Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time

Huaxiu Yao [*1] **Caroline Choi** [*1] **Yoonho Lee** [1] **Pang Wei Koh** [1] **Chelsea Finn** [1]

Abstract

Distribution shifts occur when the test distribution differs from the training distribution, and can considerably degrade performance of machine learning models deployed in the real world. While recent works have studied robustness to distribution shifts, distribution shifts arising from the passage of time have the additional structure of timestamp metadata. Real-world examples of such shifts are underexplored, and it is unclear whether existing models can leverage trends in past distribution shifts to reliably extrapolate into the future. To address this gap, we curate Wild-Time, a benchmark of 7 datasets that reflect temporal distribution shifts arising in a variety of real-world applications. On these datasets, we systematically benchmark 9 approaches with various inductive biases. Our experiments demonstrate that existing methods are limited in tackling temporal distribution shift: across all settings, we observe an average performance drop of 21% from in-distribution to out-of-distribution data.

1. Introduction

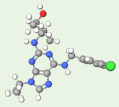
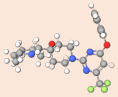
Distribution shift occurs when the test distribution differs from the training distribution, and poses significant challenges for machine learning systems deployed in the real world (Koh et al., 2021). Prior benchmarks for robustness (Koh et al., 2021; Malinin et al., 2021) focus on domain shifts, subpopulation shifts, and distribution shifts in the wild. In this work, we focus on *temporal distribution shifts*, i.e., distribution shifts that arise from the passage of time. In this problem setting, the model is trained on data from the past and evaluated on future data, and can leverage the additional structure of timestamp metadata by inferring the trends in distribution shift throughout time.

[*]Equal contribution [1]Stanford University, Palo Alto, USA. Correspondence to: Huaxiu Yao <huaxiu@cs.stanford.edu>, Caroline Choi <cchoi@stanford.edu>.

ICML 2022 Shift Happens Workshop: Crowdsourcing metrics and test datasets beyond ImageNet., Baltimore, Maryland, USA, 2022.
Copyright 2022 by the author(s).

Though temporal distribution shifts arise in many real-world applications, these kinds of shifts are not well-represented in existing datasets. The typical continual learning problem setting (Adel et al., 2019; Chaudhry et al., 2018; Kirkpatrick et al., 2017; Schwarz et al., 2018; Zenke et al., 2017; Chaudhry et al., 2019; Lopez-Paz & Ranzato, 2017; Rebuffi et al., 2017) assumes that both input features and labels are available at a given timestamp, where the observable data is used to fine-tune the model. However, in many real-world settings, features and labels often arrive asynchronously, requiring robust models which can extrapolate into the future. Furthermore, many popular continual learning benchmarks consist of a manually delineated set of tasks and artificial temporal variations. These include small-image sequences with disparate label splits (e.g., Split TinyImageNet (Le & Yang, 2015)), different kinds of image transformations to MNIST digits (e.g., Rainbow MNIST (Finn et al., 2019)), or different visual recognition targets (Li et al., 2019). While these datasets can be helpful for verifying research ideas, datasets that reflect real-world temporal distribution shifts are needed. Some recent works have investigated natural temporal distribution shifts in domains such as drug discovery (Huang et al., 2021), visual recognition (Cai et al., 2021), and sepsis prediction (Guo et al., 2022) (see detailed comparison in Appendix A). However, there does not exist a systematic study of real-world temporal distribution shifts and a benchmark spanning various domains.

This paper presents **Wild-Time** ("in-the-Wild distribution shifts over Time"), a benchmark of in-the-wild temporal distribution shifts together with two comprehensive evaluation protocols. In Wild-Time, we investigate real-world temporal distribution shifts across a diverse set of tasks (Figure 1), including portraits classification (Ginosar et al., 2015), drug-target binding affinity prediction (Huang et al., 2021), precipitation-level classification (Malinin et al., 2021), ICU patient readmission prediction (Johnson et al., 2021), ICU patient mortality prediction (Johnson et al., 2021), news tag classification (Misra & Grover, 2021), and article category classification (Clement et al., 2019a). The distribution shifts in these applications happen naturally due to the passage of time, which the datasets reflect through changing fashion and social norms (Ginosar et al., 2015), drug candidates and target proteins (Huang et al., 2021), atmospheric conditions (Malinin et al., 2021), and current events (Misra, 2018; Misra & Grover, 2021).

Datasets	Yearbook	FMoW	MIMIC-IV		Drug-BA	Precipitation	HuffPost	arXiv
			Readmission	Mortality				
Input (x)	yearbook photos	satel. image	diagnosis, treatment (ICD9)		mole., protein	meteorological features	article headline	paper title
Prediction (y)	gender	land use	readmission	mortality	binding activity	precipitation	news tag	primary category
Time Range	1930 - 2013	2002 - 2017	2008 - 2019		2013 - 2020	Oct. 18 - Jul. 19	2012 - 2018	2007 - 2022
# Examples	37,189	118,886	270,617		232,386	9,047,294	63,907	2,057,952
Train Example			Diagnosis: 560, 998, 788, 278, E878, 311, V88, V10, 266, 272 Treatment: 456, 545 Readmission: No; Mortality: No				Killer Fail: How Romney's Broken Orca App Cost Him Thousand of Votes TECH	The Limitations of Deep Learning in Adversarial Settings cs.CR
Test Example			Diagnosis: 155, 456, 452, 572 Treatment: 423, 549, 990, 990 Readmission: Yes; Mortality: Yes				Possible Autopilot Use Probed After Tesla Crashes at 60mph TECH	Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation cs.LG

Time

Figure 1. The Wild-Time benchmark includes 7 datasets with 8 tasks. For each task, we train models on the past and evaluate it in the future. We list the input, prediction, time range and the number of examples for each task.

We evaluate Wild-Time with a fixed time split (Eval-Fix). Specifically, Eval-Fix evaluation provides a single train and test split, as in standard supervised learning, and is geared toward the broader machine learning community. Each model is trained on data before a given timestamp and evaluated on data collected afterwards. We evaluate several representative continual learning and invariant learning approaches on these datasets. In particular, we extend domain generalization methods to the temporal distribution shift setting and evaluate these methods on all datasets. From these evaluations, we conclude: (1) most invariant learning approaches do not show substantial improvements compared to standard ERM training; (2) continual learning approaches do not make models more robust to temporal distribution shift. We hope that Wild-Time will accelerate the development of temporally robust models that can be safely deployed in the wild.

2. Problem and Evaluation Settings

Following (Koh et al., 2021), we view the entire data distribution as a mixture of T timestamps $\mathcal{T} = \{1, \ldots, T\}$. Each timestamp t is associated with a data distribution P_t over (x, y), where x and y represent input features and labels, respectively, and all examples are sampled from the data distribution P_t. To formulate the temporal distribution shift setting, we define the training distribution as $P^{tr} = \sum_{t=1}^{T} \lambda_t^{tr} P_t$, and the test distribution as $P^{ts} = \sum_{t=1}^{T} \lambda_t^{ts} P_t$. Note that, here, timestamp differs from the notion of "domain" used in other works on distribution shift (Ahuja et al., 2021). In the temporal shift setting, we do not require distribution shift between consecutive timestamps, i.e., we can have $P_t = P_{t-1}$. We evaluate models on a single, fixed train-test time split. Concretely, we denote the split timestamp as t_s. The train and test sets are $\mathcal{T}^{tr} = \{t \leq t_s | \forall t\}$, $\mathcal{T}^{ts} = \{t > t_s | \forall t\}$, respectively. We evaluates performance using two metrics, average and worst-time performance.

3. Datasets

In this section, we briefly discuss the datasets and tasks included in Wild-Time, which reflect natural temporal distribution shifts. We provide detailed descriptions of all datasets in Appendix B.

Yearbook. We study changes in fashion and social norms over time on the Yearbook dataset, which consists of 37,921 frontal-facing American high school yearbook photos (Ginosar et al., 2015). Each photo is a $32 \times 32 \times 1$ grey-scale image associated with a binary label y, which represents the student's gender. The training set includes data from before 1970, and the test set comprises data after 1970, which corresponds to 40 and 30 years, respectively.

FMoW-WildT. We study changes in satellite imagery overr time on the Functional Map of the World (FMoW) dataset (Christie et al., 2018), adapted from the WILDS benchmark (Koh et al., 2021). Each input x is a satellite image, and the corresponding label y is one of 62 land use categories. We choose the year 2009 to split the training and test sets.

MIMIC-IV-WildT. We study healthcare-related temporal distribution shifts on the MIMIC-IV dataset. We consider two classification tasks: (1) **MIMIC-Readmission** aims to predict the risk of being readmitted to the hospital within 15 days. (2) **MIMIC-Mortality** aims to predict in-hospital mortality for each patient. We treat each admission as one record, and for each record, we concatenate the ICD9 codes (Organization, 2004) of diagnosis and treatment. The train set consists of patient data from $2008 - 2013$, while the test set consists of data from $2014 - 2020$.

Drug-BA. We study the temporal shift of drug-target binding activity prediction on the TDC domain generalization benchmark, where the input x contains the information of both compounds and the target proteins, the label y indicates the binding activity value. We use 2016 to split training and test sets, i.e., data from years $2013 - 2018$ are used for training and the test set comprises data from 2019 and 2020.

Precipitation-WildT. Adapted from (Malinin et al., 2021), we study precipitation-level prediction over the course of a year. Given a tabular data of 123 meteorological features, the task is to classify among nine classes of precipitation levels.

Huffpost. We study temporal shifts arising from current events on the Huffpost dataset, which focuses on news classification (Misra & Grover, 2021). Each input feature x is a news headline and the output y is the news category. We choose the year 2015 as the split timestamp.

arXiv. We study the arXiv dataset (Clement et al., 2019b), where the task is to predict the primary category of arXiv preprints given the paper title as input. The entire dataset includes 172 preprint categories over the years $2007 - 2022$.

4. Baselines for Temporal Distribution Shifts

Many algorithms have been proposed to improve a model's robustness to distribution shifts or improve a model's performance on a stream of data. For our evaluation, we choose several representative methods from three main categories – classical supervised learning, continual learning, and invariant learning. These methods have been successful on domain generalization and continual learning benchmarks. In particular, we extend the selected invariant learning approaches to the temporal distribution shift setting. See Appendix C for a detailed discussion of these baseline algorithms and how they are applied them to Wild-Time tasks.

Classical Supervised Learning. We evaluate the performance of empirical risk minimization (ERM) and fine-tuning (FT) on all tasks.

Continual Learning. Continual learning, also known as lifelong learning or incremental learning, aims to effectively learn from non-stationary distributions via a sequential stream of data (Adel et al., 2019; Chaudhry et al., 2018; Kirkpatrick et al., 2017; Schwarz et al., 2018). The goal is to accumulate and reuse knowledge in future learning without forgetting information needed for previous tasks, a phenomenon known as catastrophic forgetting (Kirkpatrick et al., 2017), which may enable such models to robustly extrapolate into the future in the temporal shift setting. We evaluate three representative algorithms, including regularization-based (EWC (Kirkpatrick et al., 2017), SI (Zenke et al., 2017)) and memory-based (A-GEM (Chaudhry et al., 2019)) methods.

Temporally Invariant Learning. Invariant learning methods learn representations or predictors that are invariant across different domains. In Wild-Time, we select four representative invariant learning methods: CORAL (Sun & Saenko, 2016), IRM (Arjovsky et al., 2019), LISA (Yao et al., 2022), and GroupDRO (Sagawa et al., 2020). We adapt these methods to a temporal setting by treating substreams of data as domains, and refer to the temporal versions as CORAL-T, IRM-T, and GroupDRO-T. We introduce these four approaches and discuss how we adapt them to the temporal distribution shift setting in Appendix C.

5. Experiments

We benchmark the performance of all baselines on each dataset in Wild-Time. See Appendix D for more details.

5.1. Experimental Setup

Data Split. The training and test sets are non-overlapping subsets of the entire dataset such that the training data timestamps are earlier than the test data timestamps. Temporal out-of-distribution (OOD) robustness is measured by performance on the test set. To compare OOD with in-distribution (ID) performance, we measure the average per-timestep performance on a held-out set of 10% training examples (20% for Drug-BA, MIMIC-Mortality, and MIMIC-Readmission) from each training ID timestamp.

Evaluation Metrics. We measure accuracy in most classification tasks, including Yearbook, FMoW, MIMIC-Readmission, Precipitation, HuffPost, and arXiv. For the MIMIC-Mortality task, we use ROC-AUC due to label imbalance. Root Mean Square Error (RMSE) is used in all regression tasks, including Drug-BA and Weather-Temp.

5.2. Results

Table 1 shows the performance of all baselines on the Wild-Time datasets. For each task, we visualize the OOD performance on every test timestamp in Figure 3 (Appendix D). The following high-level observations summarize our findings:

- In all tasks, OOD performance is substantially lower than ID performance.
- In FMoW, MIMIC-Readmission, and MIMIC-Mortality, model performance degrades with time (Figures 3b, 3d, 3e), as models exhibit higher OOD accuracy on timestamps closer to that of the training data. In Yearbook (Figure 3a), performance fluctuates significantly, with models achieving higher OOD accuracy at later timestamps (e.g., 1991-1996) compared to earlier

Table 1. The in-distribution versus out-of-distribution test performance of each method evaluated on Wild-Time. The average and standard deviation (value in parentheses) are computed over three random seeds. We bold the best OOD performance for each dataset.

	Yearbook (Accuracy (%) ↑)			FMoW (Accuracy (%) ↑)		
	ID Avg.	OOD Avg.	OOD Worst	ID Avg.	OOD Avg.	OOD Worst
Fine-tuning	95.43 (1.65)	81.98 (1.52)	**69.62 (3.38)**	39.76 (1.36)	26.98 (0.12)	20.84 (0.62)
EWC	96.36 (0.47)	80.07 (0.22)	66.61 (1.95)	39.68 (0.95)	27.13 (0.48)	21.38 (0.43)
SI	96.40 (0.83)	78.70 (3.78)	65.18 (2.44)	39.69 (0.69)	27.10 (0.39)	21.08 (0.51)
A-GEM	97.18 (0.43)	81.04 (1.40)	67.07 (2.23)	35.63 (5.97)	26.48 (0.54)	20.48 (0.46)
ERM	97.99 (1.40)	79.50 (6.23)	63.09 (5.15)	58.25 (0.36)	**37.19 (0.33)**	**27.79 (0.64)**
GroupDRO-T	96.04 (0.45)	77.06 (1.67)	60.96 (1.83)	40.47 (1.03)	27.49 (0.66)	22.09 (0.59)
LISA	96.56 (0.97)	**83.65 (4.61)**	68.53 (5.79)	53.44 (0.41)	36.43 (0.45)	26.95 (0.38)
CORAL-T	98.19 (0.58)	77.53 (2.15)	59.34 (1.46)	48.18 (0.53)	32.49 (0.57)	25.25 (0.60)
IRM-T	97.02 (1.52)	80.46 (3.53)	64.42 (4.38)	48.97 (1.05)	34.78 (0.33)	26.91 (0.41)
	MIMIC-Readmission (Accuracy (%) ↑)			MIMIC-Mortality (AUC (%) ↑)		
	ID Avg.	OOD Avg.	OOD Worst	ID Avg.	OOD Avg.	OOD Worst
Fine-tuning	57.02 (3.68)	48.84 (4.25)	44.62 (4.92)	89.49 (0.78)	71.71 (5.03)	62.34 (7.82)
EWC	60.68 (3.44)	51.41 (1.62)	47.24 (2.28)	87.38 (1.28)	69.04 (2.03)	58.89 (2.08)
SI	57.64 (1.78)	43.43 (7.34)	37.01 (11.43)	86.73 (1.57)	66.47 (1.56)	55.49 (1.31)
A-GEM	60.25 (9.15)	42.22 (2.67)	34.38 (3.69)	88.69 (0.78)	70.19 (6.38)	60.77 (9.21)
ERM	69.90 (3.83)	58.51 (4.06)	55.84 (4.42)	90.86 (0.52)	69.74 (4.51)	59.43 (6.85)
GroupDRO-T	73.80 (0.72)	**66.91 (0.91)**	**65.68 (1.32)**	89.13 (1.25)	73.06 (2.32)	65.52 (3.74)
LISA	66.62 (3.46)	55.99 (2.89)	53.73 (2.67)	90.30 (1.13)	**78.11 (0.93)**	**71.69 (1.93)**
CORAL-T	75.23 (2.74)	64.50 (3.03)	61.97 (3.38)	89.74 (0.94)	70.81 (3.22)	62.19 (4.63)
IRM-T	72.47 (3.56)	59.67 (2.19)	56.73 (1.82)	88.07 (2.41)	67.02 (4.37)	57.08 (5.81)
	Drug-BA (R ↑)			Precipitation (Accuracy (%) ↑)		
	ID Avg.	OOD Avg.	OOD Worst	ID Avg.	OOD Avg.	OOD Worst
Fine-tuning	0.648 (0.003)	0.314 (0.007)	0.229 (0.020)	49.17 (0.75)	46.08 (0.94)	44.15 (0.89)
EWC	0.601 (0.004)	0.299 (0.013)	0.202 (0.004)	49.84 (1.22)	46.21 (1.73)	44.11 (2.36)
SI	0.643 (0.001)	0.319 (0.005)	0.234 (0.020)	49.96 (1.11)	46.70 (1.15)	45.16 (0.89)
A-GEM	0.639 (0.004)	0.294 (0.039)	0.203 (0.033)	47.70 (0.22)	45.41 (0.96)	43.11 (0.90)
ERM	0.787 (0.009)	**0.357 (0.012)**	0.244 (0.028)	50.60 (0.44)	**47.83 (0.58)**	**46.11 (0.60)**
GroupDRO-T	0.697 (0.002)	0.342 (0.006)	0.246 (0.007)	48.37 (0.41)	45.89 (0.44)	43.60 (0.25)
LISA	N/A	N/A	N/A	49.88 (0.38)	46.84 (0.73)	45.21 (0.67)
CORAL-T	0.745 (0.007)	0.355 (0.016)	**0.271 (0.017)**	49.28 (1.20)	46.97 (0.60)	45.04 (0.78)
IRM-T	0.716 (0.005)	0.355 (0.004)	0.252 (0.015)	48.77 (1.27)	46.38 (1.61)	44.30 (1.26)
	HuffPost (Accuracy (%) ↑)			arXiv (Accuracy (%) ↑)		
	ID Avg.	OOD Avg.	OOD Worst	ID Avg.	OOD Avg.	OOD Worst
Fine-tuning	76.79 (0.51)	69.59 (0.10)	68.91 (0.49)	51.42 (0.15)	50.31 (0.39)	48.19 (0.41)
EWC	76.26 (0.32)	69.42 (1.00)	68.61 (0.98)	51.34 (0.13)	**50.40 (0.11)**	**48.18 (0.18)**
SI	76.97 (0.30)	**70.46 (0.27)**	69.05 (0.52)	51.52 (0.19)	50.21 (0.40)	48.07 (0.48)
A-GEM	77.15 (0.07)	70.22 (0.50)	**69.15 (0.88)**	51.57 (0.18)	50.30 (0.37)	48.14 (0.40)
ERM	79.40 (0.05)	70.42 (1.15)	68.71 (1.36)	53.78 (0.16)	45.94 (0.97)	44.09 (1.05)
GroupDRO-T	78.04 (0.26)	69.53 (0.54)	67.68 (0.78)	49.78 (0.22)	39.06 (0.54)	37.18 (0.52)
LISA	78.20 (0.53)	69.99 (0.60)	68.04 (0.75)	50.72 (0.31)	47.82 (0.47)	45.91 (0.42)
CORAL-T	78.19 (0.31)	70.05 (0.63)	68.39 (0.88)	53.25 (0.12)	42.32 (0.60)	40.31 (0.61)
IRM-T	78.38 (0.51)	70.21 (1.05)	68.71 (1.13)	46.30 (0.53)	35.75 (0.90)	33.91 (1.09)

timestamps (e.g., 1981-1986). In Drug-BA, Precipitation, HuffPost, and arXiv, models achieve the best performance on the earliest test timestamps. Nevertheless, there is a significant gap between the OOD performance and best ID performance for all datasets and methods. This performance gap changes in a continual manner over time, indicating that the nature of the distribution shift is correlated with the provided timestamps.

- Most invariant learning approaches (CORAL-T, GroupDRO-T, IRM-T) did not show clear improvements over ERM. LISA performs slightly better than these invariant learning approaches, which corroborates findings in the original paper (Yao et al., 2022).
- Incremental training approaches (Fine-tuning, EWC, SI, A-GEM) improve worst OOD performance on the arXiv and HuffPost datasets. Incremental training tends to bias the trained model toward the last few timestamps. As factual knowledge in text data changes gradually, incremental training approaches seem to be more suitable for these gradual temporal shifts. In all tasks other than Yearbook, incremental training methods perform worse than invariant learning approaches, indicating that invariant learning approaches may be promising for constructing temporally robust models.

References

Adel, T., Zhao, H., and Turner, R. E. Continual learning with adaptive weights (claw). *arXiv preprint arXiv:1911.09514*, 2019.

Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., and Zhou, J. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

Cai, Z., Sener, O., and Koltun, V. Online continual learning with natural distribution shifts: An empirical study with visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8281–8290, 2021.

Carpenter, K. A., Cohen, D. S., Jarrell, J. T., and Huang, X. Deep learning and virtual drug screening. *Future medicinal chemistry*, 10(21):2557–2567, 2018.

Chaudhry, A., Dokania, P. K., Ajanthan, T., and Torr, P. H. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 532–547, 2018.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., and Alemi, A. A. On the use of arxiv as a dataset, 2019a.

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., and Alemi, A. A. On the use of arxiv as a dataset. *arXiv preprint arXiv:1905.00075*, 2019b.

Ettema, R. G., Peelen, L. M., Schuurmans, M. J., Nierich, A. P., Kalkman, C. J., and Moons, K. G. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. *Circulation*, 122(7):682–689, 2010.

Fang, C., Xu, Y., and Rockmore, D. N. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Fang, F., Dutta, K., and Datta, A. Domain adaptation for sentiment classification in light of multiple sources. *INFORMS Journal on Computing*, 26(3):586–598, 2014.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. In *International Conference on Machine Learning*, pp. 1920–1930, 2019.

Ginosar, S., Rakelly, K., Sachs, S., Yin, B., and Efros, A. A. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–7, 2015.

Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E., Posada, J., Aftandilian, C., Shah, N., and Sung, L. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):1–10, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Jia, X., Wang, M., Khandelwal, A., Karpatne, A., and Kumar, V. Recurrent generative networks for multi-resolution satellite data: An application in cropland monitoring. In *IJCAI*, 2019.

Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S.-W., Wei, X., Arnold, A., and Ren, X. Lifelong pretraining: Continually adapting language models to emerging corpora. *arXiv preprint arXiv:2110.08534*, 2021.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. Mimic-iv, 2021. URL https://physionet.org/content/mimiciv/1.0/.

Kaushik, P., Gain, A., Kortylewski, A., and Yuille, A. Understanding catastrophic forgetting and remembering in continual learning with optimal relevance mapping. *arXiv preprint arXiv:2102.11343*, 2021.

Ke, Z., Liu, B., Wang, H., and Shu, L. Continual learning with knowledge transfer for sentiment classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 683–698. Springer, 2020.

Ke, Z., Xu, H., and Liu, B. Adapting bert for continual learning of a sequence of aspect sentiment classification tasks. *arXiv preprint arXiv:2112.03271*, 2021.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, X., Zhou, Y., Wu, T., Socher, R., and Xiong, C. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. *arXiv preprint arXiv:1904.00310*, 2019.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

Lin, B. Y., Wang, S., Lin, X. V., Jia, R., Xiao, L., Ren, X., and Yih, W.-t. On continual model refinement in out-of-distribution data streams. *arXiv preprint arXiv:2205.02014*, 2022.

Lin, Z., Shi, J., Pathak, D., and Ramanan, D. The clear benchmark: Continual learning on real-world imagery. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., and Gao, J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1903–1911, 2017.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. Technical report, 2013.

Malinin, A., Band, N., Chesnokov, G., Gal, Y., Gales, M. J., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

Martin, E. J., Polyakov, V. R., Zhu, X.-W., Tian, L., Mukherjee, P., and Liu, X. All-assay-max2 pqsar: activity predictions as accurate as four-concentration ic50s for 8558 novartis assays. *Journal of chemical information and modeling*, 59(10):4450–4459, 2019.

Misra, R. News category dataset, 06 2018.

Misra, R. and Grover, J. *Sculpting Data for ML: The first act of Machine Learning*. 01 2021. ISBN 978-0-578-83125-1.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Organization, W. H. *International Statistical Classification of Diseases and Related Health Problems: Alphabetical index*, volume 3. World Health Organization, 2004.

Öztürk, H., Özgür, A., and Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537, 2018.

Sharifi, A. Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*, 101(3):891–896, 2021.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

Tsymbal, A. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Yang, C., Xiao, C., Glass, L., and Sun, J. Change matters: Medication change prediction with recurrent residual networks. *arXiv preprint arXiv:2105.01876*, 2021.

Yao, H., Huang, L.-K., Zhang, L., Wei, Y., Tian, L., Zou, J., Huang, J., et al. Improving generalization in meta-learning via task augmentation. In *International Conference on Machine Learning*, pp. 11887–11897. PMLR, 2021a.

Yao, H., Wei, Y., Huang, L.-K., Xue, D., Huang, J., and Li, Z. J. Functionally regionalized knowledge transfer for low-resource drug discovery. *Advances in Neural Information Processing Systems*, 34, 2021b.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *ICML*, 2022.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3987–3995. JMLR. org, 2017.

A. Comparison with Existing Benchmarks

Wild-Time offers a unified framework to facilitate the development of models robust to in-the-wild temporal distribution shifts. We discuss how Wild-Time is related to existing distribution shift and continual learning benchmarks.

Relation to Distribution Shift Benchmarks. Distribution shift has been widely studied in the machine learning community. Early works presented small-scale benchmarks to study distribution shifts in sentiment analysis (Fang et al.,

2014) and object detection (Saenko et al., 2010). Subsequent distribution shift benchmarks focused on larger-scale, real-world data. The first line of such benchmarks induce distribution shifts by applying different kinds of transformations to object recognition datasets. These benchmarks include: (1) ImageNet-A (Hendrycks et al., 2021b), ImageNet-C (Hendrycks & Dietterich, 2019), and CIFAR-10.1 (Recht et al., 2018), which add noise or adversarial examples to the original Imagenet (Russakovsky et al., 2015) and CIFAR (Krizhevsky et al., 2009) datasets, respectively; (2) Colored MNIST (Arjovsky et al., 2019), which changes the color of digits from the original MNIST dataset. More recent works created domain generalization benchmarks by collecting sets of images with different styles or backgrounds, such as PACS (Li et al., 2017), DomainNet (Peng et al., 2019), VLCS (Fang et al., 2013), Office-Home (Venkateswara et al., 2017), ImageNet-R (Hendrycks et al., 2021a), BREEDS (Santurkar et al., 2020), Waterbirds (Sagawa et al., 2020), and MetaShift (Liang & Zou, 2022). While these datasets are useful testbeds for verifying the efficacy of new algorithms, they do not reflect natural distribution shifts that arise in real-world applications.

Recently, a few works have constructed datasets and benchmarks for real-world distribution shifts. The WILDS benchmark consists of ten datasets spanning a wide range of real-world applications, such medical image recognition, sentiment classification, land-use classification with satellite image, and code autocompletion, with a focus on domain shifts and subpopulation shifts (Koh et al., 2021). UWILDS extends WILDS and introduces unlabeled data to help boost model robustness to distribution shifts (Sagawa et al., 2021). SHIFTS (Malinin et al., 2021) is composed of three datasets, concerning weather prediction, machine translation, and self-driving vehicle motion prediction. Unlike these works that focus on general distribution shifts, we target temporal distribution shifts arising in real-world applications. A few recent works have started investigating model robustness over time, in real-world applications such as healthcare-related prediction (Guo et al., 2022), drug discovery (Huang et al., 2021), image-based geo-localization (Cai et al., 2021), machine reading comprehension (Lin et al., 2022), and tweet hashtag prediction (Jin et al., 2021). Unlike prior datasets that target specific applications, Wild-Time presents a comprehensive benchmark comprised of 7 datasets from diverse domains and offers systematic evaluation protocols.

Relation to Continual Learning Benchmarks. Continual learning methods are often benchmarked on image classification datasets. Some popular benchmarks such as RainbowMNIST (Finn et al., 2019) and permuted MNIST (Kaushik et al., 2021) apply various image transformations to a small-scale image dataset to obtain a sequence of tasks. Others such as Split CIFAR100 (Krizhevsky et al., 2009), Split TinyImagenet (Le & Yang, 2015), F-CelebA (Ke et al., 2020), and Stanford Cars (Krause et al., 2013) split a large image dataset into multiple non-overlapping class sets, where each is regarded as one task. A third collection of related benchmarks treats each object recognition dataset as a different task. For example, Visual Domain Decathlon (Li et al., 2019) consists of 10 datasets from various domains, such as Aircraft (Maji et al., 2013), SVHN (Netzer et al., 2011), Omniglot (Lake et al., 2015), VGG-Flowers (Nilsback & Zisserman, 2008), CLEAR (Lin et al., 2021). In the natural language processing (NLP) domain, continual learning benchmarks such as ASC (Ke et al., 2021) and DSC (Ke et al., 2020) have been used to evaluate the performance of large-scale pretrained models over time. Unlike these prior benchmarks, Wild-Time presents a collection of datasets that reflect natural temporal distribution shifts arising in real-world applications.

B. Detailed Dataset Description

Yearbook. Social norms, fashion styles, and population demographics change over time. This is captured in the Portraits dataset, which consists of 37,921 frontal-facing American high school yearbook photos (Ginosar et al., 2015). We exclude portraits from earlier years due to the limited number of examples in these years, resulting in 33,431 examples from the 1930 to 2013. Each photo is a $32 \times 32 \times 1$ grey-scale image associated with a binary label y, which represents the high schooler's gender. The training set includes data from before 1970, and the test set comprises data after 1970, which corresponds to 40 and 30 years, respectively.

FMoW-WildT. Machine learning models can be used to analyze satellite imagery and aid humanitarian and policy efforts by monitoring croplands (Jia et al., 2019) and predicting crop yield (Sharifi, 2021) and poverty levels (Jean et al., 2016). Due to human activity, satellite imagery changes over time, requiring models that are robust to temporal distribution shifts.

We study this problem on the Functional Map of the World (FMoW) dataset (Christie et al., 2018), adapted from the WILDS benchmark (Koh et al., 2021). Specifically, given a satellite image, the task is to predict the type of land usage. Each input x is a satellite image, and the corresponding label y is one of 62 land use categories. We choose the year 2009 to split the training and test sets. A key difference between FMoW-WildT and FMoW-WILDS is that FMoW-WildT includes timestamp metadata, focusing on distribution shifts over time and partition the in-distribution and out-of-distribution data by year in order to systematically evaluate robustness to temporal distribution shifts.

MIMIC-IV-WildT. Many machine learning healthcare applications have emerged in the last decade, such as predicting disease risk (Ma et al., 2017), medication changes

(Yang et al., 2021), patient subtyping (Baytas et al., 2017), in-hospital mortality (Guo et al., 2022), and length of hospital stay (Ettema et al., 2010). However, a key obstacle in deploying machine learning-based clinical decision support systems is temporal dataset shift associated with changes in healthcare over time (Guo et al., 2022).

We study this problem on MIMIC-IV, one of the largest public healthcare datasets that comprises medical records of over 40,000 patients. In MIMIC-WildT, we treat each admission as one record, resulting in 216,487 healthcare records from $2008 - 2020$. Specifically, we consider two classification tasks:

- **MIMIC-Readmission** aims to predict the risk of being readmitted to the hospital within 15 days.
- **MIMIC-Mortality** aims to predict in-hospital mortality for each patient.

For each record, we concatenate the corresponding ICD9 codes (Organization, 2004) of diagnosis and treatment. The label is a binary value that indicates whether the patient is readmitted or passed away for MIMIC-Readmission and MIMIC-Mortality, respectively. The train set consists of patient data from $2008 - 2013$, while the test set consists of data from $2014 - 2020$.

Drug-BA. Drug discovery brings new candidate medications to potentially billions of people. A crucial step in the drug discovery process is virtual screening, in which we predict the binding activity value of a compound against the target protein of a disease (Carpenter et al., 2018; Svetnik et al., 2003). Recent binding activity prediction models investigate the binding pairs between existing compounds and target proteins (Martin et al., 2019; Öztürk et al., 2018; Yao et al., 2021a;b). In practice, new target proteins or new classes of compounds appear over time, requiring machine learning models that are robust to domain shifts across time.

We study this temporal shift of drug-target binding activity prediction on the TDC domain generalization benchmark, where the input x contains the information of both compounds and the target proteins, the label y indicates the binding activity value. We use 2016 to split training and test sets, i.e., data from $2013 - 2018$ are used for training and the test set comprises data from 2019 and 2020. We remove the year 2021 from the original benchmark it only has one month's worth of data.

Precipitation-WildT. Precise weather forecasting provides effective guidance for daily activity. The temporal distribution shift issue has been widely observed in weather prediction tasks (Malinin et al., 2021; Tsymbal, 2004).

We adapt the original Shifts Weather Prediction dataset (Malinin et al., 2021) for Precipitation-WildT, using measurements taken from October 2018 - August 2019. The Precipitation-WildT dataset consists of 123 heterogeneous meteorological features, 1 target (precipitation class), and 1 metadata attribute (time). We partition the dataset by month. We data from October 2018 - April 2019 (7 months) for training, and data from May 2019 - August 2019 (4 months) for test.

Huffpost. In many language models which deal with information correlated with time, temporal distribution shifts cause performance degradation in downstream tasks such as Twitter hashtag classification (Jin et al., 2021) or question answering systems (Lin et al., 2022). The performance drops along the temporal dimension reflect changes in the style or content of news.

We study temporal shifts in news articles in the Huffpost dataset, which aims to categorize news articles from their headlines (Misra & Grover, 2021). Specifically, each input feature x is a news headline and the output y is the news category. We only keep news categories that appear in every year, resulting in 11 categories in total. We choose year 2015 as the split timestamp.

arXiv. Similar to changes in news over time, arXiv preprints also change over time. For example, "neural network attack" was originally a popular keyword in the security community, but it gradually became more prevalent in the machine learning community. We study these temporal distribution shift in the arXiv dataset (Clement et al., 2019b), where the task is to predict the primary category of arXiv pre-prints given the pre-print title as input. The entire dataset includes 172 pre-print categories from $2007 - 2022$.

C. Detailed Baselines for Temporal Distribution Shifts

Here, we detail the baselines for temporal distribution shifts.

C.1. Classical Supervised Learning.

Empirical Risk Minimization (ERM). We directly train a machine learning model with ERM.

Fine-tuning (FT). We incrementally fine-tune the model at every timestamp and evaluate in the future.

C.2. Continual Learning.

Elastic Weight Consolidation (EWC) overcomes catastrophic forgetting by slowing down learning on weights based on how important they are for previous tasks (Kirkpatrick et al., 2017).

Synaptic Intelligence (SI) accumulates information relevant to tasks over time (Zenke et al., 2017). Specifically, SI tracks past and current parameter values to estimate the

importance of each "synapse" to previous tasks, and consolidates the important synapses at each new task to prevent catastrophic forgetting.

Averaged Gradient Episodic Memory (A-GEM) leverage a small episodic memory to constrain new updates to not interfere with previous tasks by minimizing average episodic memory loss at each timestep (Chaudhry et al., 2019).

C.3. Temporally Invariant Learning (Appendix B.3).

We first detail the four invariant learning baselines and then discuss how to adapt them into temporal distribution shift setting.

GroupDRO uses distributionally robust optimization to minimize the loss on the worst-case group during training, where each group is defined as a domain-class pair (Sagawa et al., 2020).

LISA uses selective augmentation to eliminate the effect of domain-associated information and achieves invariant learning (Yao et al., 2022). Concretely, LISA interpolates examples either with the same domain but different labels (intra-domain LISA) or with the same label but different domains (intra-label LISA).

CORAL penalizes differences between the means and covariances of the different domains' feature distributions with an explicit regularizer (Sun & Saenko, 2016).

IRM aims to improve performance over different domains by learning an invariant casual predictor from multiple training environments (Arjovsky et al., 2019).

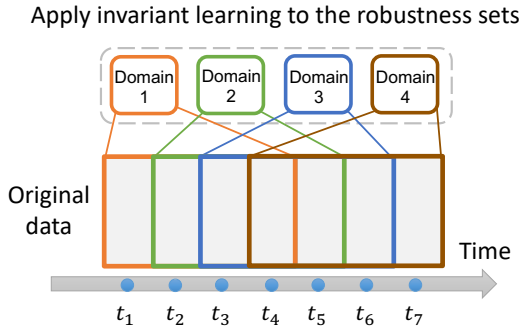


Figure 2. Construction of a temporal robustness set. Here, we have $T = 7$ timesteps and sliding window length $L = 3$. By applying sliding window-based segmentation, we obtain 4 substreams of data, where each substream is treated as a "domain". We apply the invariant learning approaches to this robustness set.

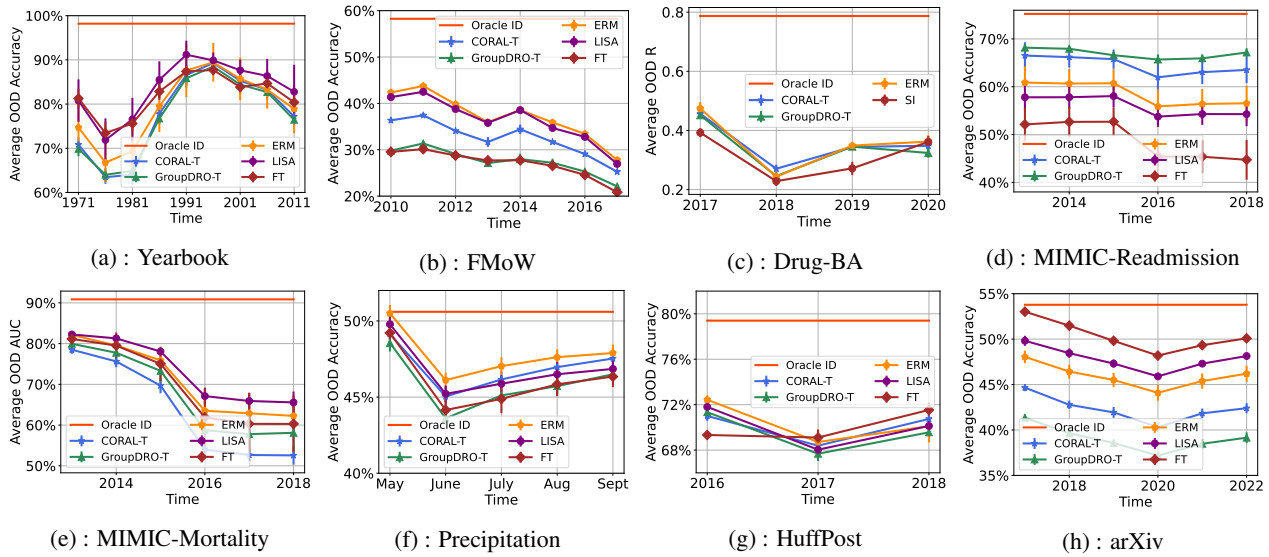
As mentioned in Section 2, in the temporal distribution shift setting, the "timestamp" information is not the same as the notion of a "domain", as the distribution shift may not happen between consecutive timestamps. Hence, the data streams in our benchmark are unsegmented and do not include domain boundaries. This setting poses new challenges to the above invariant learning approaches, which rely on domain labels.

To address this challenge, we adapt the above invariant learning approaches to the temporal distribution shift setting. We leverage timestamp metadata to create a temporal robustness set consisting of substreams of data, where each substream is treated as one domain. Specifically, as shown in Figure 2, we define a sliding window $\mathcal{G}$ with length L. For a data stream with T timestamps, we apply the sliding window $\mathcal{G}$ to obtain $T - L + 1$ substreams. We treat each substream as a "domain" and apply the above invariant algorithms on the robustness set. We name the adapted CORAL, GroupDRO and IRM as CORAL-T, GroupDRO-T, IRM-T, respectively. Note that, we do not adapt LISA since the intra-label LISA performs well without using domain information, which is also mentioned in the original paper. See Appendix B.4 for further details on temporal adaptation of these invariant learning algorithms.

D. More Experimental Settings and Results

D.1. Hyperparameter Settings

For each method, we tune hyperparameters using cross-validation. Instead of using an in-distribution validation set, we hold out 10% of the data of each training timestamp (20% for Drug-BA, MIMIC-Readmission, and MIMIC-Mortality) to construct the out-of-distribution validation set. Here, we use examples from the remaining 90% of the data timestamps to train the model and evaluate the out-of-distribution performance on the validation set for hyperparameter tuning. We repeat this process several times via cross-validation. After tuning all hyperparameters, we use the entire training set to train the model.



(a) : Yearbook (b) : FMoW (c) : Drug-BA (d) : MIMIC-Readmission

(e) : MIMIC-Mortality (f) : Precipitation (g) : HuffPost (h) : arXiv

Figure 3. Out-of-distribution performance per test timestamp. We select five representative baselines – ERM, FT, CORAL-T, GroupDRO-T, LISA, and show the corresponding performance. Oracle ID represent the best ID performance over all compared baselines. FT: Fine-tuning.