

# Rule-Based Error Detection and Correction to Operationalize Movement Trajectory Classification

Bowen Xi, Kevin Scaria, Paulo Shakarian

Arizona State University  
Tempe, AZ  
{bowenxi, kscaria, pshak02}@asu.edu

## Abstract

Classification of movement trajectories has many applications in transportation. Supervised neural models represent the current state-of-the-art. Recent security applications require this task to be rapidly employed in environments that may differ from the data used to train such models for which there is little training data. We provide a neuro-symbolic rule-based framework to conduct error correction and detection of these models to support eventual deployment in security applications. We provide a suite of experiments on several recent and state-of-the-art models and show an accuracy improvement of 1.7% over the SOTA model in the case where all classes are present in training and when 40% of classes are omitted from training, we obtain a 5.2% improvement (zero-shot) and 23.9% (few-shot) improvement over the SOTA model without resorting to retraining of the base model.

## 1 Introduction

The identification of a mode of travel for a time-stamped sequence of global position system (GPS) known as “movement trajectories” has important applications in travel demand analysis (Huang et al. 2019), transport planning (Lin & Hsu 2014), and analysis of sea vessel movement (Fikioris et al. 2023). The current state-of-the-art has relied on supervised neural models (Kim et al. 2022). More recently this problem has been of interest for security applications such as leading to efforts such as the IARPA HAYSTAC program<sup>1</sup>. In this domain, models may be deployed in environments with different geography, transportation infrastructure, and socio-cultural dynamics than in the training data and expected to adapt to such environments with little or no labeled data specific to those circumstances. Further, such deployments may happen rapidly, precluding extensive data engineering or model retraining.

In this paper, we extend the current supervised neural methods with a lightweight error detection and correction rule (EDCR) framework providing an overall neurosymbolic system. The key intuition is that training and operation data can be used to learn rules that predict and correct errors in the supervised model. Once trained, the rules are employed operationally in two phases: first detection rules identify potentially misclassified movement trajectories. A second type

of rule to re-classify the trajectories (“correction rules”) is then used to re-assign the sample to a new class. Our key contributions are as follows: (1.) We present a strong theoretical framework for EDCR rooted in logic and rule mining and formally prove how quantities related to learned rules (e.g., confidence and support) are related to changes in class-level machine learning metrics such as precision and recall. (2.) We conduct experiments where rules trained on the same data as the original model can improve machine learning metrics across various settings and model types, including the SOTA LRCN model. Specifically, the employment of EDCRs leads to a 1.7% improvement in accuracy over the original LRCN model when data leakage between training and testing is minimized (3.) By excluding 40% of the classes during the training process, we enhance 5.2% (zero-shot) and 23.9% improvement (few-shot) compared to the SOTA model. This progress is accomplished without necessitating any retraining of the underlying base model. (4.) In addition to offering domain knowledge akin to other papers, we furnish a neural network-incorporated condition, characterized by its overarching generality, thereby enhancing the versatility of EDCR for diverse problem domains. (5.) As a side result, we extend the LRCN SOTA model of (Kim et al. 2022) with attention mechanisms that establish a new SOTA baseline in certain cases without EDCR. This model is also improved with EDCR.

The rest of the paper is outlined as follows. In Section 2, we describe the movement trajectory classification problem (MTCP) and associated classification approaches, including our new “LRCN with attention” (LRCNa) model. Then we introduce our error detecting and correcting rule framework (Section 3) which formalizes our strategy for EDC and provides analytical results that support our algorithm development. This is followed by experimental results in Section 4 followed by a discussion on related work and future directions. Additional details supporting the reproducibility of both formal results (e.g., proofs) and experiments (e.g., data preprocessing and experimental details) along with code can be found in an online appendix available at <https://github.com/lab-v2/Error-Detection-and-Correction>.

## 2 Technical Preliminaries

In this section, we introduce MTCP, describe the vector embeddings used for a neural based classifier (Dabiri & Heaslip

<sup>1</sup><https://www.iarpa.gov/research-programs/haystac>

2018; Kim et al. 2022) as well as the three neural architectures utilized CNN (Dabiri & Heaslip 2018), Long-term Recurrent Convolutional Network (LRCN) (Kim et al. 2022), and (newly introduced in this work) LRCN with attention (LRCNa).

**Movement Trajectory Classification Problem.** We define the MTCP problem as given a sequence of GPS points,  $\omega$ , and assign a movement class from  $\mathcal{C}$ . The number of classes in  $\mathcal{C}$  is  $n$ . In this work, as per others (e.g., (Dabiri & Heaslip 2018; Kim et al. 2022) )we define  $\mathcal{C} = \{\text{walk, bike, bus, drive, train}\}$ , though we will typically not refer to specific classes outside of the description of the experiments for purposes of generalizability. The current paradigm for the MTCP problem is to create a neural model  $f_\theta$  that maps sequences to movement classes using a set of weights,  $\theta$ . In this approach traditional methods (i.e., gradient descent) find a set of parameters such that a loss function is minimized based on some training set  $\mathcal{T}$  (where each sample  $\omega \in \mathcal{T}$  is associated with a ground truth class  $gt(\omega)$ ). Formally:  $\arg \min_{\theta} \mathbb{E}_{\omega \in \mathcal{T}} \text{Loss}(f_\theta(\omega), gt(\omega))$ . We also note that with each sample  $\omega$ , we will associate three predicates for each class  $i$ :  $pred_i$ ,  $corr_i$ , and  $error_i$  that we will later use to describe a logic for reasoning about error correction.

- $pred_i$ : if the model predicted class  $i$ :  $pred_i(\omega)$  is true iff  $f_\theta(\omega) = i$ .
- $corr_i$ : the correct movement class for  $\omega$ :  $corr_i(\omega)$  is true iff  $gt(\omega) = i$ .
- $error_i$  if the model had an error:  $error_i(\omega)$  is true iff  $f_\theta(\omega) \neq gt(\omega)$ . In other words: the model is wrong and predicted class  $i$ .

**Vector Embedding.** The current SOTA approaches that we examine for  $f_\theta$  rely on an embedding of a sequence  $\omega$  that consists of a stack of vectors describing the velocity, acceleration, jerk(time rate of change of acceleration), and bearing rate. In this paper, we based these calculations on prior work (Kim et al. 2022; Dabiri & Heaslip 2018) and included details in the appendix.

**CNN (Dabiri & Heaslip 2018).** Utilizing a convolutional neural network (CNN) presents a viable solution for inferring mobility modes from GPS trajectories, as it can autonomously extract highly efficient features (Dabiri & Heaslip 2018). Here, the CNN incorporates a comprehensive set of layers, including the input layer, convolutional layers, pooling layers, fully-connected layers, and dropout layers.

**LRCN (Kim et al. 2022).** To further enhance the accuracy of extracting mobility modes from GPS trajectories, the application of a Long-term Recurrent Convolutional Network (LRCN) proves beneficial (Kim et al. 2022). The layers of the LRCN model follow a hierarchical structure with three components, proceeding from bottom to top: the convolutional layers, LSTM layers, and fully connected layers.

**LRCN with Attention (new in this paper).** Due to the notable performance improvement transformer architecture (Vaswani et al. 2017) has provided on related problems, we felt it would be important to include a transformer-based approach. Hence, we created a simple extension to LRCN that utilizes attention. We shall refer to this architecture as

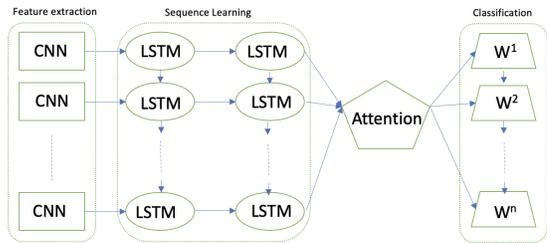


Figure 1: The LRCNa architecture introduced in this paper.

LRCNa. We provide an overview in Figure 1 in the appendix. LRCNa is a neural network architecture comprising several essential components, including convolutional layers employed for feature extraction purposes, LSTM layers, and an attention layer, which collaboratively contribute to sequence learning, and lastly, fully connected layers strategically utilized for effective classification tasks.

### 3 Error Detection and Correction Rules

A key issue with the deployment of model  $f_\theta$  is that it may encounter sequences whose distribution differs from the data used to train the model. Further, in our target application, there may not be sufficient labeled data or time to properly retrain  $f_\theta$ . We also note that in some cases,  $f_\theta$  may be inaccessible for fine-tuning (e.g., behind an API). Additionally, understanding why the results of  $f_\theta$  change is also important for our envisioned security application. As such, we are employing a rule-based approach to correcting  $f_\theta$ . The intuition is that using limited data, we will learn a set of rules (denoted  $\Pi$ ) that will be able to detect and correct errors of  $f_\theta$  by logical reasoning (Aditya et al. 2023). Then, upon deployment for some new sequence  $\omega$ , we would first compute the class  $f_\theta(\omega)$  and then use the rules in set  $\Pi$  to conclude if the result of  $f_\theta$  should be accepted and if not, provide an alternate class in an attempt to correct the mistake. In this section, we formalize the error correcting framework with a simple first order logic (FOL) and provide analytical results relating aspects of learned rules that inform our analytical approach to learning such error detecting and correcting rules. We complete the section with a discussion on how various potential “failure conditions” are extracted to create the rules to correct errors.

Throughout this section, we shall assume a set  $\mathcal{O}$  of operational sequences for which there is ground truth available after model training. The size of set  $\mathcal{O}$  is  $N$  and generally, this is expected to be much smaller than  $\mathcal{T}$  (the set of training data). Later, in our experiments, we look at cases where  $\mathcal{O} = \mathcal{T}$  and  $\mathcal{T} \subseteq \mathcal{O}$  - however these are not requirements as our results are based on model performance on  $\mathcal{O}$  - and we envision use-cases where  $\mathcal{O}$  is significantly different from  $\mathcal{T}$ . On these samples, for each class  $i$ , the model ( $f_\theta$ ) returns class  $i$  for  $N_i$  of the samples, and for each class  $i$  we have the number of true positives, false positives, true negatives, and false negatives  $TP_i, FP_i, TN_i, FN_i$ . We have precision  $P_i = TP_i/N_i$ , recall  $R_i = TP_i/(TP_i + FN_i)$ , and prior of predicting class  $i$ :  $\mathcal{P}_i = N_i/N$ .

**Language.** We assume simple first order language where

samples are represented by constant symbols, and we have unary predicates associated with each sample. This language includes a set  $C$  of  $m$  “condition” predicates  $cond_1, \dots, cond_m$  associated with each sample that can be either true or false for a given sample. Additionally, the language includes the following:

- “Correct” predicates  $corr_1, \dots, corr_i, \dots, corr_n$  which denotes the ground truth class for the sample (i.e., for a given sample one  $corr_i$  will be true and the rest false),
- “Prediction” predicates  $pred_1, \dots, pred_i, \dots, pred_n$  denotes the predicted class for the model (i.e., for a given sample one  $pred_i$  will be true and the rest false)
- “Error” predicates  $error_1, \dots, error_i, \dots, error_n$  if the sample is incorrect for class  $i$ . Note that  $error_i$  is true iff both  $corr_i$  is true and  $pred_i$  is false

**Rules** The set of rules  $\Pi$  will consist of two rules for each class: one “error detecting” and one “error correcting.” Error detecting rules will determine if a prediction by  $f_\theta$  is invalid. In essence, we can think of such a rule as changing the movement class assigned by  $f_\theta$  to some sample  $\omega$  from  $i$  to “unknown.” For a given class  $i$ , we will have an associated set of detection conditions  $DC_i$  that is a subset of conditions, the disjunction of which is used to determine if  $f_\theta$  gave an incorrect classification.

$$error_i(\omega) \leftarrow pred_i(\omega) \wedge \bigvee_{j \in DC_i} cond_j(\omega) \quad (1)$$

After the application of the error detection rules for each class, we may consider re-assigning the samples to another class using a second type of rule called the “corrective rule.” Such rules are formed based on a subset of conditions-class pairs  $CC_i \subseteq C \times C$ .

$$corr_i(\omega) \leftarrow \bigvee_{q,r \in CC_i} (cond_q(\omega) \wedge pred_r(\omega)) \quad (2)$$

Associated with the rules of both types are the following values - both are defined as zero if there are no conditions.

*Support* ( $s$ ): fraction of samples in  $\mathcal{O}$  where the body is true.

*Support w.r.t. class*  $i$  ( $s_i$ ): given the subset of samples where the model predicts class  $i$ , the fraction of those samples where the body is true (note the denominator is  $N_i$ ).

*Confidence* ( $c$ ): the number of times the body and head are true together divided by the number of times the body is true.

Now we present some analytical results that inform our learning algorithms. Our strategy for learning involves first learning detection rules (which establish conditions for which a given classification decision by  $f_\theta$  is deemed incorrect) and then learning correction rules (which then correct the detected errors by assigning a new movement class to the sample). We formalize these two tasks as follows.

*Improvement by error detecting rule.* For a given class  $i$ , find a set of conditions  $DC_i$  such that precision is maximized and recall decreases by, at most  $\epsilon$ .

*Improvement by error correcting rule.* For a given class  $i$ , find a subset  $CC_i$  of  $C \times C$  such that either precision or recall is maximized.

**Properties of Detection Rules.** First, we examine the effect on precision and recall when an error detecting rule is used. Our first result shows a bound on precision improvement. If class support ( $s_i$ ) is less than  $1 - P_i$ , which we would expect (as the rule would be designed to detect the  $1 - P_i$  portion of results that failed), then we can also show that the quantity  $c \cdot s_i$  gives us a lower bound on the improvement in precision. In the appendix, we also note that precision will always increase under a reasonable condition (specifically when  $c \geq 1 - P_i$ ). The proof of this and all other formally stated results can be found in the appendix.

**Theorem 1.** *Under the condition  $s_i \leq 1 - P_i$ , the precision of model  $f_\theta$  for class  $i$ , with initial precision  $P_i$ , after applying an error detecting rule with support  $s_i$  and confidence  $c$  increases by a function of  $s_i$  and  $c$  and is greater than or equal to  $c \cdot s_i$ .*

The error detecting rules can cause the recall to stay the same or decrease. Our next result tells us precisely how much recall will decrease.

**Theorem 2.** *After applying the rule to correct errors, the recall will decrease by  $(1 - c)s_i \frac{R_i}{P_i}$ .*

---

#### Algorithm 1: DetRuleLearn

---

**Require:** Class  $i$ , Recall reduction threshold  $\epsilon$ , Condition set  $C$

**Ensure:** Subset of conditions  $DC_i$

$DC_i := \emptyset$

$DC^* := \{c \in C \text{ s.t. } NEG_{\{c\}} \leq \epsilon \cdot \frac{N_i P_i}{R_i}\}$

**while**  $DC^* \neq \emptyset$  **do**

$c_{best} = \arg \max_{c \in DC^*} POS_{DC_i \cup \{c\}}$

Add  $c_{best}$  to  $DC_i$

$DC^* := \{c \in C \setminus DC_i \text{ s.t. } NEG_{DC_i \cup \{c\}} \leq \epsilon \cdot \frac{N_i P_i}{R_i}\}$

**end while**

**return**  $DC_i$

---

It turns out that both quantities identified in the theorem 1 and theorem 2 are submodular and monotonic - a property we can use algorithmically (formal statements and proofs are included in the appendix). Specifically, we can see that the selection of a set of rules to maximize  $c \cdot s_i$  subject to the constraint that  $(1 - c)s_i \frac{R_i}{P_i} \leq \epsilon$  is a special case of the “Submodular Cost Submodular Knapsack” (SCSK) problem and can be approximated with a simple greedy algorithm (Iyer & Bilmes 2013) with approximation guarantee with polynomial run time (Theorem 4.7 of (Iyer & Bilmes 2013)). Our algorithm DetRuleLearn is an instantiation of such an approach to creating an error detecting rule for a given class. As this algorithm will only select conditions for error detecting rules for a given movement class  $i$  that ensure that recall does not decrease more than epsilon, we can be assured it meets our requirement for recall. Here  $POS_{DC}$ ,  $NEG_{DC}$  are simply the number of samples that satisfy the conditions for some set  $DC$  as well as satisfy  $error_i(\omega)$  (for  $POS_{DC}$ ) and  $corr_i(\omega) \wedge pred_i(\omega)$  (for  $NEG_{DC}$ ) respectively. In other words, given a set of condition class pairs and the rule of interest, BOD here is the number of examples that

satisfy the body (class-condition pair) of the error detection rules, and POS here is the number of examples that satisfy the body (class-condition pair) and the head of the error detection rules.  $P_i, R_i$  are precision and recall for class  $i$  while  $N_i$  is the number of samples that the model classifies as class  $i$ .

**Properties of Corrective Rules.** In what follows, we shall examine the results for corrective rules. Here, the error correcting rule with predicate  $corr_j$  in the head will have a disjunction of elements of set  $CC_i \subseteq C \times \mathcal{C}$ . Also, note that here the support  $s$  is used instead of class support ( $s_i$ ). Here we find that both precision and recall increase with rule confidence (Theorem 3). We also show a corollary that ensures that recall is always non-decreasing for corrective rules and that precision increases when the rule confidence exceeds  $P_i$ .

**Theorem 3.** *For the application of error correcting rules, both precision and recall increase if and only if rule confidence ( $c$ ) increases.*

It is clear that confidence is the right quantity to optimize for error correcting rules as it will get both precision and recall. With these results in mind, we can optimize both precision and recall using an error correcting rule (with respect to the class specified in the rule head) but optimizing for confidence. Note that this does not consider the precision and recall for the class specified in the rule body (however, we shall assume that the impact on precision and recall for the class in the body was handled with the application of the initial error detection rules). However, it is noteworthy that confidence is not monotonic as we add conditions to set  $CC_i$  as the precision can decrease. We will consider an initial set of condition-class pairs  $CC_{all}$  that is a subset of  $C \times \mathcal{C}$ . For a given class for which we create an error correcting rule, we select  $CC_i$  from this larger set. To do so, we adapt the simple ‘‘Deterministic USM’’ algorithm of (Buchbinder et al. 2012) that we call 2. Note here that  $POS_{CC}$  is the number of samples that satisfy the rule body and head ( $corr_i(\omega)$  in this case) given a set of condition-class pairs  $CC$  while  $BOD_{CC}$  is the number of samples that satisfy the body formed with set  $CC$ .

**Learning Detection and Correction Rules Together.** Error correcting rules created using `CorrRuleLearn` will provide optimal improvement to precision and recall for the rule in the target class, but in the case of multi-class problems, it will cause recall to drop for some other classes. However, we can combine both error detecting and correcting rules to overcome this difficulty. The intuition is first to create error detecting rules for each class, which effectively re-assigns any sample into an ‘‘unknown’’ class. Then, we create a set  $CC_{all}$  (used as input for `CorrRuleLearn`) based on the conditions selected by the error detecting rules. In this way, we will not decrease recall beyond what occurs in the application of error detecting rules.

### Conditions for Error Detection and Correction

In this section, we describe the methods we used to create conditions (set  $C$ ) from dataset  $\mathcal{O}$ . As mentioned in section ??, in addition to offering domain-specific knowledge,

---

#### Algorithm 2: CorrRuleLearn

---

**Require:** Class  $i$ , Set of condition-class pairs  $CC_{all}$

**Ensure:** Subset of condition-class pairs  $CC_i$

```

 $CC_i := \emptyset$ 
 $CC'_i := CC_{all}$ 
Sort each  $(c, j) \in CC_{all}$  from greatest to least by
 $\frac{POS_{\{(c,j)\}}}{BOD_{\{(c,j)\}}}$  and remove  $\frac{POS_{\{(c,j)\}}}{BOD_{\{(c,j)\}}} \leq P_i$ 
for  $(c, j) \in CC_{all}$  selected in order of the sorted list do
   $a := \frac{POS_{CC_i \cup \{(c,j)\}}}{BOD_{CC_i \cup \{(c,j)\}}} - \frac{POS_{CC_i}}{BOD_{CC_i}}$ 
   $b := \frac{POS_{CC'_i \setminus \{(c,j)\}}}{BOD_{CC'_i \setminus \{(c,j)\}}} - \frac{POS_{CC'_i}}{BOD_{CC'_i}}$ 
  if  $a \geq b$  then
     $CC_i := CC_i \cup \{(c, j)\}$ 
  else
     $CC'_i := CC'_i \setminus \{(c, j)\}$ 
  end if
end for
if  $\frac{POS_{CC_i}}{BOD_{CC_i}} \leq P_i$  then
   $CC_i := \emptyset$ 
end if
return  $CC_i$ 

```

---



---

#### Algorithm 3: DetCorrRuleLearn

---

**Require:** Recall reduction threshold  $\epsilon$ , Condition set  $C$

**Ensure:** Set of rules  $\Pi$

```

 $\Pi := \emptyset$ 
 $CC_{all} := \emptyset$ 
for Each class  $i$  do
   $DC_i := \text{DetRuleLearn}(i, \epsilon, C)$ 
  if  $DC_i \neq \emptyset$  then
     $\Pi := \Pi \cup \{error_i(\omega) \leftarrow pred_i(\omega) \wedge \bigvee_{j \in DC_i} cond_j(\omega)\}$ 
  end if
  for  $cond \in DC_i$  do
     $CC_{all} := CC_{all} \cup \{(cond, i)\}$ 
  end for
end for
for Each class  $i$  do
   $CC_i := \text{CorrRuleLearn}(i, CC_{all})$ 
  if  $CC_i \neq \emptyset$  then
     $\Pi := \Pi \cup \{corr_i(\omega) \leftarrow \bigvee_{q,r \in CC_i} (cond_q(\omega) \wedge pred_r(\omega))\}$ 
  end if
end for
return  $\Pi$ 

```

---

our contribution extends to the provision of a condition integrated with a neural network, referred to as the model based in our paper. This condition, marked by its comprehensive generality, serves to amplify the adaptability of the EDCR across a spectrum of diverse problem domains.

**Model Based** The field of Deep Learning witnesses a continuous influx of new and improved models for solving complex problems. The prevailing trend involves the adoption of the latest and supposedly superior models, often leading to the abandonment of previously successful ones. We present a method that challenges this paradigm, proposing a technique to harness the potential of older, proven models to augment the performance of the latest and most advanced models. We employ a collection of diverse pre-existing neural models as a set of conditions to enhance the efficacy of the current model. More specifically, a more coarse-grain model can also provide insight into the conditions. As such, we utilized a binary classifier for each class for a given sample. Hence, given class  $i$ , we have a binary classifier  $g_i$  which returns “true” for sample  $\omega$  if  $g_i$  assigns it as  $i$  and “false” otherwise. In this way, for each sample  $\omega$  we have a  $g_i(\omega)$  condition for each of the classes. We used the LRCNa architecture for the binary classifier and the details are in the appendix.

**Domain Knowledge** Harnessing domain expertise in outlier analysis can yield valuable insights and conditions. Specifically, our attention was drawn to the maximum velocity records within our dataset. Consequently, for each class denoted as  $i$ , we formulated a set of conditions encapsulated by  $s_i$ , each of which is linked to the maximum velocity criterion. So, for a given sample  $\omega$ ,  $s_i(\omega)$  is true if the velocity for  $\omega$  is greater than the maximum velocity observed in set  $\mathcal{O}$  and false otherwise.

## 4 Experimental Evaluation

**GeoLife Dataset.** The proposed methodology is validated and assessed using GPS trajectories obtained from the GeoLife project, which involved data collected from 69 users (Zheng et al. 2008). Details on the preprocessing of the data can be found in the appendix.

**Training and Test Splits.** Previous work such as (Kim et al. 2022) is known to have data leakage based on the split between training and test primarily due to segments of a movement sequence existing in both training and test sets resulting from ransom assignment to each. To address this data leakage issue, we examine our algorithms under various conditions based on ordering and overlap. For ordering, we examine random (which can allow previous behavior of the same agent in the training set, as in previous work) and sequential (which orders the agents to avoid this issue). For overlap, we examine no overlap between the training and test sets, segment overlap that allows training and test samples to overlap each other(as in previous work), and data point overlap (that allows for data points of a trajectory to span both training and test).

**Compute and Implementation.** All experiments were performed on a 2000 MHz AMD EPYC 7713 CPU, and a NVIDIA GA100 GPU using Python 3.10 with PyTorch.

**All Classes Observed.** In our first set of experiments, we examined how error detecting and correcting rules (EDCR) can affect the performance of the underlying model. In Table 1 we examine the accuracy of each model, both with and without EDCR. Models enabled with EDCR performed the same or better with improvement most noticeable when samples are sequential (which has less data leakage between training and test). In terms of overall performance, LRCNa with EDCR performed the best in five of six cases with LRCN with EDCR performing the best in the sixth. Of particular importance, in the “no overlap - sequential” case - the least likely to exhibit data leakage - EDCR improves the performance of both LRCNa and LRCN, 1.6% and 1.7% respectively. Additionally, we scrutinized the F1 scores in Table 2 for all models, both with and without EDCR, revealing more improvement in performance metrics compared to accuracy.

**Hyperparameter Sensitivity.** In the “all classes observed” set of experiments, we also examined hyperparameter sensitivity for  $\epsilon$ . Recall that  $\epsilon$  is interpreted as the maximum decrease in recall. We observed and validated the theoretical reduction(TR) in recall empirically and the experiments show us that in all cases, recall was no lower than the threshold specified by the hyperparameter  $\epsilon$  though recall decreases as  $\epsilon$  increases. In many cases, the experimental evaluation reduced recall significantly less than expected. In Figure 2, as the value of  $\epsilon$  (x-axis) ranges from 0 to 0.10, it is evident that the decline in recall for all classes remains within the confines of 0.10. Likewise, precision only increases with  $\epsilon$ , which is aligned with our theoretical results. We show precision, recall, and F1 by class for the “no overlap - sequential” of LRCNa in Figure 2. Though the algorithm DetCorrRuleLearn calls for a single  $\epsilon$  hyperparameter, it is possible to set it differently for each class (e.g., lower values for classes where recall is important, higher values for classes where false positives are expensive). This may be beneficial as F1 for different classes seemed to peak for different values of  $\epsilon$ . We leave the study of heterogeneous  $\epsilon$  settings to future work.

**Removal of Movement Classes from Training.** Our experimental focus was on assessing how the introduction of EDCR impacts model performance in scenarios where certain movement classes are excluded from training. In Figure 3, we trained the CNN, LRCN, and LRCNa models without incorporating the walk and drive classes. Remarkably, employing EDCR without any supplementary data yielded a 5.2%(zero-shot) improvement over the base models, and a 23.9% (few-shot) improvement over the SOTA model without resorting to retraining of the base model, with even more pronounced results than in the initial experiment set. Utilizing a mere 30% of data from previously unseen classes, EDCR demonstrates a 21.3% to elevate the performance of the baseline model, all achieved without the need for direct access to the model itself. This outcome implies the potential for conducting few-shot learning, enabling the adaptation of  $f_\theta$  to novel scenarios with impressive efficacy. This enhancement significantly boosts accuracy using limited data for unseen samples, without extensive model modifications. This

	No Overlap		Segment Overlap		Data point Overlap	
	Random	Sequential (least leakage)	Random (prev. studies)	Sequential	Random	Sequential
LRCNa (ours)	0.747	0.751	<b>0.971</b>	0.758	<b>0.921</b>	0.760
LRCNa+EDCR (ours)	0.759 (+1.6%)	<b>0.763</b> (+1.6%)	<b>0.971</b> ( $\pm 0\%$ )	<b>0.769</b> (+1.5%)	<b>0.921</b> ( $\pm 0\%$ )	<b>0.780</b> (+2.6%)
LRCN (prev. SOTA)	0.749	0.747	0.952	0.767	0.887	0.774
LRCN+EDCR (ours)	<b>0.761</b> (+1.6%)	0.760 (+1.7%)	0.952 ( $\pm 0\%$ )	0.768 (+0.1%)	0.889 (+0.2%)	0.783 (+1.1%)
CNN	0.742	0.755	0.851	0.763	0.853	0.779
CNN+EDCR (ours)	0.743 (+0.1%)	0.755 ( $\pm 0\%$ )	0.866 (+1.8%)	0.763 ( $\pm 0\%$ )	0.862 (+1.0%)	0.779 ( $\pm 0\%$ )

Table 1: Accuracy when all classes are represented in training and test sets under various data leakage cases. EDCR means “error detecting and correcting rules” were used on the model output and numbers in parens show the percent change in accuracy from EDCR over the base model. Bold numbers are the best in each case.

	No Overlap		Segment Overlap		Data point Overlap	
	Random	Sequential (least leakage)	Random (prev. studies)	Sequential	Random	Sequential
LRCNa (ours)	0.727	0.734	<b>0.971</b>	0.742	<b>0.906</b>	0.715
LRCNa+EDCR (ours)	0.742 (+2.06%)	<b>0.751</b> (+2.32%)	<b>0.971</b> ( $\pm 0\%$ )	0.757 (+2.02%)	<b>0.906</b> ( $\pm 0\%$ )	0.749 (+4.76%)
LRCN (prev. SOTA)	0.732	0.738	0.951	0.751	0.864	0.737
LRCN+EDCR (ours)	<b>0.75</b> (+2.46%)	0.741 (+0.41%)	0.951 ( $\pm 0\%$ )	<b>0.76</b> (+1.2%)	0.864 (+0%)	<b>0.755</b> (+2.44%)
CNN	0.722	0.737	0.846	0.745	0.826	0.748
CNN+EDCR (ours)	0.723 (+0.14%)	0.737 ( $\pm 0\%$ )	0.866 (+2.36%)	0.745 ( $\pm 0\%$ )	0.83 (+0.48%)	0.748 ( $\pm 0\%$ )

Table 2: macro F1 when all classes are represented in training and test sets under various data leakage cases. EDCR means “error detecting and correcting rules” were used on the model output and numbers in parens show the percent change in macro F1 from EDCR over the base model. Bold numbers are the best in each case.

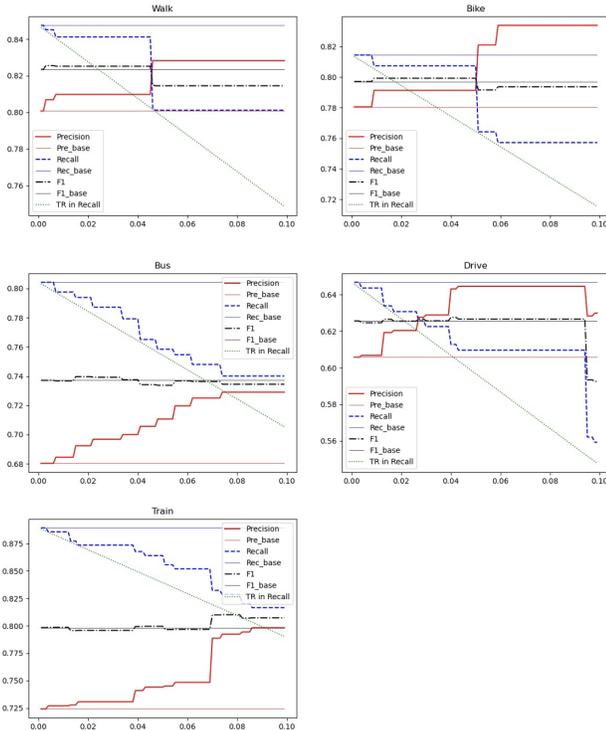


Figure 2: LRCNa Results for application of error detection and correction rules as a function of  $\epsilon$  (no overlaps with sequential selection). TR in Recall is the theoretical reduction in recall based on analytic results.

is crucial when direct model access is limited, for example through an API.

## 5 Related Work and Conclusion

As described earlier, the MTCP problem was previously studied in (Dabiri & Heaslip 2018; Kim et al. 2022), which introduces the LRCN and CNN architectures, respectively. Earlier work has also explored this problem with other machine learning approaches (Zheng et al. 2008; Wang et al. 2017; Simoncini et al. 2018). Note that error detection and correction have not previously been explored in these earlier works. Also note that both this prior work and this paper differ from trajectory generation (Janner et al. 2021; Chen et al. 2021; Itkina & Kochenderfer 2022) - which differs from trajectory classification.

Earlier work on machine learning introspection (Daftary et al. 2016; Ramanagopal et al. 2018) examined error detection on various perceptual models. Unlike this work, these approaches were not applied to the MTCP, only focused on error detection, and did not provide theoretical guarantees of improvement. Another area of related work is machine learning verification that (Ivanov et al. 2021; Jothimurugan et al. 2021; Ma et al. 2020)) that looks to ensure the output of an ML model meets a logical specification. Like our work, some of these contributions (e.g. (Ma et al. 2020)) adjust the output of a machine learning model to meet a logic-based specification. However, to our knowledge, there has been no work on the use of machine learning verification to correct a machine learning model as this work does. Other related areas include meta-learning and domain generalization (Hospedales et al. 2021; Zhou et al. 2022; Vanschoren 2018; Maes & Nardi 1988) which attempt to account for changes in the distribution of data and/or selec-

tion of a model that was trained on data similar to the current problem. While our approach can use additional data, it does not depend on training data generated by different distributions. To our knowledge, these other methods have not been applied to MTCP. Recent studies on abductive learning (Huang et al. 2023; Dai et al. 2019) and neural symbolic reasoning (Cornelio et al. 2022) incorporate error correction mechanisms rooted in inconsistency with domain knowledge as logical rules. These approaches typically necessitate direct access to the perceptual model for effective implementation. In contrast, our work takes a distinct approach by avoiding reliance on predefined learning rule pairs and eliminating the need for direct access to the perceptual model. We conjecture that these approaches could be complementary to EDCR, and we leave it to future work to explore how they can work together.

**Conclusion.** A key near-term direction for future work is the employment of these methods in government-administered tests of the IARPA HAYSTAC program which will provide an assessment of utility more closely related to real-world use cases. Likewise, an extension related to the aforementioned IARPA program would be to identify a sequence of movement classes in the case where an agent’s mode of transit may change. For example, Here we would look to apply our error detection and correction framework to recently introduced models such as those described in (Zeng et al. 2023). Separately, we framed rule learning as a pair of sub-modular maximization problems, but there are several options for algorithms beyond this paper. Finally, the use of rules for error detection and correction of machine learning models presented here may be useful in domains such as vision.

## 6 Acknowledgments

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0032. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. Additionally, some of the authors are supported by ONR grant N00014-23-1-2580 as well as internal funding from ASU Fulton Schools of Engineering.

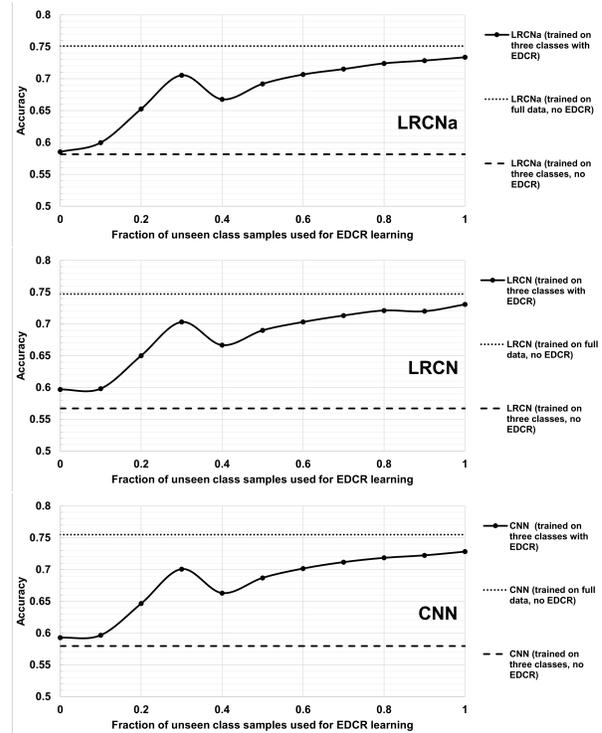


Figure 3: Results for experiments with two movement classes removed from training.

## References

- Aditya, D., Mukherji, K., Balasubramanian, S., Chaudhary, A., and Shakarian, P. PyReason: Software for open world temporal logic. In *AAAI Spring Symposium, 2023*.
- Buchbinder, N., Feldman, M., Naor, J., and Schwartz, R. A tight linear time (1/2)-approximation for unconstrained sub-modular maximization. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 649–658, 2012. doi: 10.1109/FOCS.2012.73.
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *CoRR*, abs/2106.01345, 2021. URL <https://arxiv.org/abs/2106.01345>.
- Cornelio, C., Stuehmer, J., Hu, S. X., and Hospedales, T. Learning where and when to reason in neuro-symbolic inference. In *The Eleventh International Conference on Learning Representations, 2022*.
- Dabiri, S. and Heaslip, K. Inferring transportation modes from gps trajectories using a convolutional neural network. *Transportation research part C: emerging technologies*, 86: 360–371, 2018.
- Daftry, S., Zeng, S., Bagnell, J. A., and Hebert, M. Introspective perception: Learning to predict failures in vision systems, 2016. URL <http://arxiv.org/abs/1607.08665>.
- Dai, W.-Z., Xu, Q., Yu, Y., and Zhou, Z.-H. Bridging machine learning and logical reasoning by abductive learning. *Advances in Neural Information Processing Systems*, 32, 2019.

- Fikioris, G., Patroumpas, K., Artikis, A., Pitsikalis, M., and Paliouras, G. Optimizing vessel trajectory compression for maritime situational awareness. *GeoInformatica*, 27(3): 565–591, 2023.
- Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.
- Huang, H., Cheng, Y., and Weibel, R. Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, 101:297–312, 2019.
- Huang, Y.-X., Dai, W.-Z., Jiang, Y., and Zhou, Z.-H. Enabling knowledge refinement upon new concepts in abductive learning. 2023.
- Itkina, M. and Kochenderfer, M. J. Interpretable self-aware neural networks for robust trajectory prediction, 2022.
- Ivanov, R., Carpenter, T., Weimer, J., Alur, R., Pappas, G., and Lee, I. Verisig 2.0: Verification of neural network controllers using taylor model preconditioning. In *Computer Aided Verification: 33rd International Conference, CAV 2021, Virtual Event, July 20–23, 2021, Proceedings, Part I*, pp. 249–262. Springer-Verlag, 2021. ISBN 978-3-030-81684-1. doi: 10.1007/978-3-030-81685-8\_11. URL [https://doi.org/10.1007/978-3-030-81685-8\\_11](https://doi.org/10.1007/978-3-030-81685-8_11).
- Iyer, R. and Bilmes, J. Submodular optimization with submodular cover and submodular knapsack constraints. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, pp. 2436–2444, Red Hook, NY, USA, 2013. Curran Associates Inc.
- Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, 2021.
- Jothimurugan, K., Bansal, S., Bastani, O., and Alur, R. Compositional reinforcement learning from logical specifications. In *Advances in Neural Information Processing Systems*, 2021.
- Kim, J., Kim, J. H., and Lee, G. Gps data-based mobility mode inference model using long-term recurrent convolutional networks. *Transportation Research Part C: Emerging Technologies*, 135:103523, 2022.
- Lin, M. and Hsu, W.-J. Mining gps data for mobility patterns: A survey. *Pervasive and mobile computing*, 12:1–16, 2014.
- Ma, M., Gao, J., Feng, L., and Stankovic, J. Stlnet: Signal temporal logic enforced multivariate recurrent neural networks. *Advances in Neural Information Processing Systems*, 33:14604–14614, 2020.
- Maes, P. and Nardi, D. Meta-level architectures and reflection. 1988.
- Ramanagopal, M. S., Anderson, C., Vasudevan, R., and Johnson-Roberson, M. Failing to learn: Autonomously identifying perception failures for self-driving cars. 3(4):3860–3867, 2018. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2018.2857402. URL <http://arxiv.org/abs/1707.00051>.
- Simoncini, M., Taccari, L., Sambo, F., Bravi, L., Salti, S., and Lori, A. Vehicle classification from low-frequency gps data with recurrent neural networks. *Transportation Research Part C: Emerging Technologies*, 91:176–191, 2018.
- Vanschoren, J. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- Vincenty, T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93, 1975.
- Wang, H., Liu, G., Duan, J., and Zhang, L. Detecting transportation modes using deep neural network. *IEICE TRANSACTIONS on Information and Systems*, 100(5):1132–1135, 2017.
- Zeng, J., Yu, Y., Chen, Y., Yang, D., Zhang, L., and Wang, D. Trajectory-as-a-sequence: A novel travel mode identification framework. 146:103957, 2023. ISSN 0968-090X. doi: <https://doi.org/10.1016/j.trc.2022.103957>. URL <https://www.sciencedirect.com/science/article/pii/S0968090X22003709>.
- Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 312–321, 2008.
- Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

## A Appendix

### Details on Vector Embedding of Sequences

We begin with a set of GPS points where each point is a tuple of timestamp ( $t$ ), latitude ( $lat$ ), and longitude ( $long$ ),  $P_i = (t_i, P_i^{lat}, P_i^{long})$ . Each point  $P_i$  also has an associated class label  $c \in \mathcal{C}$ . To embed these tuples as vector embeddings that can be consumed by the neural model  $f_\theta$ , three essential preprocessing steps must be performed. These steps include normalizing the data size to meet the input requirements, extracting movement behaviors from the GPS points, and refining the data. In this section, we draw upon previous approaches (Zheng et al., 2008a,b; Dabiri et al., 2018; Kim, 2022) to guide the data preprocessing process.

As part of the data size normalization step we sequentially group chronologically ordered GPS points into uniform lengths of 40. The class label  $c$  of every point in this sequence is the same and the entire sequence represents the movement trajectory of that class for 40 time units. The resulting sequence  $\omega \in S$ , where  $S$  is the set of all sequences that are curated.

To capture patterns of movement behaviors from GPS points the distance time-series vector is computed as follows.  $D_i^j$  is the distance between two GPS point tuples  $P_i^j$  and  $P_{i-1}^j$ , where  $j \in S$  and  $i \in \omega$ , and is computed using the Vincenty Distance formula (Vincenty 1975). Here  $D_{10}^3$  represents the distance between two points  $P_{10}$  and  $P_9$  from the  $3^{rd}$  sequence. There could be cases where a distance time-series vector falls short of 40 data points. To maintain a consistent length of sequence  $\omega$  we pad the shorter  $D^j$  vector with zeros.

Additionally, we extract the velocity ( $V$ ), acceleration ( $A$ ), jerk ( $J$ ) and bearing rate ( $BR$ ) time-series vectors for each sequence as follows:

$$V_i^j = \frac{\text{Vincenty}(P_{i-1}^j, P_i^j)}{t_i - t_{i-1}} \quad (1)$$

$$A_i^j = \frac{V_i^j - V_{i-1}^j}{t_i - t_{i-1}} \quad (2)$$

$$J_i^j = \frac{A_i^j - A_{i-1}^j}{t_i - t_{i-1}} \quad (3)$$

$$BR_i^j = | \text{Bearing}_i - \text{Bearing}_{i-1} | \quad (4)$$

$$\text{where Bearing}_i = \arctan(y, x) \quad (5)$$

$$y = \sin(P_i^{long} - P_{i-1}^{long}) * \cos(P_i^{lat}) \quad (6)$$

$$x = \cos(P_{i-1}^{lat}) * \sin(P_i^{lat}) - \sin(P_{i-1}^{lat}) * \quad (7)$$

$$\cos(P_i^{lat}) * \cos(P_i^{long} - P_{i-1}^{long})$$

We finally stack the vectors  $V^j$ ,  $A^j$ ,  $J^j$  and  $BR^j$  for each sequence  $\omega$ , which is passed as the input to the neural model  $f_\theta$  as detailed in section 2.

### Formal Statements of Additional Theorems Corollaries for Error Detection Rules

**Corollary 1.** *If and only if  $c \geq 1 - P_i$  then the rule will cause precision not to decrease.*

**Corollary 2.** *If  $P_i \geq 1 - c$  (the minimum condition for precision improvement from Corollary 1) then recall decreases by at most  $s_i R_i$ .*

**Theorem 4.** *For a given error detecting rule, the quantity  $c \cdot s_i$  is a normalized polymatroid function w.r.t. set  $DC$ .*

**Corollary 3.** *The quantity  $(1 - c)s_i \frac{R_i}{R_i}$  (decrease in recall) is a normalized polymatroid function w.r.t. set  $DC$ .*

**Corollary 4.** *GreedyRuleSelect provides an approximation of  $cs$  that is within  $1/|C|$  of optimal.*

### Formal Statements of Additional Theorems Corollaries for Error Correction Rules

**Corollary 5.** *Precision increases for class  $i$  with the application of an error correcting rule if and only if  $c > P_i$ .*

**Corollary 6.** *Recall is non-decreasing for class  $i$  with the application of an error correcting rule.*

**Theorem 5.** *Confidence is submodular with respect to  $CC_i$ .*

**Corollary 7.** *For an arbitrarily small constant  $\epsilon$ , DetUSM-PosRuleSelect provides a  $1/3 + \epsilon$  approximation of confidence if the returned confidence is greater than the initial precision.*

### Proof of Theorem 1

Under the condition  $s_i \leq 1 - P_i$ , the precision of model  $f_\theta$  for class  $i$ , with initial precision  $P_i$ , after applying an error correcting rule with support  $s_i$  and confidence  $c$  increases by a function of  $s_i$  and  $c$  and is greater than or equal to  $c \cdot s_i$ .

*Proof.* CLAIM 1: The precision of model  $f_\theta$  for class  $i$ , with initial precision  $P_i$ , after applying an error correcting rule with support  $s_i$  and confidence  $c$  increases by:

$$\frac{s_i}{1 - s_i}(c + P_i - 1) \quad (8)$$

The total number of items that  $f_\theta$  will attempt to classify as  $i$  before error correction is  $N_i = TP_i + FP_i$ . Out of those,  $s_i \cdot N_i$  will be corrected by the rule. However, a fraction of  $(1 - c)$  will be samples that would have been true positives if not corrected. Hence, the new precision can be written as follows:

$$\frac{TP_i - (1 - c)s_i \cdot N_i}{N_i - s_i \cdot N_i} \quad (9)$$

As  $P_i \cdot N_i = TP_i$ , we have:

$$\frac{P_i \cdot N_i - (1 - c)s_i \cdot N_i}{N_i(1 - s_i)} \quad (10)$$

$$= \frac{P_i - (1 - c)s_i}{(1 - s_i)} \quad (11)$$

Now we subtract from that quantity the initial precision.

$$\frac{P_i - (1-c)s_i}{1-s_i} - P_i \quad (12)$$

$$= \frac{P_i - (1-c)s_i}{1-s_i} - \frac{(1-s_i)P_i}{1-s_i} \quad (13)$$

$$= \frac{-s_i + s_i c + P_i s_i}{1-s_i} \quad (14)$$

$$= \frac{s_i}{1-s_i} (c + P_i - 1) \quad (15)$$

CLAIM 2: If  $s_i \leq 1 - P_i$  then  $c \cdot s_i$  is a lower bound on the improvement in precision.

BWOC, then by Claim 1 we have.

$$\frac{s_i}{1-s_i} (c + P_i - 1) < c \cdot s_i \quad (16)$$

$$c + P_i - 1 < c(1-s_i) \quad (17)$$

$$c + P_i - 1 < c - c \cdot s_i \quad (18)$$

$$c \cdot s_i < 1 - P_i \quad (19)$$

$$c \cdot s_i < s_i \quad (20)$$

However, as  $c \leq 1$  this is a contradiction.

The proof of the theorem then follows directly from claim 2.  $\square$

### Proof of Corollary 1

If and only if  $c \geq 1 - P_i$  then the rule will cause precision not to decrease.

*Proof.* Suppose, BWOC, the statement is false. By Theorem 1 then the following must be true.

$$\frac{P_i - s_i(1-c)}{1-s_i} - P_i < 0 \quad (21)$$

$$P_i - s_i(1-c) < P(1-s_i) \quad (22)$$

$$s_i c - s_i < -P_i s_i \quad (23)$$

$$P_i < 1 - c \quad (24)$$

However, as  $P_i \geq 1 - c$  this cannot hold.

Likewise, suppose BWOC that  $c < 1 - P_i$  and BWOC the statement is false:

$$\frac{P_i - s_i(1-c)}{1-s_i} - P_i > 0 \quad (25)$$

$$P_i - s_i(1-c) > P(1-s_i) \quad (26)$$

$$s_i c - s_i > -P_i s_i \quad (27)$$

$$P_i > 1 - c \quad (28)$$

Again, a contradiction.  $\square$

### Proof of Theorem 4

For a given error detecting rule, the quantity  $c \cdot s_i$  is a normalized polymatroid function w.r.t. set  $DC$ .

*Proof.* CLAIM 1:  $c \cdot s_i = POS/N_i$  where  $POS$  is the number of samples where both the rule body and head are satisfied.

Let  $BOD$  be the number of samples that the body of the rule

is true. This gives us  $c \cdot s_i = \frac{POS}{BOD} \frac{BOD}{N_i}$  which is equivalent to the statement of the claim. CLAIM 2: The quantity  $c \cdot s_i$  is submodular w.r.t. set  $DC$ .

We show this by the submodularity of  $POS$  as  $N_i$  is a constant as well as the result of Claim 1. BWOC,  $POS$  is not submodular for some set  $DC$ . We use the symbol  $POS(DC)$  to denote this and assume the existence of two sets of conditions  $DC_1, DC_2$ . Then, the following must be true:

$$POS(DC_1) + POS(DC_2) < POS(DC_1 \cup DC_2) \quad (29)$$

Which can be re-written as:

$$\left| \bigcup_{cond \in DC_1} \{x | cond(\omega) \wedge pred(\omega)\} \right| + \quad (30)$$

$$\left| \bigcup_{cond \in DC_2} \{x | cond(\omega) \wedge pred(\omega)\} \right| \quad (31)$$

This quantity is less than the following:

$$\left| \bigcup_{cond \in DC_1 \cup DC_2} \{x | cond(\omega) \wedge pred_x\} \right| \quad (32)$$

However, this would imply there is at least one element in  $DC_1 \cup DC_2$  not in either  $DC_1$  or  $DC_2$  which is a contradiction. CLAIM 3:  $c \cdot s_i$  monotonically increases with  $DC$ .

By claim 1, as the quantity equals  $POS/N_i$  and  $N_i$  is a constant, we just need to show monotonicity of  $POS$ . Clearly  $POS$  increases monotonically as additional elements in  $DC$  can only make it increase. CLAIM 4: When  $DC = \emptyset$ ,  $c \cdot s_i = 0$ .

Follows directly from the fact that we define  $s_i$  as zero is no conditions are used.

Proof of theorem. Follows directly from claims 2-4.  $\square$

### Proof of Theorem 2

After applying the rule to correct errors, the recall will decrease by

$$(1-c)s_i \frac{R_i}{P_i} \quad (33)$$

*Proof.* The number of corrections made by the rule is  $s_i(TP_i + FP_i)$  with  $(1-c)$  fraction of these being incorrect (increasing false negatives). Note that the sum  $TP_i + FN$  does not change after error correction, as any "corrected" false positive becomes a false negative, and false negatives do not otherwise change from error correction. Therefore, the new recall is:

$$\frac{TP_i - s(1-c)(TP_i + FP_i)}{TP_i + FN_i} \quad (34)$$

When this quantity is subtracted from the original recall ( $R_i$ ), we obtain:

$$s_i(1-c) \left( R_i + \frac{FP_i}{TP_i + FN_i} \right) \quad (35)$$

$$(36)$$

We note that  $FP_i = \frac{TP_i}{P_i} - TP_i = \frac{TP_i - P_i \cdot TP_i}{P_i}$  which gives us:

$$s_i(1-c) \left( R_i + \frac{TP_i}{P(TP_i + FN_i)} - \frac{TP_i \cdot P_i}{P_i(TP_i + FN_i)} \right) \quad (37)$$

$$= s_i(1-c) \left( R_i + \frac{R_i}{P_i} - R_i \right) \quad (38)$$

$$= (1-c)s_i \frac{R_i}{P_i} \quad (39)$$

□

### Proof of Corollary 2

If  $P_i \geq 1-c$  (the minimum condition for precision improvement from Corollary 1 then recall decreases by at most  $s_i R_i$ .

*Proof.* Suppose BWOC the statement is false. By Theorem 2, recall decrease by  $(1-c)s_i \frac{R_i}{P_i}$ . This gives us:

$$(1-c)s_i \frac{R_i}{P_i} > s_i R_i \quad (40)$$

Precision cannot be less than  $1-c$ , so recall must then decrease by:

$$(1-c)s_i \frac{R_i}{1-c} > s_i R_i \quad (41)$$

$$s_i R_i > s_i R_i \quad (42)$$

□

### Proof of Corollary 3

The quantity  $(1-c)s_i \frac{R_i}{P_i}$  (decrease in recall) is a normalized polymatroid function w.r.t. set  $DC$ .

*Proof.* Note that  $BOD$  is the number of samples that satisfy the body, while  $POS$  is the number of samples that satisfy the body and head,  $NEG = POS - BOD$ .

$$(1-c)s_i \frac{R_i}{P_i} = \left(1 - \frac{POS}{BOD}\right) \frac{BOD}{N_i} \frac{R_i}{P_i} \quad (43)$$

$$= \frac{NEG}{BOD} \frac{BOD}{N} \frac{R_i}{P_i} \quad (44)$$

$$= NEG \frac{1}{N_i} \frac{R_i}{P_i} \quad (45)$$

As  $\frac{1}{N_i} \frac{R_i}{P_i}$  is a constant, we need to show the submodularity of  $NEG$  which follows the same argument for  $POS$  as per Claim 2 of Theorem 4. Likewise,  $NEG$  is monotonic (mirroring the argument of Claim 3 of Theorem 4) and normalized by the definition of  $s_i$  in the case where there are no conditions. The statement of the theorem follows. □

### Proof of Theorem 3

For the application of positive rules, precision increases if and only if rule confidence ( $c$ ) increases.

*Proof.* CLAIM 1: Precision increases by  $\frac{cs-P_i s}{P_i+s}$ .

The new precision is equal to the following:

$$\frac{TP_i + csN}{M_i + sN} \quad (46)$$

The improvement of the precision can be derived as follows.

$$\frac{TP_i + csN}{M_i + sN} - P_i = \quad (47)$$

$$= \frac{TP_i + csN - P_i M_i - P_i sN}{M_i + sN} \quad (48)$$

$$= \frac{TP_i + csN - TP_i - P_i sN}{M_i + sN} \quad (49)$$

$$= \frac{csN - P_i sN}{M_i + sN} \quad (50)$$

$$= \frac{cs - P_i s}{P_i + s} \quad (51)$$

CLAIM 2: If count of samples satisfying both rule body and head (the numerator of confidence) increases, then precision increases.

Suppose BWOC the claim is not true. Then for some value of  $POS$  for which the improvement in precision is greater than  $POS' = POS + 1$ . Note that, in this case, the number of samples satisfying the body also increases by 1. First, we know that we can re-write the result of claim 1 as follows.

$$\frac{POS - P_i BOD}{M_i + BOD} \quad (52)$$

Therefore, using the result from Claim 1, the following relationship must hold.

$$\frac{POS - P_i BOD}{M_i + BOD} > \frac{POS + 1 - P_i BOD - P_i}{M_i + BOD + 1} \quad (53)$$

$$POS - P_i BOD > M_i(1 - P_i) + BOD(1 - P_i) \quad (54)$$

$$POS > M(1 - P_i) + BOD \quad (55)$$

This gives us a contradiction, as  $M(1 - P_i) \geq 0$  and  $POS \leq BOD$  by definition.

CLAIM 3: If the difference in precision increases, the number of samples satisfying both rule body and head must increase.

By definition, the only way for this to occur is if  $BOD$  increases and  $POS$  does not - as they can both increase or only  $BOD$  increase. If neither there is no change, and it is not possible for  $POS$  to increase without  $BOD$ . Therefore the following must be true.

$$\frac{POS - P_i BOD}{M_i + BOD} < \frac{POS - P_i BOD - P_i}{M_i + BOD + 1} \quad (56)$$

However, this is clearly a contradiction the expression on the right is clearly smaller (the numerator is smaller as  $P_i$  is positive, and the denominator is larger).

CLAIM 4: Precision increases if and only if  $c$  increases.

Follows directly from claims 1-3.

CLAIM 5: When adding more samples that satisfy the body of the rule, confidence increases if and only if  $POS$  increases.

Note that confidence is defined as  $POS/BOD$ . Clearly, there confidence decreases if  $BOD$  increases but not  $POS$  and it is not possible for  $POS$  to increase alone. Therefore, BWOC, the following must hold true.

$$\frac{POS + k}{BOD + k} < \frac{POS}{BOD} \quad (57)$$

$$BODk < POSk \quad (58)$$

$$BOD < POS \quad (59)$$

This is a contradiction as  $BOD \geq POS$ .

Going other way, suppose BWOC confidence increases but POS does not. We get:

$$c_2 > c_1 \quad (60)$$

$$\frac{POS}{BOD_2} > \frac{POS}{BOD_1} \quad (61)$$

$$BOD_1 > BOD_2 \quad (62)$$

However, by the statement, as we add more samples that satisfy the body of the rule, we must have  $BOD_1 \leq BOD_2$ . Hence a contradiction.

CLAIM 6: Recall increases if and only if  $POS$  increases.

As we can write the new recall in this case simply as the following, the claim immediately follows.

$$\frac{TP_i + POS}{TP_i + FN_i} \quad (63)$$

CLAIM 7: Recall increases if and only if  $c$  increases.

Follows directly from claims 5-6.

Proof of theorem.

Follows directly from claims 4 and 7.  $\square$

### Proof of Corollary 4

GreedyRuleSelect provides an approximation of  $c_s$  that is within  $1/|C|$  of optimal.

*Proof.* Follows directly from Theorem 4.7 of (Iyer & Bilmes 2013).  $\square$

### Proof of Corollary 7

For an arbitrarily small constant  $\epsilon$ , DetUSMPosRuleSelect provides a  $1/3 + \epsilon$  approximation of confidence if the returned confidence is greater than the initial precision.

*Proof.* Follows directly from the fact that confidence is zero when  $CC_i = \emptyset$  and Theorem 2.3 of (Buchbinder et al. 2012).  $\square$

### Conditions for Error Detection and Correction

This section describes the various methods we used to create conditions (set  $C$ ) in detail with examples.

**Model based.** In this study, we employed multiple models, denoted as  $M$ , each corresponding to a specific class. These models were constructed using our LRCNa architecture, as detailed in this paper. However, during the training process, we adapted the model  $M$  to perform binary class classification. To illustrate, for the `drive` class, we divided the training data  $\mathcal{T}$  into two distinct datasets: one exclusively containing samples labeled as `drive`, and the other encompassing samples labeled as `walk`, `bike`, `bus`, `train`, collectively forming the `non_drive` class. We employ this binary class classification approach to establish a set of conditions  $C$ .

In the realm of Deep Learning, the constant evolution of models poses the challenge of choosing the most optimal solution for a given problem. It is a common practice to discard older SOTA models in favor of newer ones. However, this paper introduces a novel approach aimed at leveraging

the capabilities of older, proven models to enhance the performance of the latest SOTA models.

In the context of classification problems, the conventional practice involves employing a threshold of 0.5 for evaluating final results. As illustrated in many receiver operating characteristic (ROC) curves, it is evident that precision generally escalates with an increase in the threshold. Consequently, a higher threshold is advocated as a standard in older state-of-the-art models to enhance their performance.

Examining the ROC curve as an illustrative example, with a threshold of 0.5, the True Positive Rate (TPR) approximates 0.65. Elevating the threshold to 0.9 corresponds to an increased TPR of approximately 0.8. In the event of the introduction of a new state-of-the-art model with a TPR below 0.8 at the 0.5 threshold, adopting the 0.9 threshold from the prior model is recommended. Here, values predicted above 0.9 are considered true positives, while those below 0.9 are designated as unknown predictions. For the latter, the state-of-the-art model can be employed for prediction.

Similar principles are applicable when utilizing the False Positive Rate curve and reducing the threshold. A lowered threshold yields a higher true-false prediction ratio, thereby offering a basis for refining predictions. This methodology, originally designed for binary classification, is adaptable for enhancing predictions in the realm of multiple classifications as well.

**Domain knowledge.** Leveraging domain knowledge pertaining to outliers, we focused on the maximum velocity values present in our dataset. Notably, the highest speed records were associated with the `drive` labels. To ensure fair and consistent comparisons across the dataset, we conducted data normalization based on the maximum speed observed in the `drive` data. The highest velocity recorded in our dataset is 1, associated with the label `drive`. Following closely is the `train` label, exhibiting a maximum velocity of 0.751.

In our datasets, any sample with a speed exceeding the maximum speed recorded for the `train` (0.751 in our dataset) is unambiguously classified as a `drive`. In a broader context, we apply the following condition: For instance, if a sample's maximum speed measures 0.73—falling below both the maximum speeds of 0.751 attributed to the `train` class and 1 associated with the `drive` class, yet surpassing those of other categories—it indicates that the sample is likely to be categorized as either `drive` or `train`. We proceed to assess its multiclass prediction values. The class with the higher prediction value will ultimately determine our final classification for the sample.

**Model based.** In this study, we employed multiple models, denoted as  $M$ , each corresponding to a specific class. These models were constructed using our LRCNa architecture, as detailed in this paper. However, during the training process, we adapted the model  $M$  to perform binary class classification. To illustrate, for the `drive` class, we divided the training data  $\mathcal{T}$  into two distinct datasets: one exclusively containing samples labeled as `drive`, and the other encompassing samples labeled as `walk`, `bike`, `bus`, `train`, collectively forming the `non_drive` class. We employ this binary class classification approach to establish a set of conditions  $C$ .