

# Multi-Modal Anomaly Detection for Unstructured and Uncertain Environments

Tianchen Ji, Sri Theja Vuppala, Girish Chowdhary, Katherine Driggs-Campbell

University of Illinois at Urbana-Champaign  
{tj12,sritheja,girishc,krdc}@illinois.edu

**Abstract:** To achieve high-levels of autonomy, modern robots require the ability to detect and recover from anomalies and failures with minimal human supervision. Multi-modal sensor signals could provide more information for such anomaly detection tasks; however, the fusion of high-dimensional and heterogeneous sensor modalities remains a challenging problem. We propose a deep learning neural network: supervised variational autoencoder (SVAE), for failure identification in unstructured and uncertain environments. Our model leverages the representational power of VAE to extract robust features from high-dimensional inputs for supervised learning tasks. The training objective unifies the generative model and the discriminative model, thus making the learning a one-stage procedure. Our experiments on real field robot data demonstrate superior failure identification performance than baseline methods, and that our model learns interpretable representations.

**Keywords:** Anomaly Detection, Feature Learning, Field Robots

## 1 Introduction

Agriculture is currently facing a labor crisis. Recent research has revealed that small, low-cost robots (agbots) deployed beneath crop canopies can coordinate to create more sustainable agro-ecosystems [1, 2]. However, one big challenge for field robot development arises from the nature of highly uncertain environment in fields, where robots are likely to encounter various types of anomalies or abnormal cases. A reliable robotic system should have the ability to detect not only the presence but also the possible causes of an anomaly [3]. Multi-modal sensor signals can provide valuable information about robot surroundings; however, the fusion of high-dimensional and heterogeneous modalities remains a challenging problem [4].

Deep-learning based anomaly detection (AD) algorithms have become increasingly popular [5]. The focus of many previous works have attempted to cast the AD problem as an one-class classification problem or as the detection of out-of-distribution samples [4, 6, 7, 8]. In this work, we concentrate on multi-class classification, which potentially guides the robot to take corresponding recovery maneuvers or call for assistance. Both SVM-based and neural-network-based algorithms have been proposed for multi-class classification for the task of anomaly detection [3, 9, 10, 11]. However, these approaches may struggle with learning multi-modal distributions [11, 12], lack the representational power for high-dimensional data [3, 10], or require a two-stage training procedure [9].

Researchers often seek to reduce the dimension of high-dimensional inputs before applying the detection or classification [13, 14, 15, 16]. A simple and common approach for dimensionality compression is principal component analysis (PCA). However, several prior works have shown that reconstruction-based methods outperformed PCA in robust feature extraction [13, 17]. A relevant approach in the reconstruction-based autoencoder domain is the variational autoencoder (VAE) [18, 19]. VAEs learn the underlying distribution of the input data using variational inference. Recent research efforts have made noteworthy progress in leveraging the representational power of deep autoencoders in unimodal classification tasks [20, 21]. Furthermore, multi-modal integration learning has been explored in out-of-distribution sample detection [4], safe robot navigation [12], robot behavior-learning [13], and deep reinforcement learning [22].

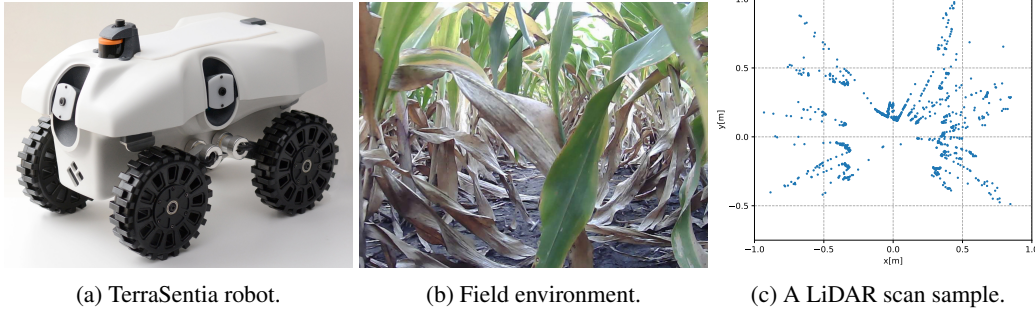


Figure 1: **(a)** Our AD module is evaluated on the data collected by TerraSentia, a 15kg and 0.31m wide field robot. **(b)** A picture of robot’s surroundings from the front camera in a corn field. Low-hanging leaves or weeds may block the LiDAR and camera view as the robot navigates under canopy. **(c)** A LiDAR scan sample in the dataset. The origin represents the robot position.

In this paper, we approach the anomaly or failure detection problem using a multi-class classifier, which simultaneously monitors the emergence of robot failures and finds possible reasons behind failures. We use a compact agricultural robot, TerraSentia, as our testbed for our model (Figure 1). Low localization accuracies between crop rows and uncertainties in sensory signals (e.g., LiDAR) due to weeds and hanging leaves in field environments lead to a higher failure rate than in indoor or static environments.

We propose the use of a supervised variational autoencoder (SVAE) model for the failure identification task, which utilizes the representational power of VAE to perform multi-class classification for multi-modal input modalities. Through unifying the optimization of the generative model and discriminative model in a single objective function, our method yields a one-stage training procedure. We demonstrate higher classification accuracy than baselines and interpretable learning results on data collected by the field robot.

## 2 Related Work

Anomaly detection, also called novelty detection or outlier detection, is an important problem that has been researched within diverse application domains [23]. In robotics, the AD problem is also related to failure detection or fault detection and an anomaly detector is often defined as a method to identify when the current execution differs from past successful experiences [4]. Early research efforts often make use of robot or sensor models for the AD task. Multiple Model Adaptive Estimation (MMAE) uses a bank of Kalman filters to predict the outcome of several faulty patterns. A neural network is then trained to identify the current state of the robot based on the residual between the predicted readings and the actual sensor readings [24]. Soika examines inconsistencies between different sensors’ statements using probabilistic sensor models to detect failures [25]. Vemuri et al. models the faults as nonlinear functions of the measured variables in system dynamics where anomalies are declared whenever off-nominal system behaviors are observed [26]. These approaches were designed for anomalies that may affect the system or sensor models; however, the generalization of these model-based methods to anomalies caused by external factors (e.g., environmental interference) remains unclear.

In the domain of supervised deep anomaly detection, deep architecture with support vector data description (SVDD) [11] uses hidden layers with SVDD to capture features that are helpful for classification. Nevertheless, Deep SVDD shows inferior performance when compared to deep autoencoders in terms of learning the joint distribution over input data due to its unimodal assumption [12]. Recently, Erfani et al. proposed to perform subspace projection and landmark selection to extract robust features from the data before training the classifier. The proposed method is capable of learning complex probability distributions of high-dimensional data. However, the optimization of the feature extractor and classifier are separate and thus training is a two-stage procedure [9].

Thanks to the representational power of deep autoencoders, VAEs have been widely used in AD tasks. LSTM-VAE replaces the feed-forward network in VAEs with LSTMs to learn joint distributions of multi-modal observations and their temporal dependencies [4]. The reconstruction proba-

bility, which considers the variability of the variable distributions, is used as a new anomaly score to detect anomalies [27]. Conditional VAEs have also been applied to identify anomalous samples given associated observable information in specific applications [28].

VAE-based methods have also been proposed to perform multi-class classification tasks. Kingma et al. propose M1 and M2 models for semi-supervised learning problem [20]. However, M1 requires two-stage training, and M2 assumes two independent encoder networks for labels and other latent variables, which may not lead to an optimal solution for supervised learning tasks. Semi-supervised VAEs, which can also be used for supervised learning, connect a classification layer to the topmost encoder layer providing unified training [21]. Nevertheless, without the use of a deep architecture for the classification layer, it is unclear how well the model generalizes to multi-modal and high-dimensional input modalities. In this work, we utilize the representational power of VAEs to improve the classification accuracy for multi-modal inputs with a one-stage training procedure.

### 3 Methodology

Our goal is to detect and classify failures or anomalies that occur during robot navigation in the field. We introduce a general data-driven anomaly detector architecture, which can be deployed on the robot and be trained efficiently based on variational inference.

We construct our anomaly detector as a deep supervised multi-class classifier. We assume that the sensor data is multi-modal, consisting of a high-dimensional input modality  $\mathbf{x}_h \in \mathbb{R}^H$  (e.g., LiDAR) and a low-dimensional input modality  $\mathbf{x}_l \in \mathbb{R}^L$  (e.g., wheel encoders). The AD module outputs zero when the robot is in normal operation, and outputs non-zero values when an anomaly occurs. We further divide the anomalies to  $C$  sub-classes to facilitate the robot to take corresponding recovery maneuvers for different types of anomalies or failures. At each time step, the AD module maps from a set of current sensor data  $\mathbf{x}_t \in \mathbb{R}^{H+L}$  to a corresponding class label  $y_t \in \{0, 1, \dots, C\}$ .

Our anomaly detector is composed of two parts: a feature generator (FG) and a classifier. A schematic overview of the AD module is depicted in Figure 2 (left). We define the FG as a function  $g : \mathbf{x}_h \mapsto \mathbf{z}$ , which maps the high-dimensional inputs  $\mathbf{x}_h \in \mathbb{R}^H$  to some latent variables  $\mathbf{z} \in \mathbb{R}^d$ . In most cases, we prefer  $d \ll H$  to ease the training of the classifier. With proper selection and training of the FG, the model is expected to learn meaningful and robust representations / features of the original inputs in the latent space.<sup>1</sup>

The latent variables  $\mathbf{z}$  generated by the FG, along with other low-dimensional inputs  $\mathbf{x}_l$  are then used as inputs to the classifier: in our case, a feed-forward network. Here, we do not constrain the low-dimensional inputs  $\mathbf{x}_l$  to be the raw sensor data, instead  $\mathbf{x}_l$  can also involve modalities derived from the raw data which may help the classifier make the correct inference [12].

#### 3.1 Supervised VAE Architecture

The classification performance of the AD module presented in Figure 2 (left) heavily depends on the quality of the feature generator. The latent variables extracted should be distinct for different classes while preserving the nature of the inputs  $\mathbf{x}_h$ . We propose a supervised variational autoencoder (SVAE) to unify the training of the FG and the classifier (Figure 2 (right)). During inference, the encoder is used as the FG while the decoder is abandoned.

SVAEs are inspired by the variational autoencoder framework. The decoder uses a generative model of the form:

$$p_\theta(\mathbf{x}_h|\mathbf{z}) = f(\mathbf{x}_h; \mathbf{z}, \theta) \quad (1)$$

where  $f(\mathbf{x}_h; \mathbf{z}, \theta)$  is a likelihood function whose probability distribution is formed by a non-linear transformation of the latent variable  $\mathbf{z}$  with parameters  $\theta$ . Here, we choose the non-linear transformation to be a multilayer perceptron (MLP) and the likelihood function  $f$  to be a Gaussian distribution:

$$f(\mathbf{x}_h; \mathbf{z}, \theta) = \mathcal{N}(\mathbf{x}_h | \text{MLP}(\mathbf{z}; \theta), \sigma^2 \cdot I) \quad (2)$$

where  $\text{MLP}(\mathbf{z}; \theta)$  is a mean vector, and  $\sigma$  is a hyperparameter.

<sup>1</sup>The FG can be implemented with fully connected layers of a neural network, or with a fully-convolutional neural network when dealing with image-like inputs [12].

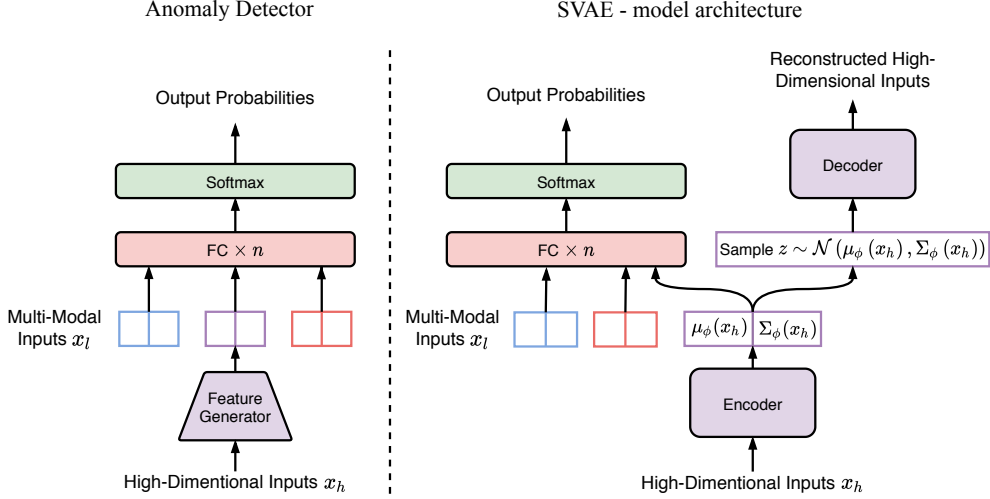


Figure 2: *Left*: The high-dimensional inputs are projected onto a latent space to extract features. The classifier makes an inference based on the compressed representation of the high-dimensional data and other low-dimensional data. *Right*: The VAE is combined with a classifier during training time. The reparameterization trick is omitted.

We represent our inference model as an approximate posterior distribution  $q_\phi(\mathbf{z}, y|\mathbf{x})$  with variational parameters  $\phi$ . For the latent variable  $\mathbf{z}$ , we use a Gaussian inference network as the encoder:

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x}_h) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}_h), \text{diag}(\boldsymbol{\Sigma}_\phi(\mathbf{x}_h))) \quad (3)$$

where  $\boldsymbol{\mu}_\phi(\mathbf{x}_h)$  is a mean vector,  $\boldsymbol{\Sigma}_\phi(\mathbf{x}_h)$  is a variance vector, and the functions  $\boldsymbol{\mu}_\phi: \mathbb{R}^H \mapsto \mathbb{R}^d$  and  $\boldsymbol{\Sigma}_\phi: \mathbb{R}^H \mapsto \mathbb{R}^d$  are parameterized as MLPs as in the decoder.

The classifier is connected to the output layer of the encoder. We specify this inference model for  $y$  as a categorical distribution:

$$q_\phi(y|\mathbf{x}) = \text{Cat}(y|\boldsymbol{\pi}_\phi(\boldsymbol{\mu}_\phi(\mathbf{x}_h), \boldsymbol{\Sigma}_\phi(\mathbf{x}_h), \mathbf{x}_l)) \quad (4)$$

where  $\text{Cat}(y|\boldsymbol{\pi}_\phi)$  is the categorical distribution,  $\boldsymbol{\pi}_\phi(\cdot)$  is a probability vector and is represented as an MLP.

Unlike the prior work in generative semi-supervised model for the digit classification task [29] where the inference networks  $q_\phi(\mathbf{z}|\mathbf{x})$  and  $q_\phi(y|\mathbf{x})$  are assumed to be independent with each other, we expect that the inference model  $q_\phi(y|\mathbf{x})$  depends on  $q_\phi(\mathbf{z}|\mathbf{x})$ . In the former case, the class specification is separated from the writing style of the digit in order to facilitate the generation for specific digits. However, in our case, the features generated by  $q_\phi(\mathbf{z}|\mathbf{x})$  relate to classification criteria which are utilized by the classifier to make an inference.

### 3.2 Training

Our model focuses on two different but related tasks: classification and reconstruction. The inference model  $q_\phi(y|\mathbf{x})$  generates a probability distribution over class labels  $y$  while the generative model  $p_\theta(\mathbf{x}_h|\mathbf{z})$  tries to reconstruct the inputs  $\mathbf{x}_h$ . For our application, the main goal is to find a high-quality inference model for the class label given the current sensor data; however, we found it very useful to learn a generative model simultaneously. We claim that the reconstruction task serves as a regularization [30, 31, 32], which forces the encoder to learn global features of the high-dimensional inputs that are critical for both the classifier and the decoder. As a positive side effect, by compressing the input dimensions, the model learns to filter out sensor noise and to extract much more robust high-level ideas from those inputs.

Denoting the labeled dataset by  $\mathcal{D}$ , we specify the overall loss function for SVAE on the entire dataset as:

$$\mathcal{L} = \sum_{(\mathbf{x}, y) \in \mathcal{D}} [-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}_h|\mathbf{z})] + D_{\text{KL}}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})] - \alpha \cdot \log q_\phi(y|\mathbf{x})] \quad (5)$$



where  $p_\theta(\mathbf{z})$  is the prior distribution of the latent variable  $\mathbf{z}$ , and the hyperparameter  $\alpha$  controls the relative weight between the generative and discriminative learning. As in the typical setting of VAE with continuous latent variables [19], we choose  $p_\theta(\mathbf{z})$  as a standard Gaussian distribution  $\mathbf{z} \sim \mathcal{N}(0, I)$ .

The overall training objective has two parts, each of which focuses on different tasks. The first two terms in equation (5), which is the negative of the evidence lower bound (ELBO) in vanilla VAEs, evaluate the reconstruction loss of the high-dimensional inputs  $\mathbf{x}_h$  and contribute to the optimization of the decoder and the encoder. The second KL divergence term can be viewed as a regularization.

The third term in the objective function (5) penalizes the classification error, and contributes to the optimization of the classifier and the encoder. We set  $\alpha = 0.1 \cdot N$ , where  $N$  is the total number of datapoints. An interesting observation from our experiments is that increasing the relative weight  $\alpha$  for the classification task does not necessarily generate higher classification accuracy. We argue that with less attention on the reconstruction task, the model tends to lose the generalization property [30].

The inference model and the generative model can be optimized jointly by stochastic gradient descent of the unified objective function (5). To enable the backpropagation through the sampling process within the network (see Figure 2), the reparameterization trick is used to move the sampling to a stochastic input layer. We refer readers to [19] for details.

## 4 Experimental Results

The proposed model for anomaly detection was evaluated on data collected with the TerraSentia robot in corn and sorghum fields from August to October 2019. The robot navigates through rows of crops under cluttered canopies without damaging the plants. During data collection, the robot was either teleoperated by a human operator or driven by navigation algorithms autonomously while a failure may or may not occur in a run.

In all experiments using SVAEs, we implement the encoder with only one hidden layer and 128 hidden units. The decoder follows the same structure as the encoder. We construct the classifier with one hidden layer and 64 hidden units. We choose a 2-dimensional latent space,  $\mathbf{z} \in \mathbb{R}^2$ . ReLU activation functions are applied and an Adam optimizer with a constant learning rate of 0.0005 is used to train the network<sup>2</sup> [33].

### 4.1 TerraSentia Robot and Input Modalities

TerraSentia is a low-cost, ultracompact and ultralight field robot. The robot is equipped with Hokuyo UST-10LX, a 2D LiDAR sensor which covers a 270° range with 0.25° angular resolution, maximum distance reading of 30 m, and 40 Hz update rate. The LiDAR is placed 0.15m to the front of the robot and is adjusted to have a symmetric view about the longitude axis of the robot. The robot is driven by a two-motor set on the left and right wheels. Wheel velocities are measured by a two-channel Hall-effect encoder which provides 64 counts per revolution for each motor.

In our experiments, the AD module takes the LiDAR data of dimension 1080 as the high-dimensional input  $\mathbf{x}_h$ . The low-dimensional input  $\mathbf{x}_l$  has two components: (i) velocities from left and right wheel encoders; (ii) other modalities derived from the raw sensor data. In this work, we choose the second part of  $\mathbf{x}_l$  as an array consisting of average distances from the robot to the surrounding objects within a set of fixed ranges of scanning angles:

$$\mathbf{x}_l = \left[ v_{\text{left}}; v_{\text{right}}; \bar{d}_{[\frac{\pi}{3}, \frac{5\pi}{12}]}; \bar{d}_{[\frac{5\pi}{12}, \frac{\pi}{2}]}; \bar{d}_{[\frac{\pi}{2}, \frac{7\pi}{12}]}; \bar{d}_{[\frac{7\pi}{12}, \frac{2\pi}{3}]} \right] \in \mathbb{R}^6 \quad (6)$$

where the subscript of  $\bar{d}$  denotes the ranges of LiDAR scanning angle.<sup>3</sup> Note that the additional input modalities derived from LiDAR data is the only part that we design specifically for our robotic platform. However, such additional inputs can be designed similarly on other robots without any change of the model architecture.

<sup>2</sup>The implementation of SVAE is available at <https://github.com/tianchenji/Multimodal-SVAE>.

<sup>3</sup>This additional modalities derived from LiDAR might alleviate the reliance on the raw LiDAR data; however, our experiments suggested that the combination of the two generates the best results.



Figure 3: Snapshots of anomalous cases from the front camera on the robot in our dataset.

## 4.2 Dataset

The raw LiDAR data is clipped at 1.8 meters and normalized to  $[0,1]$ . We attach a class label to each synchronized data point  $\mathbf{x}$ . Focusing on the anomalous cases induced by environmental interference, we divide the operation mode during the runs into the following four situations (Figure 3):

- 0) *normal*: The robot follows the center line and navigates safely towards the destination.
- 1) *row collision*: The robot deviates from the center line and crashes into crops on either side of the narrow row (lane width of 0.8m).
- 2) *untraversable obstacle*: The robot stops in front of the obstacles which obstruct the center line. Such obstacles may have larger size than the robot or are rooted in the ground, making the robot unable to traverse. Typical examples include weeds, lodged plants, etc.
- 3) *traversable obstacle*: The robot stops or significantly slows down in front of the obstacles. However, the robot can drive over the obstacle with larger motor torques. Such obstacles may include lodged crop stems, stones, uneven terrains, etc.

The above four classes are labeled as an integer from 0 to 3, respectively. Note that each class corresponds to a different recovery maneuver. For instance, the robot can reverse and re-plan the path for row collisions, call for help for untraversable obstacles, or increase motor torque to run over the traversable obstacles. As a result, the output of the AD module can be used as a high-level decision for the planning and control algorithms.

We split the datapoints into a training set (80%) and a test set (20%). The training set and test set came from independent runs. Under-sampling of normal cases and over-sampling of anomalous cases were performed to obtain a more balanced dataset for training purposes.

## 4.3 Results

**Failure Identification Performance.** We test the performance of SVAE on the test set, along with the following baseline methods:

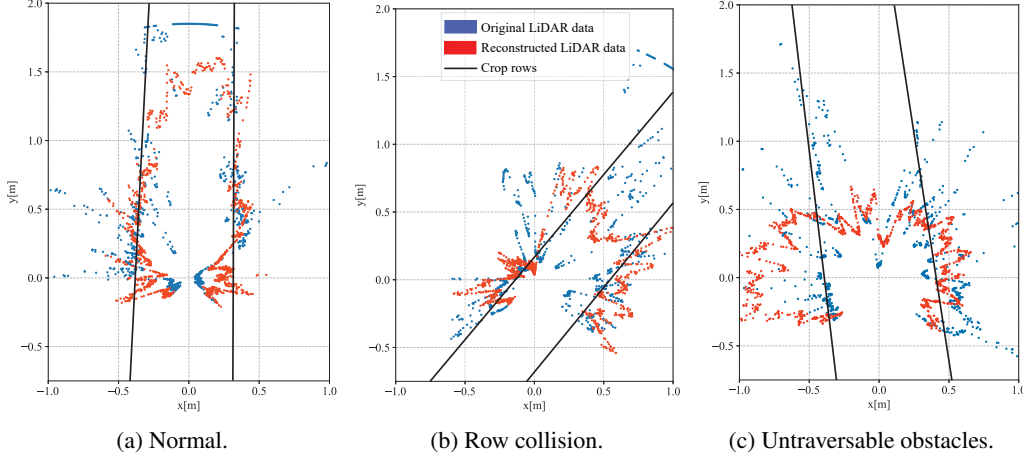
- 1) MLP: A multilayer perceptron with three hidden layers mapping from multi-modal inputs  $\mathbf{x}$  directly to class labels  $y$ .
- 2) PCA + MLR: The high-dimensional input  $\mathbf{x}_h$  is compressed through principal component analysis (PCA), then concatenated with low-dimensional input  $\mathbf{x}_l$  as the inputs to multinomial logistic regression (MLR).
- 3) VAE Fixed Features + MLP: This model was proposed by Kingma et al. [29]. The training is a two-stage procedure. First, a VAE is trained by maximizing the ELBO. Then the encoder is used as a feature extractor, and a classifier is trained on the features.
- 4) Uni-modal SVAE: The multi-modal inputs  $\mathbf{x}_h$  and  $\mathbf{x}_l$  are treated identically by concatenating and feeding them to the encoder. The classifier makes an inference merely based on the encoder output. To account for the imbalanced input dimensions, we repeat each element in  $\mathbf{x}_l$  by  $H$  times to match the size of  $\mathbf{x}_h$ .

Quantitatively, we compare different methods using the following two metrics: (i) precision, including the classification accuracy for each class and the average value; (ii) Kappa coefficient [34], which is a measure of reliability or agreement between the ground truth and the trained model.<sup>4</sup>

<sup>4</sup>Kappa coefficient can take any value between -1 to 1. A score of 0 means there is random agreement between the two, whereas a score of 1 means there is a complete agreement among the two. Kappa coefficient is usually regarded as a classification accuracy normalized by the imbalance of the classes in the dataset.

Table 1: Classification results with different anomaly detection methods

Model	normal	row collision	untraversable obstacle	traversable obstacle	average	Kappa coefficient
MLP	98.96 $\pm$ 0.56	59.04 $\pm$ 10.50	4.22 $\pm$ 1.88	69.39 $\pm$ 13.04	57.90 $\pm$ 4.47	0.63 $\pm$ 0.05
PCA+MLR	97.87 $\pm$ 0.09	78.08 $\pm$ 0.00	2.22 $\pm$ 0.00	<b>96.84 <math>\pm</math> 0.48</b>	68.75 $\pm$ 0.11	0.73 $\pm$ 0.00
VAE+MLP	98.47 $\pm$ 0.69	79.86 $\pm$ 6.39	49.89 $\pm$ 5.98	87.46 $\pm$ 3.77	78.92 $\pm$ 1.69	0.81 $\pm$ 0.03
SVAE (uni)	<b>99.47 <math>\pm</math> 0.29</b>	80.41 $\pm$ 7.91	45.00 $\pm$ 3.68	84.91 $\pm$ 1.22	77.45 $\pm$ 1.87	0.82 $\pm$ 0.01
SVAE	98.94 $\pm$ 0.59	<b>82.19 <math>\pm</math> 6.02</b>	<b>58.45 <math>\pm</math> 4.20</b>	88.42 $\pm$ 3.50	<b>82.00 <math>\pm</math> 1.52</b>	<b>0.84 <math>\pm</math> 0.02</b>

Figure 4: Reconstructed LiDAR point clouds from the latent space. The encoder compresses  $\mathbf{x}_h \in \mathbb{R}^{1080}$  to  $\mathbf{z} \in \mathbb{R}^2$ , then the decoder takes  $\mathbf{z}$  to reconstruct  $\mathbf{x}_{\text{recon}} \in \mathbb{R}^{1080}$ . The robot is at the origin.

All four baselines, along with the SVAE, are evaluated with randomly initialized weights over 10 runs. Table 1 summarizes the results. As shown, the SVAE achieves the highest average classification accuracy and best Kappa coefficient. Moreover, the individual classification accuracy of our model for the four classes is either the highest or comparable to the highest among the five models. Note that the MLP generates the overall largest standard deviation whereas PCA+MLR has the lowest standard deviation due to fewer parameters. An interesting observation is that PCA+MLR performs well in classifying three out of the four cases while generating low accuracy for untraversable obstacles. We analysed the confusion matrix and found that 81.11% of the untraversable obstacles were classified as row collision.<sup>5</sup> We hypothesize that this is due to ignoring the global features of LiDAR clouds and thus making inference mainly based on local details. The robot should see a tilted path from the LiDAR in the case of row collision, whereas the path should be more aligned with the robot’s longitude axis in the case of untraversable obstacles. As will be shown in Figure 4, the SVAE learns such global features and thus producing more promising results.

In the SVAE, the encoder is trained to learn the features beneficial to classification and reconstruction, whereas that in VAE+MLP learns the features only useful to the reconstruction task. Such joint learning generally guides SVAE to perform better on classification task. The effectiveness of the multi-modal aspect of our model was verified by the superior classification performance of the SVAE over the uni-modal SVAE. By feeding the inputs  $\mathbf{x}_h$  and  $\mathbf{x}_l$  to the model at different stages, the encoder is able to focus on learning the features of high-dimensional inputs, without being distracted by low-dimensional input modalities. The relatively large variance of the shape and size of the untraversable obstacles, along with a smallest number of samples in the training set, might be the reasons for its lowest classification accuracy among the four cases.

**Learned Feature Space.** In addition to the quantitative evaluation of our proposed model above, we now present a qualitative results of the learned feature space in the SVAE. The main goal is to visualize what features the latent variable  $\mathbf{z}$  is learning from the high-dimensional input  $\mathbf{x}_h$ .

After training, we feed the data  $\mathbf{x}$  in the test set through the model to generate reconstructed LiDAR point clouds using the decoder. Note that the decoder here serves the purpose of qualitative analysis, and is usually ignored in the deployment of the model. Figure 4 shows the results for some typical scenarios. With only two numbers representing the original point cloud of dimension 1080, the

<sup>5</sup>Confusion matrices for all five models are provided in Appendix.

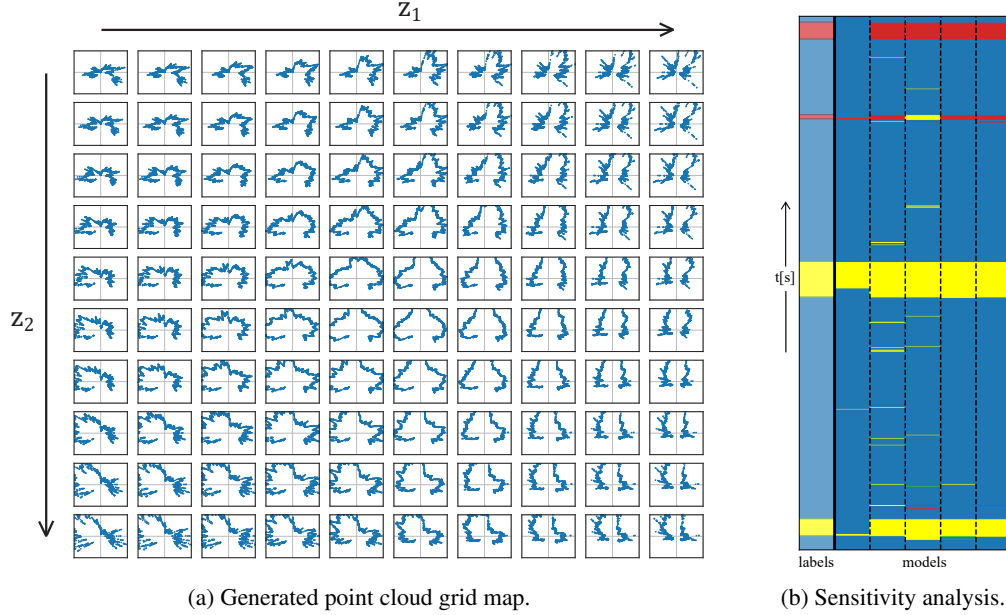


Figure 5: **(a)** High-level features for LiDAR point clouds obtained by varying latent variables  $\mathbf{z}$ . **(b)** The outputs of the five models and ground truth during a run in the test set. The colors of each column represents the actual or inferred labels. We provide more details in Appendix. From left to right: ground truth, MLP, PCA+MLR, VAE+MLP, SVAE (uni-modal), and SVAE. Blue - normal case, Yellow - row collision, Green - untraversable obstacle, red - traversable obstacle.

decoder can still manage to reconstruct reasonable results, with an average reconstruction error of 0.388m per LiDAR point over the test set. With a more detailed observation, we find that the reconstructions for normal cases and untraversable obstacles are more symmetric about the  $y$ -axis, while the row collision case shows a more tilted path. Moreover, the reconstructions for the normal cases show more free space in front, whereas that for the untraversable obstacle show less.

We further explore the physical meaning of the latent variable by feeding different combinations of  $(z_1, z_2)$  in the 2-D space to the decoder. Figure 5 (a) shows the point clouds grid map with  $z_1, z_2$  increasing in the direction of the arrows. As can be seen,  $z_1$  learns to represent how wide the robot’s front view is, making the point clouds in front further away as it increases. In contrast,  $z_2$  manages to represent the orientation of the crop rows, making the point clouds orient from right to left as it increases. The model learns these two important features automatically during training without any specific design or hand tuning.

**Sensitivity Analysis.** In practice, a crucial property of the AD module is its sensitivity to the anomalies. Ideally, the frequency of the declaration of abnormal cases should be similar to that of the actual anomalies. A highly sensitive AD module may frequently intervene during the normal operation of the robot, making it unable to work under uncertain environments with noisy measurements. On the contrary, insensitive AD module may miss some actual failures which could do harm to the robot. In Figure 5 (b), we investigate the sensitivity by analysing the output of the AD modules in a run. As shown, the SVAE interrupts the robot only when necessary and captures most of the anomalies correctly.

## 5 Conclusion

In this work, we introduced the SVAE, a new family of multi-class classifier for multi-modal failure or anomaly detection which combines VAEs with discriminative models. The training of our model is a simple, one-stage procedure. During test time, high-dimensional modalities are compressed through the encoder to the latent space, and then concatenated with other multi-modal modalities as inputs to the classifier to make an inference. The experiments on field robot data showed that SVAEs are capable of extracting robust and interpretable features from high-dimensional modalities, and that the resultant classifier outperforms baseline methods in failure identification tasks.

## Acknowledgments

This work was supported in part by the National Robotics Initiative 2.0 (NIFA#2021-67021-33449), NSF CPS (NSF#1739874, NIFA#2018-67007-28379), and the Illinois Center for Digital Agriculture. We would like to thank EarthSense for providing the field robot data, and the reviewers for their thorough and constructive comments. We also thank Peter Du and Zhe Huang for their thoughtful feedback on paper drafts.

## References

- [1] V. A. Higuti, A. E. Velasquez, D. V. Magalhaes, M. Becker, and G. Chowdhary. Under canopy light detection and ranging-based autonomous navigation. *Journal of Field Robotics*, 36(3): 547–567, 2019.
- [2] E. Kayacan and G. Chowdhary. Tracking error learning control for precise mobile robot path tracking in outdoor environment. *Journal of Intelligent & Robotic Systems*, 95(3-4):975–986, 2019.
- [3] A. L. Christensen, R. O’Grady, M. Birattari, and M. Dorigo. Fault detection in autonomous robots based on fault injection and learning. *Autonomous Robots*, 24(1):49–67, 2008.
- [4] 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*, 3(3): 1544–1551, 2018.
- [5] R. Chalapathy and S. Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.
- [6] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pages 4393–4402, 2018.
- [7] R. Chalapathy, A. K. Menon, and S. Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.
- [8] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS ONE*, 11(4):e0152173, 2016.
- [9] S. M. Erfani, M. Baktashmotlagh, M. Moshtaghi, V. Nguyen, C. Leckie, J. Bailey, and K. Ramamohanarao. From shared subspaces to shared landmarks: A robust multi-source classification approach. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [10] V. Jumutc and J. A. Suykens. Multi-class supervised novelty detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2510–2523, 2014.
- [11] S. Kim, Y. Choi, and M. Lee. Deep learning with support vector data description. *Neurocomputing*, 165:111–117, 2015.
- [12] L. Wellhausen, R. Ranftl, and M. Hutter. Safe robot navigation via multi-modal anomaly detection. *IEEE Robotics and Automation Letters*, 5(2):1326–1333, 2020.
- [13] K. Noda, H. Arie, Y. Suga, and T. Ogata. Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems*, 62(6):721–736, 2014.
- [14] A. Rodriguez, D. Bourne, M. Mason, G. F. Rossano, and J. Wang. Failure detection in assembly: Force signature analysis. In *IEEE International Conference on Automation Science and Engineering*, pages 210–215, 2010.
- [15] V. Sukhoy, V. Georgiev, T. Wegter, R. Sweidan, and A. Stoytchev. Learning to slide a magnetic card through a card reader. In *IEEE International Conference on Robotics and Automation*, pages 2398–2404, 2012.
- [16] F. Kamalov and H. H. Leung. Outlier detection in high dimensional data. *Journal of Information & Knowledge Management*, 19(01):2040013, 2020.



- [17] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [19] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- [20] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.
- [21] F. Berkhahn, R. Keys, W. Ouertani, N. Shetty, and D. Geißler. Augmenting variational autoencoders with sparse labels: A unified framework for unsupervised, semi-(un) supervised, and supervised learning. *arXiv preprint arXiv:1908.03015*, 2019.
- [22] P. Chang, S. Liu, H. Chen, and K. Driggs-Campbell. Robot sound interpretation: Combining sight and sound in learning-based control. *arXiv preprint arXiv:1909.09172*, 2019.
- [23] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009.
- [24] P. Goel, G. Dedeoglu, S. I. Roumeliotis, and G. S. Sukhatme. Fault detection and identification in a mobile robot using multiple model estimation and neural network. In *IEEE International Conference on Robotics and Automation*, pages 2302–2309, 2000.
- [25] M. Soika. A sensor failure detection framework for autonomous mobile robots. In *IEEE/RSJ International Conference on Intelligent Robot and Systems*, pages 1735–1740, 1997.
- [26] A. T. Vemuri and M. M. Polycarpou. Neural-network-based robust fault diagnosis in robotic systems. *IEEE Transactions on Neural Networks*, 8(6):1410–1420, 1997.
- [27] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.
- [28] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini. Anomaly detection with conditional variational autoencoders. In *IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1651–1657, 2019.
- [29] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.
- [30] L. Le, A. Patterson, and M. White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems*, pages 107–117, 2018.
- [31] T. Liu, D. Tao, M. Song, and S. J. Maybank. Algorithm-dependent generalization bounds for multi-task learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2): 227–241, 2016.
- [32] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [34] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

## A Additional Failure Identification Results

To further analyze the classification performance of the SVAE along with other baseline methods, we computed the confusion matrices for the five models over the test set. As in Section 4.3, we compute the mean and standard deviation of the classification results over 10 runs with different weight initializations. Table 2–6 summarize the results in percentages.

As shown in the first row in Table 2, around 25% of each anomalous case were mis-classified as normal. Without a feature generator extracting features from high-dimensional inputs, MLP struggles with distinguishing generic anomalous cases from the normal case. By contrast, the other four models generate nearly zero probabilities of classifying an abnormal case as normal. This observation indicates the importance of dimensionality compression when dealing with high-dimensional and multi-modal sensor modalities. As mentioned before, PCA+MLR fails to learn the global features of the LiDAR data and thus confuses row collisions and untraversable obstacles. Such claim is confirmed by the similar classification results in the two columns of row collision and untraversable obstacle in Table 3. The confusion matrices for VAE+MLP and SVAE share the most similarities among the five models. However, the SVAE shows the better failure identification performance due to the joint learning of the discriminative model and the generative model. Due to the increased input size and the resultant larger network, the uni-modal SVAE encounters a longer training time and the loss of interpretability of the latent space, without showing a superior classification performance over the SVAE.

Table 2: Confusion matrix for MLP over the test set

		Actual			
		normal	row collision	untraversable	traversable
Predicted	normal	98.96 $\pm$ 0.56	26.03 $\pm$ 11.15	24.67 $\pm$ 6.78	27.54 $\pm$ 11.53
	row collision	0.72 $\pm$ 0.39	59.04 $\pm$ 10.50	69.67 $\pm$ 6.79	3.07 $\pm$ 4.01
	untraversable	0.05 $\pm$ 0.09	0.00 $\pm$ 0.00	4.22 $\pm$ 1.88	0.00 $\pm$ 0.00
	traversable	0.28 $\pm$ 0.26	14.93 $\pm$ 6.63	1.44 $\pm$ 1.57	69.39 $\pm$ 13.04

Table 3: Confusion matrix for PCA+MLR over the test set

		Actual			
		normal	row collision	untraversable	traversable
Predicted	normal	97.87 $\pm$ 0.09	0.00 $\pm$ 0.00	1.11 $\pm$ 0.00	0.00 $\pm$ 0.00
	row collision	1.50 $\pm$ 0.09	78.08 $\pm$ 0.00	81.11 $\pm$ 0.00	3.16 $\pm$ 0.48
	untraversable	0.25 $\pm$ 0.00	2.74 $\pm$ 0.00	2.22 $\pm$ 0.00	0.00 $\pm$ 0.00
	traversable	0.37 $\pm$ 0.00	19.18 $\pm$ 0.00	15.56 $\pm$ 0.00	96.84 $\pm$ 0.48

Table 4: Confusion matrix for VAE+MLP over the test set

		Actual			
		normal	row collision	untraversable	traversable
Predicted	normal	98.47 $\pm$ 0.69	0.00 $\pm$ 0.00	1.11 $\pm$ 0.00	0.00 $\pm$ 0.00
	row collision	1.07 $\pm$ 0.69	79.86 $\pm$ 6.39	26.67 $\pm$ 8.36	11.84 $\pm$ 3.95
	untraversable	0.10 $\pm$ 0.06	11.78 $\pm$ 1.16	49.89 $\pm$ 5.98	0.70 $\pm$ 1.93
	traversable	0.35 $\pm$ 0.13	8.36 $\pm$ 6.07	22.33 $\pm$ 2.94	87.46 $\pm$ 3.77

Table 5: Confusion matrix for SVAE (uni-modal) over the test set

		Actual			
		normal	row collision	untraversable	traversable
Predicted	normal	99.47 $\pm$ 0.29	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.09 $\pm$ 0.28
	row collision	0.17 $\pm$ 0.11	80.41 $\pm$ 7.91	44.78 $\pm$ 3.52	14.13 $\pm$ 0.28
	untraversable	0.18 $\pm$ 0.21	6.30 $\pm$ 1.32	45.00 $\pm$ 3.68	0.88 $\pm$ 1.09
	traversable	0.18 $\pm$ 0.07	13.29 $\pm$ 7.42	10.22 $\pm$ 0.47	84.91 $\pm$ 1.22

Table 6: Confusion matrix for SVAE over the test set

		Actual			
		normal	row collision	untraversable	traversable
Predicted	normal	$98.94 \pm 0.59$	$0.00 \pm 0.00$	$1.11 \pm 0.00$	$0.00 \pm 0.00$
	row collision	$0.62 \pm 0.58$	$82.19 \pm 6.02$	$19.56 \pm 8.28$	$11.49 \pm 3.52$
	untraversable	$0.08 \pm 0.03$	$11.92 \pm 0.92$	$58.45 \pm 4.20$	$0.09 \pm 0.28$
	traversable	$0.36 \pm 0.15$	$5.89 \pm 6.43$	$20.89 \pm 5.34$	$88.42 \pm 3.50$

In practice, robots can always apply larger motor torques to try running over the obstacles without damaging the plants as long as the output of the AD module is not row collision. The robot would either return to normal operation if the obstacle is traversable or call for assistance if the obstacle is untraversable. Such strategy groups traversable and untraversable obstacles as a whole, and thus leads to simplified confusion matrices with higher average classification accuracy of the obstacles than in the original confusion matrices. Here, we only present the results of the SVAE in Table 7, which generates the best classification performance among the five models.

Table 7: Simplified confusion matrix for SVAE over the test set

		Actual		
		normal	row collision	obstacle
Predicted	normal	$98.94 \pm 0.59$	$0.00 \pm 0.00$	$0.49 \pm 0.00$
	row collision	$0.62 \pm 0.58$	$82.19 \pm 6.02$	$15.05 \pm 3.78$
	obstacle	$0.44 \pm 0.16$	$17.81 \pm 6.02$	$84.46 \pm 3.78$

## B Additional Sensitivity Analysis Results

Figure 6 shows the additional results of the sensitivity analysis in Section 4.3 for the five models during the run. The probabilities of classes generated by the models are represented as solid lines. As shown, the uni-modal SVAE and the SVAE generate the best results in this run. With a more detailed observation, the SVAE is more confident in detecting anomalies generally, and has higher classification accuracies for failure cases as suggested by the quantitative results in Section 4.3.

During data collection, the human operator reset the robot in the correct orientation and resumed moving forward after an anomaly occurred. Such post-anomaly stage before resuming the normal operation is difficult to classify due to the human intervention. To maintain the continuity in time in sensitivity analysis, we keep the output of the AD module unchanged as the previous time step when such discontinuity in time occurs.

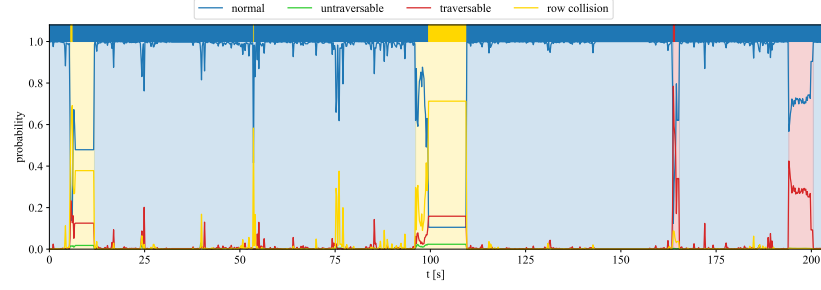
## C Baseline Methods Architectures

We briefly introduce the architectures for the baseline methods implemented in Section 4:

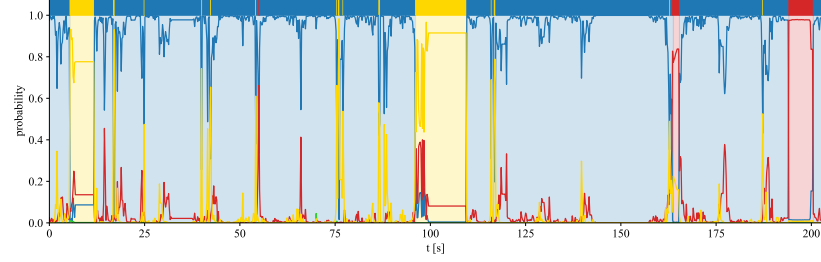
- 1) MLP: We used three hidden layers. The number of hidden units are [1024, 512, 512].
- 2) PCA+MLR: We compressed the LiDAR data to 30 components, with 74.48% of variance retained. The classifier is implemented as multinomial logistic regression.<sup>6</sup>
- 3) VAE+MLP: For the encoder in the VAE, we used the same architecture as in the SVAE with one hidden layer and 128 hidden units. The decoder mirrors the encoder. We chose a 2-dimensional latent space. The MLP also follows the same architecture as in the SVAE with one hidden layer and 64 hidden units.
- 4) SVAE (uni): To account for the larger input, we used two hidden layers of size [2048, 1024] for the encoder. The decoder mirrors the encoder. We chose the latent size to be 128. The classifier has the same structure as in the SVAE with one hidden layer and 64 hidden units.

ReLU activation functions were applied. The learning rate for each baseline method was determined individually by selecting the one generating the best average classification accuracy over the four classes, thus may not be strictly the same.

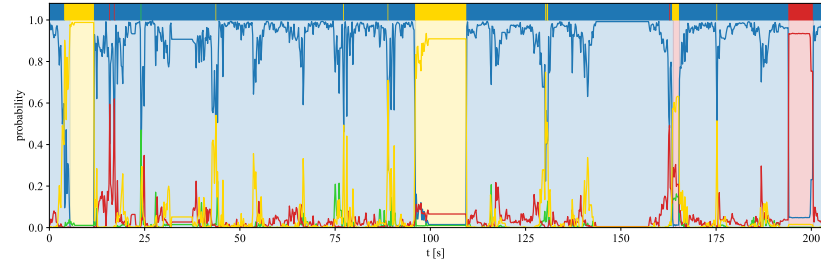
<sup>6</sup>D. Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197-200, 1992.



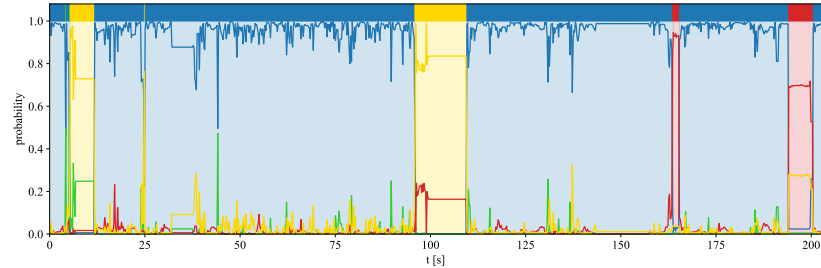
(a) MLP.



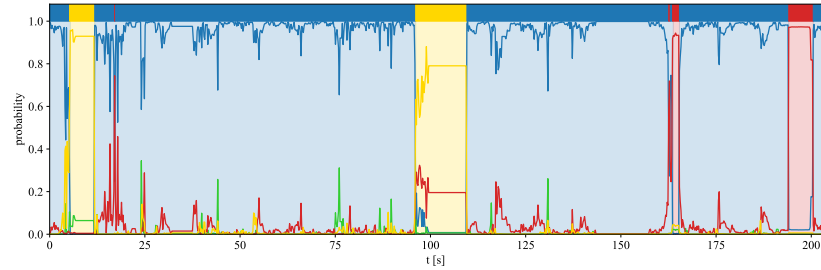
(b) PCA+MLR.



(c) VAE+MLP.



(d) SVAE (uni-modal).



(e) SVAE.

Figure 6: Detailed sensitivity analysis results. The colors of the top bars represent the outputs (i.e., inferred labels) of different models. The colors of the shaded areas represent the ground truth labels.