances and containers used in the training data were not repeated in the test data. Variations in the grasping height of the pouring container were introduced, identifying three levels proportional to its height, and data samples were collected at different grasping heights. Similarly, three levels were identified for the receiving container, and the granular substance was poured at different levels. Following a goal-oriented approach, the samples were labeled, as in [11].

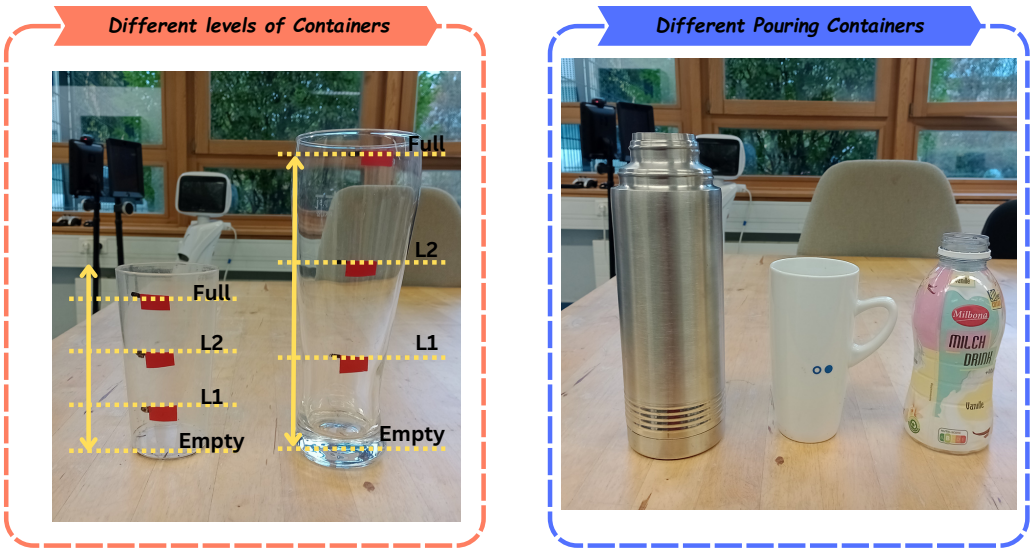


Figure 3.2: Three levels based on the height of containers

The training dataset comprises 110 normal pouring samples, while the test dataset consists of 72 pouring samples. Among these, 38 pouring samples belong to the normal category, and the remaining 34 samples belong to the anomaly category. Three anomaly types were considered and are illustrated in Figure 3.3: the first involves removing the receiving container in the middle of the pouring activity; the second involves the collision of the robotic hand or pouring container with the receiving container; the last refers to the overflow of the granular pouring substance.

Data collection, conducted at various times of the day, introduced variations in lighting conditions. Audio modalities feature natural background noise, including human conversations. Vision modalities encompass samples where the receiving container is significantly occluded by the robotic arm performing the pouring task. The substantial background noise in sensor modalities enables the experimentation of "Mixture of Expert" sensor fusion schemes.

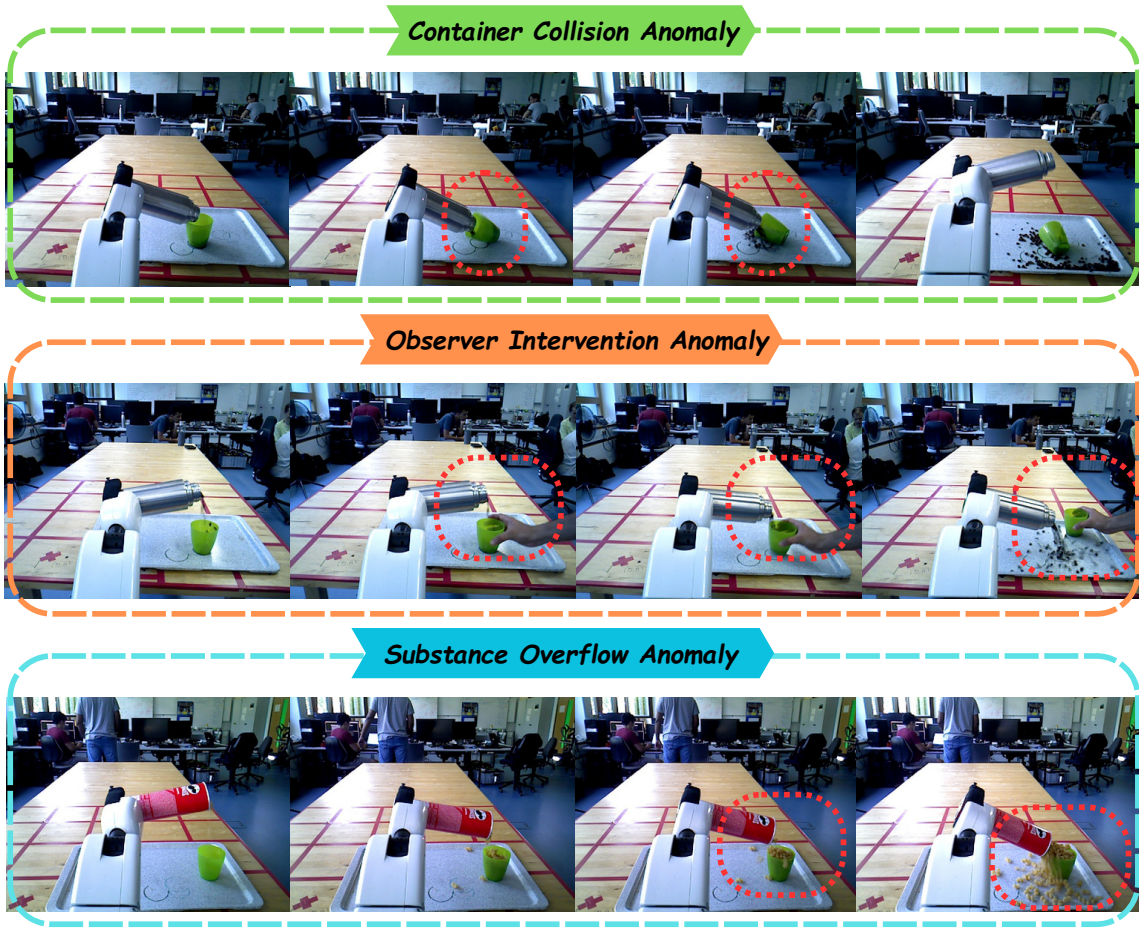


Figure 3.3: *Illustrations of various types of pouring anomalies present in the test dataset*

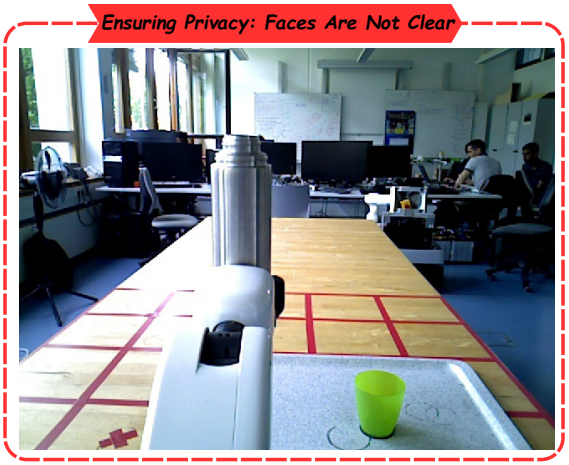


Figure 3.4: *Illustrations of various types of pouring anomalies present in the test dataset*

Although human conversations are recorded in the audio modality, the conversations are not clear and only serve as background noise. Care was taken to safeguard the privacy of individuals in the dataset. The HSR robot executes both successful and unsuccessful granular pouring tasks with various granular substances and containers, recording synchronized RGB-D, audio, and relevant metadata information in ROS bag files. The dataset was recorded using data collection software, Referee-Box [a], developed by the research faculty of the university. The software records the ROS Bag along with its metadata, and processed data containing audio, depth, and video modalities along with metadata are made available to the research community.

[a] https://github.com/HEART-MET/metrics_refbox

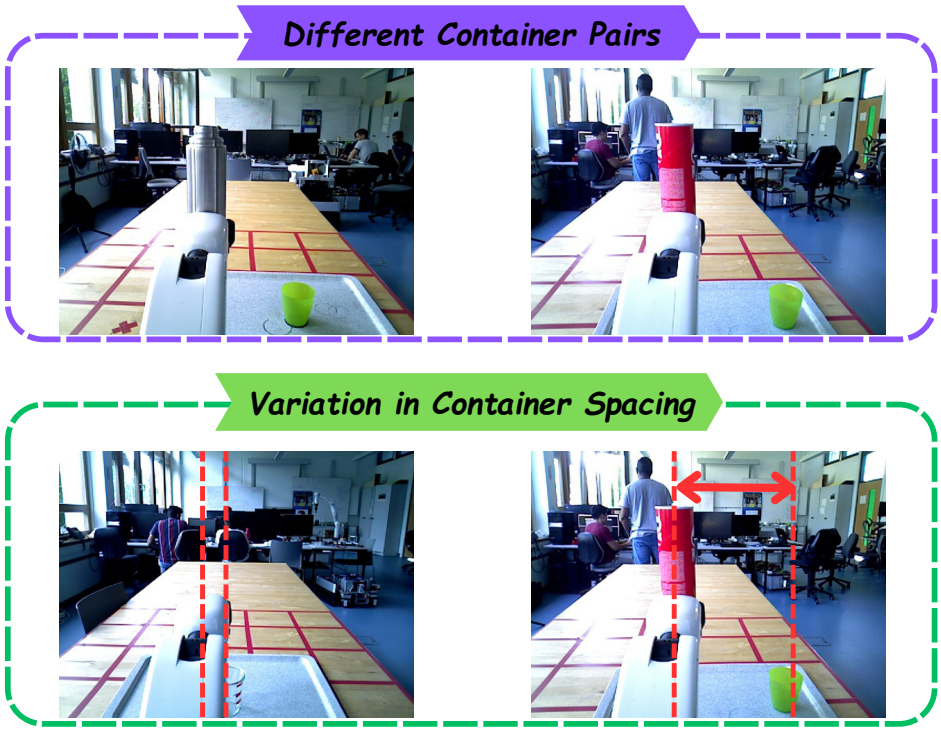


Figure 3.5: *Different pairs of containers were used for data collection, and the distance between containers was also varied*

3.2 Feature Extraction

We focus on three sensor modalities, namely, audio, video, and depth. The following section explains the process of feature extraction from each of these modalities using pre-trained deep learning models.

Video Feature Extraction

The video and depth modalities are decomposed into individual frames, where each video sample is characterized by a defined start time, t_{start}, and end time, t_{end}, indicating the time window during which the pouring activity took place. Frames captured within this time window are uniformly sampled at a rate (R) given by $R = \frac{1}{N-1}(t_{\text{end}} - t_{\text{start}})$, where N is a user-defined number of frames, always a multiple of 8. The resulting sequence of frames, denoted as $\{F_{t_i}\}$, collectively forms the representation of the video sample. Here, $t_i = t_{\text{start}} + i \times R$ represents the time associated with each frame, and i is the index of the sampled frame. Furthermore, each frame within a sampled sequence undergoes a inference transformation as recommended by pytorch.org [1]. The resulting sequence is then used as input for the ResNet(2+1)D

[1] https://pytorch.org/vision/main/models/generated/torchvision.models.video.r2plus1d_18.*html*

architecture [26], which outputs a vector $\mathbf{v} \in \mathbb{R}^{512}$.

Table 3.1: *Inference Transforms.*

Operation	Description
Resize	Resize video frames to the size [128, 171] using Bilinear interpolation
Central Crop	[112, 112]
Rescale	Normalize the pixel values in the range [0.0, 1.0]
Normalize	Mean = [0.43216, 0.394666, 0.37645] Standard Deviation = [0.22803, 0.22145, 0.216989]
Permute Dimensions	Input (B, T, C, H, W) → Output (..., C, T, H, W) C: Number of Channels, T: Number of Frames

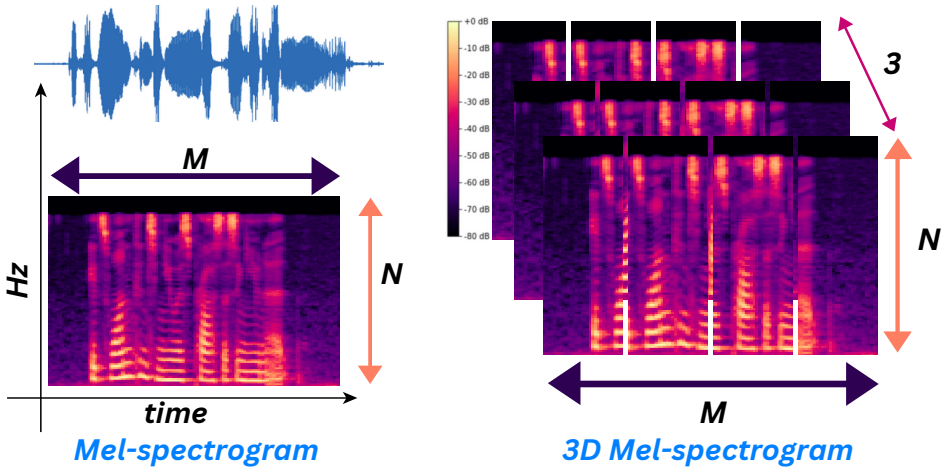


Figure 3.6: *Mel-spectrogram is a 2D matrix, and a 3D mel-spectrogram is a 3D array with dimensions M×N×3, where the third dimension corresponds to three channels*

Audio Feature Extraction

Mel-spectrograms represent the frequency content of an audio signal over time as a single-channel image [27]. Given that human audio perception demonstrates a high sensitivity to various frequencies, this method of representing audio signals brings digital audio perception several steps closer to human audio perception [28]. As a result, mel-spectrograms are widely employed across various applications, ranging from speech recognition [27] to multi-modal anomaly detection [29].

Mel-spectrograms are suitable input for neural network architectures designed to process single-channel images. However, numerous pretrained CNN models are designed to process RGB images or 3D arrays. To extract features from these popular pretrained models, the visual representation of audio signals as 3D arrays is essential. This three-dimensional representation is achieved by initially extracting two-dimensional

mel-spectrograms from audio signals and then stacking the same mel-spectrogram three times to form a 3D array, as illustrated in Figure 3.6. The 3D array is used as input to ResNet18, which outputs a vector of dimensions $\mathbf{a} \in \mathbb{R}^{512}$.

Depth Feature Extraction

Depth images are monochromatic 2D arrays of pixels D, where the value $D(i, j)$ stored in each pixel (i, j) represents the distance of a point in the scene from the camera. The Toyota HSR robot is equipped with an RGB-D camera with a least count of 300 mm and a maximum measurable distance of 9500 mm. All values in the range of 0-300 mm are clipped to 0 mm, and any value greater than 9500 mm is clipped to 9500 mm.

Depth feature extraction follows a process analogous to video feature extraction, wherein depth frames within the time window of the start and end pour time are uniformly sampled. Each depth image in a sampled sequence is stacked to create a 3D array, mirroring the approach used for mel-spectrograms. The resulting mini-batch of depth frames serves as input to the ResNet18 pretrained model for feature extraction.

3.3 Feature Extractors

Our method draws inspiration from the late fusion method introduced in [5] for multi-modal anomaly detection. We employ two pre-trained feature extractors: ResNet 18 for audio and two inpedendent branches of ResNet (2+1)D for video and depth modality, respectively.

ResNet(2+1)D: The ResNet (2+1)D architecture uses (2+1)D convolutions instead of 3D convolutions to process the video modality. (2+1)D convolutions, introduced by [26], refer to 2D spatial convolutions followed by 1D temporal convolutions. (2+1)D convolution offers two advantages over 3D convolutions. With the same number of parameters, adding an extra ReLU activation function between 2D and 1D convolutions increases the number of non-linearities. This, in turn, improves the approximation of nonlinear functions as the number of non-linear activation functions in the neural network architecture increases.

The selection of ResNet18 as the feature extractor for depth modality is arbitrary and is solely employed to illustrate the late fusion of three modalities.

Algorithm 1 Multimodal Late Fusion for Pouring Anomaly Detection

Input:
List of audio signals $\{A^{(1)}(t), A^{(2)}(t), ..., A^{(M)}(t)\}$
List of video frames sets $\{V^{(1)}_{1,..,T}, V^{(2)}_{1,..,T}, ..., V^{(M)}_{1,..,T}\}$
List of depth frames sets $\{D^{(1)}_{1,..,T}, D^{(2)}_{1,..,T}, ..., D^{(M)}_{1,..,T}\}$
ResNet18 pretrained model for audio feature extraction ϕ_a
ResNet(2+1)D pretrained network for video feature extraction ϕ_v
ResNet model for depth feature extraction ϕ_d
Start time t_{start}, End time t_{end}
Number of frames $N = 64$
Number of coreset samples l
Output:
$\mathcal{D}_M \leftarrow \{\mathbf{a}^{(1)}, \mathbf{v}^{(1)}, \mathbf{d}^{(1)}, \mathbf{a}^{(2)}, \mathbf{v}^{(2)}, \mathbf{d}^{(2)}, ..., \mathbf{a}^{(M)}, \mathbf{v}^{(M)}, \mathbf{d}^{(M)}\}$
Procedure:
for m from 1 to M **do** ▷ Loop over instances
    Step 1: `Extract Audio Features` for Instance m
        $A^{(m)}_{\text{clipped}} \leftarrow \texttt{ClipAudio}(A^{(m)}(t), t_{\text{start}}, t_{\text{end}})$
        $S^{(m)} \leftarrow \texttt{ComputeMelSpectrograms}(A^{(m)}_{\text{clipped}})$
        $S^{(m)}_{3D} \leftarrow \texttt{Stack}(S^{(m)}, S^{(m)}, S^{(m)})$
        $\mathbf{a}^{(m)} \leftarrow \phi_a(S^{(m)}_{3D})$ ▷ Output: $\mathbf{a}^{(m)} \in \mathbb{R}^{512}$
    Step 2: `Extract Video Features` for Instance m
        $\{F^{(m)}_t\} \leftarrow \{F^{(m)}_t \in V^{(m)}_{1,..,T} \,|\, t_{\text{start}} \leq t \leq t_{\text{end}}\}$
        $\{F^{(m)}_{t_i}\} \leftarrow \texttt{UniformSampling}(\{F^{(m)}_t\}, N)$
        $\{F^{(m)}_{t_i}\} \leftarrow \texttt{transform}(\{F^{(m)}_{t_i}\})$
        $\mathbf{v}^{(m)} \leftarrow \phi_v(\{F^{(m)}_{t_i}\})$ ▷ Output: $\mathbf{v}^{(m)} \in \mathbb{R}^{512}$
    Step 3: `Extract Depth Features` for Instance m
        $\{D^{(m)}_t\} \leftarrow \{\texttt{ClipDistance}(D^{(m)}_t) \,|\, \forall D^{(m)}_t \in \{D^{(m)}_t\}\}$
        $\{D^{(m)}_t\} \leftarrow \{D^{(m)}_t \in D^{(m)}_{1,..,T} \,|\, t_{\text{start}} \leq t \leq t_{\text{end}}\}$
        $\{D^{(m)}_{t_i}\} \leftarrow \texttt{UniformSampling}(\{D^{(m)}_t\}, N)$
        $\{D^{(m)}_{t_i}\} \leftarrow \{\texttt{Stack}(D^{(m)}_{t_i}, D^{(m)}_{t_i}, D^{(m)}_{t_i}) \,|\, \forall D^{(m)}_{t_i} \in \{D^{(m)}_{t_i}\}\}$
        $\mathbf{d}^{(m)} \leftarrow \phi_d(\{D^{(m)}_{t_i}\})$ ▷ Output: $\mathbf{d}^{(m)} \in \mathbb{R}^{512}$
end for
Sensor Fusion: Training set features
    $\mathcal{D}_M \leftarrow \{\mathbf{a}^{(1)}, \mathbf{v}^{(1)}, \mathbf{d}^{(1)}, \mathbf{a}^{(2)}, \mathbf{v}^{(2)}, \mathbf{d}^{(2)}, ..., \mathbf{a}^{(M)}, \mathbf{v}^{(M)}, \mathbf{d}^{(M)}\}$
    $\mathcal{D}_{\text{coreset}} \leftarrow \texttt{CoresetSampling}(\mathcal{D}_M, l)$ ▷ Greedy Coreset Sampling

3.4 Trainable Modules

Inspired by [18], we train One-Class Support Vector Machine, OC-SVM, using high-level features extracted from audio, video, and depth modalities. OC-SVM, introduced by [30], is a variation of the infamous support vector machine, tailored for semi-supervised anomaly detection. During training, OS-SVM maps normal features into a feature space and maximizes the distance between normal samples and the origin. During testing, the distance between the input sample and the origin is used as an anomaly score. Based on the user-defined threshold for the anomaly score, the input is classified as normal or anomalous.

In recent times, Isolation Forest introduced by [31], is quite popular for semi-supervised anomaly detection along with OC-SVM and was chosen for the comparative study in this research work. Inspired by [32], we employ the greedy-coreset sampling method to train models on selective features, aiming to achieve comparable results to models trained on the entire dataset. The pseudo-code algorithm 1 explains feature extraction from three modalities and the late fusion sensor scheme.

4

Results

"If you churn the ocean of life, you are responsible for nectar as well as poison."

Bhagavata Purana

In this chapter, three experiments were conducted to investigate the contribution of various sensor modalities in anomaly detection through ablation study, compare two anomaly detection methods, and explore the significance of sampling techniques in reducing the number of samples for model training.

4.1 Ablation Study

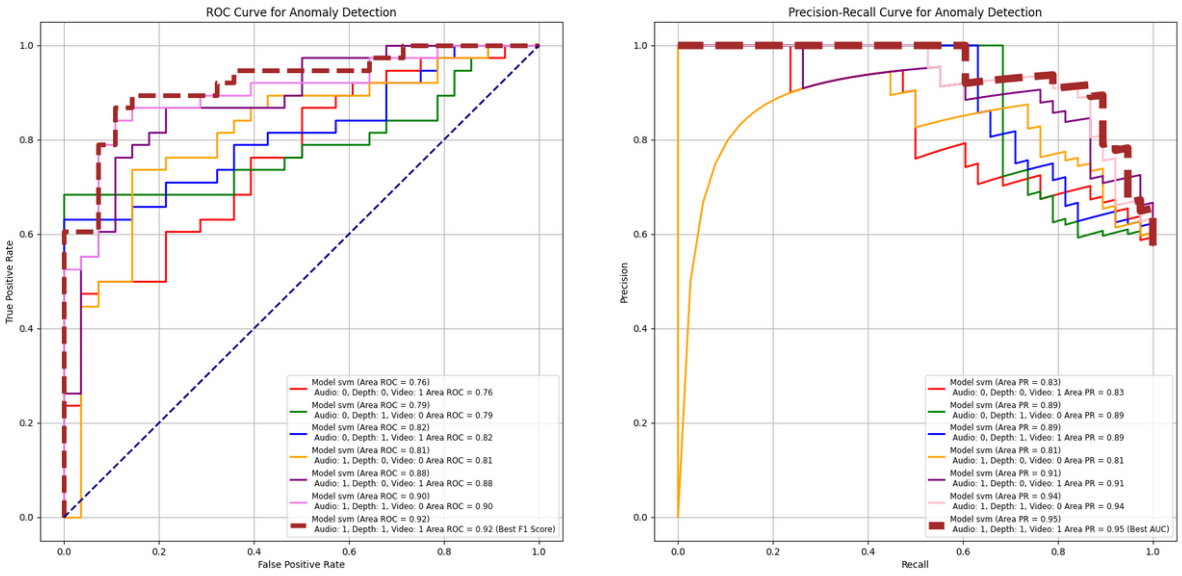


Figure 4.1: Receiver Operating Characteristic (ROC) Curve (left image) and Precision-Recall (PR) (right image) curve for OC-SVM Anomaly Detection in various sensor configurations

- **Objective:** The aim of this experiment is to study the contribution of audio, video, and depth modalities in anomaly detection.
- **Terminologies:** We trained OC-SVM and Isolation Forest using three modalities and obtained different models based on various combinations of sensor modalities. For example, unimodal configurations involve training models on just one modality, such as OC-SVM-Audio. Bimodal configuration involves training models on any two of three sensor modalities, for example, IF-Audio-Video, and trimodal configuration involves training models on all three modalities, such as OC-SVM-Audio-Depth-Video.
- **Hypothesis:** Our hypothesis asserts that the incremental addition of sensor modalities enhances the performance of anomaly detection. Given that vision (video + depth) is the most crucial modality for human experts in achieving pouring tasks, we specifically anticipate that the unimodal video configuration will outperform other unimodal configurations. Furthermore, we expect that a bimodal configuration incorporating both video and depth modalities will outperform other bimodal configurations.
- **Observations:** The plots clearly illustrate that OC-SVM exhibits incremental performance with the addition of a new sensor modality. The trimodal configuration outperformed both the bimodal and unimodal configurations. In the unimodal configuration, the audio modality demonstrated superior performance compared to the video and depth modalities. Despite the higher amount of noise in the audio modality compared to video and depth, the model performs exceptionally well with the audio modality in the unimodal configuration.

Configuration	**ROC**	**PR**	**F1-score**
OC-SVM-Audio	76%	83%	80.95%
OC-SVM-Depth	79%	89%	81.25%
OC-SVM-Video	82%	89%	77.78%
OC-SVM-Audio-Depth	81%	81%	88%
OC-SVM-Depth-Video	88%	91%	77.42%
OC-SVM-Audio-Video	90%	94%	85.71%
OC-SVM-Audio-Depth-Video	**92%**	**95%**	**89.47%**

Table 4.1: *Area under the ROC curve, area under the PR curve, and F1-score values for One-Class SVMs in different configurations*

Pouring anomalies are very well captured in the video modality compared to depth and audio, and with relatively less noise in the data, the performance of the model is quite contrary to our hypothesis in unimodal and bimodal configurations. It is worth noting that the unimodal audio configuration shows performance comparable to the bimodal video and depth configurations. The bimodal configuration of audio and depth performs the best, second only to the trimodal configuration.

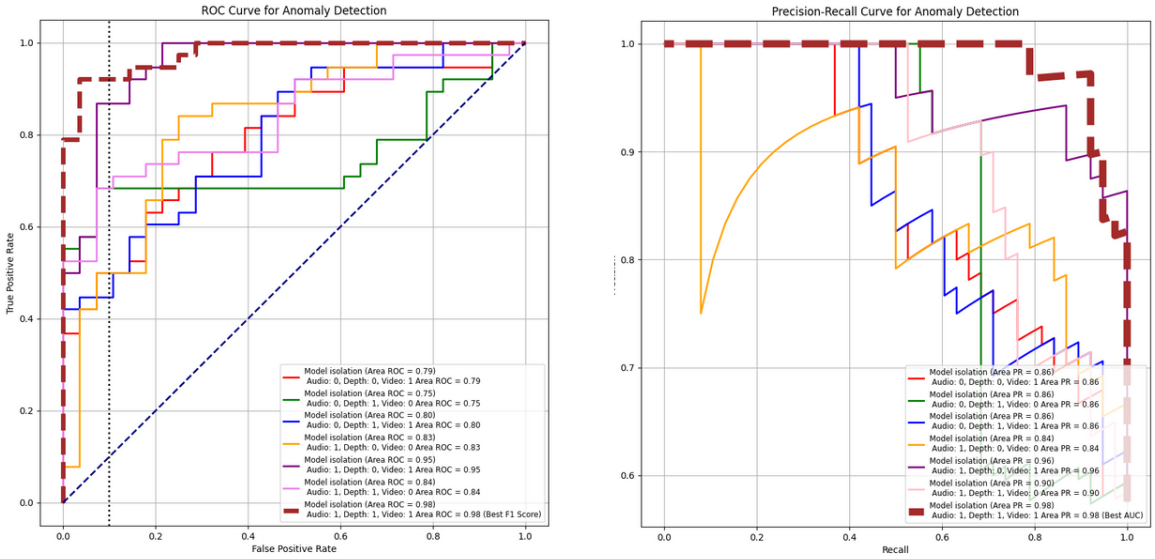


Figure 4.2: Receiver Operating Characteristic (ROC) Curve (left image) and Precision-Recall (PR) (right image) curve for Isolation Forest Anomaly Detection in various sensor configurations

Configuration	ROC	PR	F1-score
IF-Audio	83%	84%	83.12%
IF-Depth	75%	86%	78.79%
IF-Video	79%	86%	79.12%
IF-Audio-Depth	84%	90%	80.46
IF-Depth-Video	80%	86%	80.90
IF-Audio-Video	95%	96%	92.68%
IF-Audio-Depth-Video	**98%**	**98%**	**94.59%**

Table 4.2: *Area under the ROC curve, area under the PR curve, and F1-score values for Isolation Forests in different configurations*

The tri-modal configuration demonstrates superior performance with Isolation Forest, similar to OC-SVM. While the unimodal video configuration performs better than the unimodal depth configuration, the unimodal audio configuration exhibits the most superior performance. Contrary to OC-SVM results, the unimodal audio configuration even surpasses the bimodal configuration of video and depth modalities. Furthermore, in contrast to OC-SVM, the bimodal configuration of audio and video shows superior performance, second only to the tri-modal configuration.

4.2 Comparative Study

- **Objective:** The aim of this experiment is to compare the statistical performance of OC-SVM and Isolation Forest.

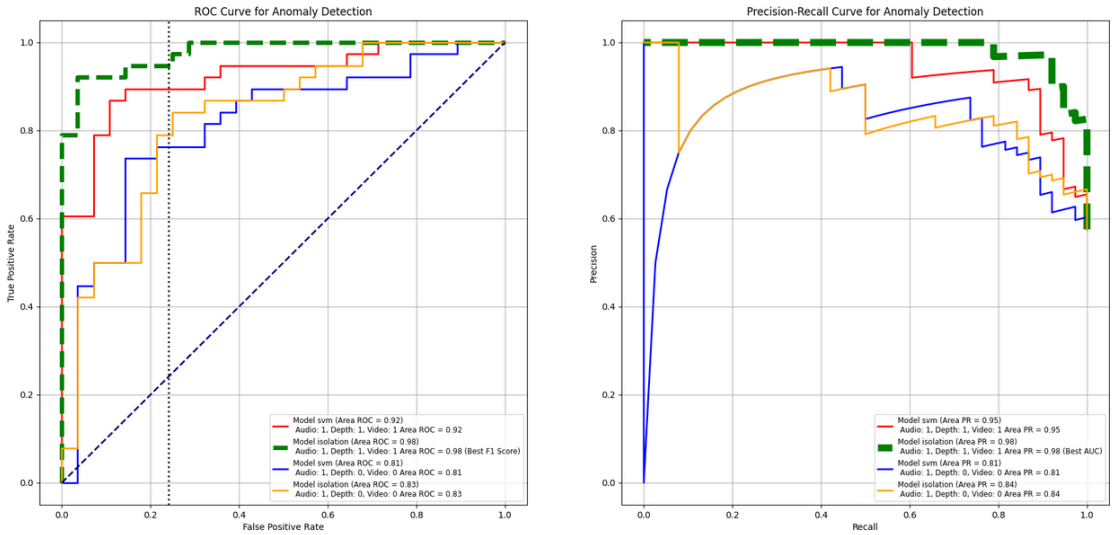


Figure 4.3: Statistical comparison of OC-SVM and Isolation Forest in unimodal audio and trimodal configuration

Configuration	ROC	PR	F1-score
SVM-Audio	81%	84%	80.95%
IF-Audio	**83%**	**84%**	**83.12%**
SVM-Audio-Depth-Video	92%	98%	89.47%
IF-Audio-Depth-Video	**98%**	**98%**	**94.59%**

Table 4.3: *Area under the ROC curve, area under the PR curve, and F1-score values for Isolation Forests and OC-SVM in unimodal audio and trimodal configurations*

- **Observations:** The plot clearly illustrates the superior performance of Isolation Forest in unimodal audio and video configurations, bimodal audio and video configuration, and trimodal configurations compared to OC-SVM. It is important to note that the bimodal audio and video configuration outperforms the trimodal configuration of OC-SVM and is comparable to the trimodal configuration of Isolation Forest.

- **Verdict:** Isolation Forest outperforms OC-SVM in different sensor modality configurations on our dataset. As mentioned in [33], the time complexity to train Isolation Forest is lower compared to the time complexity required to train OC-SVM. The reduced time complexity and demonstrated

statistical performance on our datasets, along with three other unimodal datasets mentioned in [33], clearly show the superior nature of Isolation Forest over OC-SVM.

4.3 Sampling Methods

- **Objective:** The aim of this experiment is to compare coreset-greedy sampling and uniform sampling techniques, emphasizing the importance of sampling techniques in improving model performance with a reduced number of features.
- **Hypothesis:** We assert that the model performance with a reduced number of features obtained from core-set sampling is comparable to the model trained on the entire feature set.

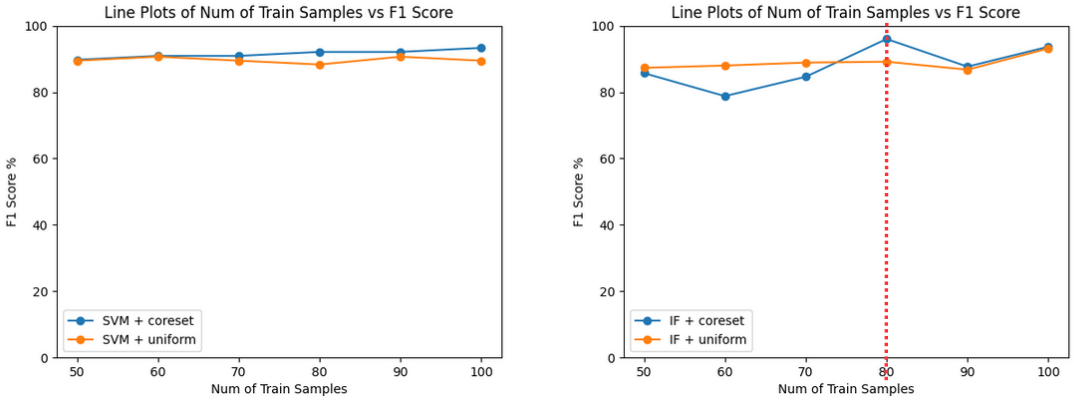


Figure 4.4: Comparison of the performance of models trained on coreset-sampled features and uniformly sampled features

- **Observations:** In the case of OC-SVM, the model's performance trained on sampled features from coreset sampling is superior to that of uniformly sampled features. However, the performance of OC-SVM is still lower when the model is trained on the full feature set.

  In the case of Isolation Forest, it is uncertain whether coreset-sampled features exhibit superior performance over uniformly sampled features. Nevertheless, it is important to note that the isolation forest exhibits a very high F1 score with 80 coreset-sampled features, comparable to the isolation forest trained on the entire feature set. Although there were minor fluctuations in the F1-score, the model can be trained on just 80 coreset-sampled features instead of a complete feature set.

- **Verdict:** When the number of features is less than 70% of the complete feature set, uniform sampling demonstrates superior performance. However, for feature sets larger than 70%, coreset sampling outperforms uniform sampling. The hypothesis is marginally supported by the Isolation

Forest at $n = 80$. We believe that the training dataset, with 110 features, is still too small and less diverse for effective utilization of sampling techniques. Expanding the dataset with more diverse samples is essential to further demonstrate the importance of sampling techniques in reducing the number of features.

5

Conclusions

"Nobody made a greater mistake than he who did nothing because he could do only a little."

Sydney Smith

5.1 Answers to Research Questions

- **RQ1: What are the deficits in the existing collection of robotic pouring datasets, and why do we need a new dataset?**

  We briefly highlight the limitations of pouring datasets as follows:

  – There is a general scarcity of real-world datasets for anomaly detection. Anomaly detection pouring datasets are limited. Pouring datasets that could be modified for anomaly detection focus on liquid pouring over granular materials. Notably, one prominent dataset resembles toy demonstrations and is only suitable for proof-of-concept. In contrast, we contributed a granular real-world pouring dataset with a focus on anomaly detection only.

  – The datasets were collected in a quiet environment using a highly instrumented setup. In contrast, our dataset was gathered in a lab environment at various times of the day, employing a Toyota Human Support Robot equipped with an external microphone on its platform. Careful measures were taken to safeguard privacy and voice conversations.

  – The datasets have limited variations in the physical attributes of containers and pouring substances. The datasets only include overflow anomalies and no other pouring anomalies. The datasets are available on request. In contrast, our dataset is multimodal and features variations in containers and granular substances. It includes three unique pouring anomalies and will be made available freely to the research community.

- **RQ2: Which sensor fusion scheme works best for multimodal deep anomaly detection?**

  – RQ2.1: Which sensor modalities should be fused for anomaly detection?
  Human experts depend on vision and hand proprioception for robust pouring activities,

occasionally incorporating audition as well. Inspired by human expertise, we recognize **audio**, **video**, **depth**, and **haptic** data as crucial modalities for anomaly detection. Through an ablation study, we demonstrated the importance of each sensor modality using the proposed dataset. A future enhancement would be to include the force-torque sensor data, which is also part of our dataset, in the ablation study.

- RQ2.2: When to fuse different types of sensor modalities for anomaly detection?
  Sensor modalities are heterogeneous in nature. Therefore, the **late fusion** and **middle fusion** techniques proposed by [11] are suitable for sensor fusion. In this research work, our focus is on the late fusion of audio, video, and depth modalities. For future comparison with middle fusion techniques, it will be required to train the feature extractors for all modalities jointly.
- RQ2.3:How to fuse different types of sensor modalities for deep anomaly detection?
  Concatenation and **Mixture of Experts** proposed by [23] are two crucial operations for fusing heterogeneous sensor modalities. Due to their relative ease of implementation, we opted for concatenation operations to fuse the heterogeneous sensor modalities. Further details regarding the importance of Mixture of Experts for heterogeneous sensor modalities will be elaborated on in the future work section.

- **RQ3: How do the chosen anomaly detection methods perform on the proposed dataset?**
  Similar to [33], we compare the distance-based anomaly detection method (OC-SVM) and the isolation-based anomaly detection method (Isolation Forest). The comparative study leads us to the conclusion that Isolation Forest is statistically superior to OC-SVM in most sensor modal configurations on the proposed dataset. However, due to the absence of local anomaly samples in the test data, the sensitive nature of anomaly detection methods towards local anomalies could not be investigated in this research work.

5.2 Contributions

- **Literature Review:** We presented a literature review on data-driven multimodal anomaly detection methods. We took inspiration from novel concepts from related literature and also discussed the limitations of these methods. Through the literature review, we sought answers to research questions **RQ1** and **RQ2**.

- **Dataset:** We contributed a multimodal granular pouring dataset for anomaly detection, collected using a Toyota Human Support Robot. This pouring dataset is the first of its kind and draws inspiration from existing collections of pouring datasets while also addressing five limitations present in those datasets.

- **Experimentation:** We implemented the Late Fusion sensor scheme and conducted an ablation study to investigate the contribution of heterogeneous sensor modalities to anomaly detection. We presented a comparative study involving two different anomaly detection methods, and lastly, we presented a small case study to analyse the importance of two sampling techniques to reduce

dependency on full feature sets. Through experimentation, we sought answers to research questions **RQ2** and **RQ3**.

5.3 Future work

- **Improving the Proposed Dataset:** Although the proposed dataset is the first of its kind for anomaly detection, it does have limitations. The dataset is limited to three pouring anomalies, namely, hand intervention by the observer, collisions between containers, and substance overflow. Notably, pouring container slippage, a significant anomaly occurring in daily life, is not included in the dataset. Human experts rely on at least two modalities, namely, vision and haptic data, to ensure robust pouring. Like [11], our dataset is shortage of haptic data that effectively captures pouring slippage anomaly.

- **Implementation of Mixture of Experts:** Concatenation is a widely used operation in multimodal anomaly detection due to its relative ease of implementation. The Mixture of Experts is crucial for fusing sensor modalities in pouring activity. It assigns weights to sensor modalities based on the amount of noise. For instance, in scenarios with negligible noise in sensor modalities, it assigns equal weight to all modalities. Otherwise, it prioritizes the less noisy modality by assigning it a greater weight compared to the extremely noisy modalities.

- **Fine Tune Pretrained Models:** Video and audio are widely used modalities in various applications, and consequently, the deep learning research community provides a broad selection of pre-trained models for these modalities. However, in the case of depth and haptic data, pre-trained models are not as prevalent.

  In this research work, we utilized a model pretrained on video modalities for action recognition from the Kinetics 700 dataset as a feature extractor for the depth modality. The weights of the pretrained model are biased for RGB input, and features extracted will not accurately depict information contained in depth images. To use these modalities effectively with late fusion, it becomes necessary to fine-tune pre-trained models specifically for depth and haptic data.

- **Rigorous Comparative Study:** In this research work, we are comparing distance-based and isolation-based anomaly detection methods. Both methods are sensitive to local anomalies, as shown experimentally in [33]. However, our dataset is deficient in local anomalies, and therefore, a rigorous comparative study to analyze the sensitivity of these methods towards local anomalies remains as future work.

References

[1] J. B. Peterson, *Maps of meaning: The architecture of belief*. Routledge, 2002.

[2] D. Altan and S. Sariel, "Clue-ai: A convolutional three-stream anomaly identification framework for robot manipulation," *IEEE Access*, 2023.

[3] J. Wilson, A. Sterling, and M. C. Lin, "Analyzing liquid pouring sequences via audio-visual neural networks," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7702–7709, IEEE, 2019.

[4] H. Liang, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang, "Making sense of audio vibration for liquid height estimation in robotic pouring," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5333–5339, IEEE, 2019.

[5] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," *ROBOMECH journal*, vol. 6, no. 1, pp. 1–15, 2019.

[6] J. P. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery IX*, vol. 5093, pp. 230–240, SPIE, 2003.

[7] T. Nakao, S. Hanaoka, Y. Nomura, M. Murata, T. Takenaga, S. Miki, T. Watadani, T. Yoshikawa, N. Hayashi, and O. Abe, "Unsupervised deep anomaly detection in chest radiographs," *Journal of Digital Imaging*, vol. 34, pp. 418–427, 2021.

[8] X. Hu, X. Wang, X. Yang, D. Wang, P. Zhang, and Y. Xiao, "An infrared target intrusion detection method based on feature fusion and enhancement," *Defence Technology*, vol. 16, no. 3, pp. 737–746, 2020.

[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[10] S. Thoduka, J. Gall, and P. G. Plöger, "Using visual anomaly detection for task execution monitoring," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4604–4610, IEEE, 2021.

[11] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "Fino-net: A deep multimodal sensor fusion framework for manipulation failure detection," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6841–6847, IEEE, 2021.

[12] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[13] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, pp. 31–31, 2005.

[14] P. Karkus, D. Hsu, and W. S. Lee, "Particle filter networks with application to visual localization," in *Conference on robot learning*, pp. 169–178, PMLR, 2018.

[15] D. Park, Z. Erickson, T. Bhattacharjee, and C. C. Kemp, "Multimodal execution monitoring for anomaly detection during robot manipulation," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 407–414, IEEE, 2016.

[16] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.

[17] D. Dasgupta and N. S. Majumdar, "Anomaly detection in multidimensional data using negative selection algorithm," in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, vol. 2, pp. 1039–1044, IEEE, 2002.

[18] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8032–8041, 2023.

[19] A. Burns, S. Xiang, D. Lee, L. Jackel, S. Song, and V. Isler, "Look and listen: A multi-sensory pouring network and dataset for granular media from human demonstrations," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2519–2524, IEEE, 2022.

[20] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[21] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

[22] W. Gong, Y. Wang, M. Zhang, E. Mihankhah, H. Chen, and D. Wang, "A fast anomaly diagnosis approach based on modified cnn and multisensor data fusion," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 13636–13646, 2021.

[23] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[24] H. Liang, C. Zhou, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, M. Stoffel, and J. Zhang, "Robust robotic pouring using audition and haptics," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10880–10887, IEEE, 2020.

[25] U. Yamaguchi, F. Saito, K. Ikeda, and T. Yamamoto, "Hsr, human support robot as research and development platform," in *The Abstracts of the international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics: ICAM 2015.6*, pp. 39–40, The Japan Society of Mechanical Engineers, 2015.

[26] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[27] B. Thornton, "Audio recognition using mel spectrograms and convolution neural networks," 2019.

[28] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3253–3257, IEEE, 2022.

[29] Y. Liu, Y. Tan, and H. Lan, "Self-supervised contrastive learning for audio-visual action recognition," *arXiv preprint arXiv:2204.13386*, 2022.

[30] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of machine Learning research*, vol. 2, no. Dec, pp. 139–154, 2001.

[31] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, pp. 413–422, IEEE, 2008.

[32] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

[33] S. S. Z. L. Rongfang Gao, Tiantian Zhang, "Research and Improvement of Isolation Forest in Detection of Local Anomaly Points," in *Journal of Physics: Conference Series*, vol. 1237, p. 052023, IOP Publishing, 2019.