Fraunhofer IESE

Fraunhofer Institute for Experimental
Software Engineering IESE

Department of Safety Engineering

StaDRe and StaDRo: Reliability and Robustness Estimation of ML-based Forecasting using Statistical Distance Measures

Mohammed Naveed Akram, Akshatha Ambekar*, Ioannis Sorokos, Koorosh Aslansefat, Daniel Schneider

*Presenting Author

Agenda

- **Introduction**
  - Introduction of SafeML
    - Distributional shift
    - Reliability and robustness measure using SafeML
- **Related Work**
- **Research Questions**
- **Proposed measures for reliability and robustness**
- **Experiments: AI-based investment prediction**
- **Results**
- **Conclusion and Future Scope**

Chapter 01

Introduction and Background

Introduction

- **Increased use of AI in safety-critical applications**
  - Trustworthy and dependable AI/ML systems.
- **Desired qualities of ML solutions**
  - Safety, reliability, robustness, explainability, transparency and security.
- **Five categories of safety problems in AI [1]**
- **ML models are vulnerable to distributional shifts.**
- **Application of ML models must be supported with context informatio that they are designed for: *Scope Compliance***

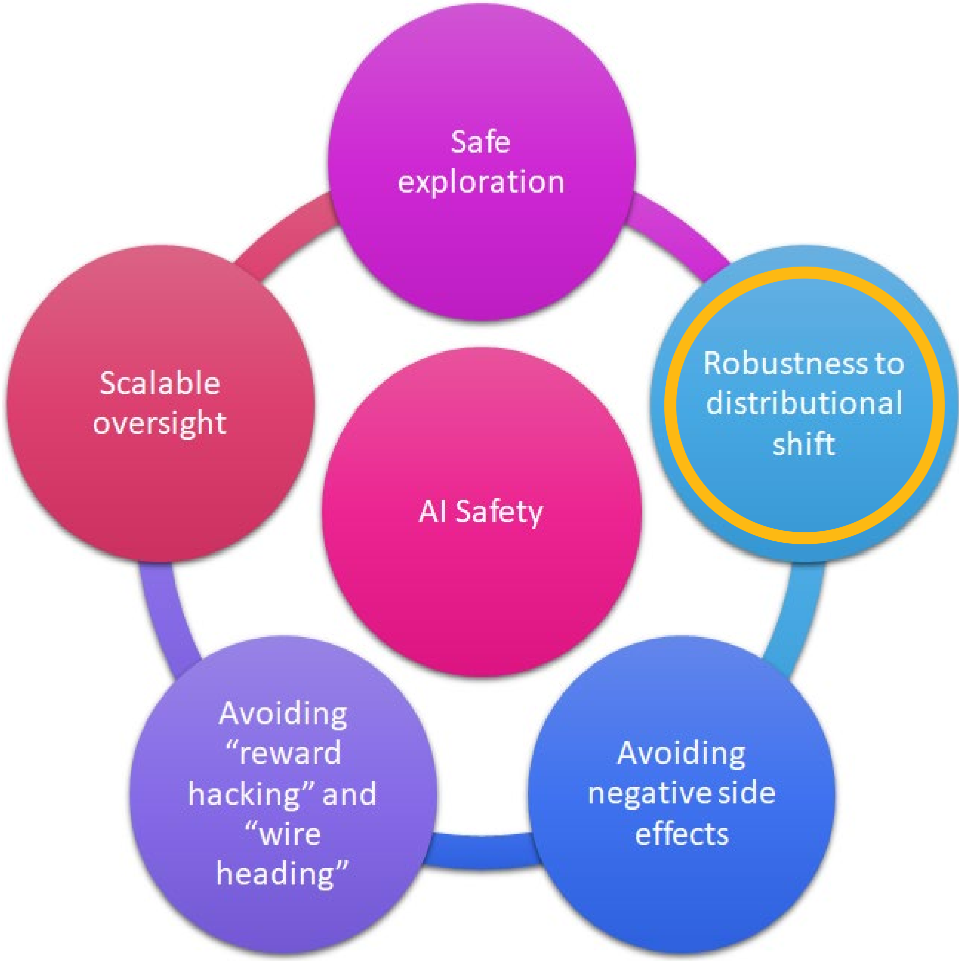
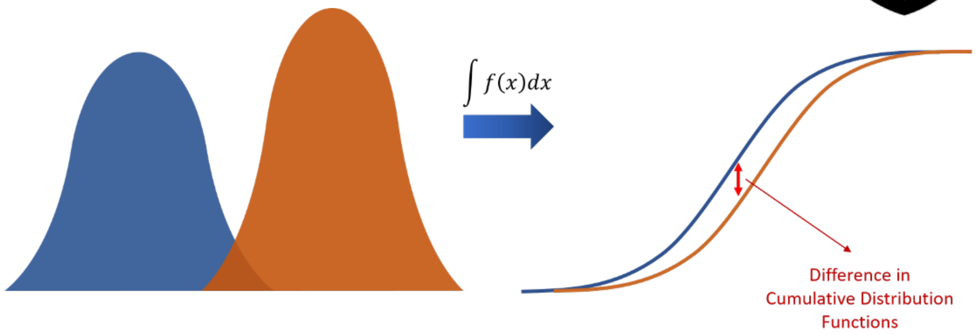

Safe exploration
Robustness to distributional shift
Scalable oversight
AI Safety
Avoiding “reward hacking” and “wire heading”
Avoiding negative side effects

Figure **1**: Five categories of AI Safety **([1])**

Addressing Scope Compliance

Introducing SafeML

- **Challenge**
  - How to monitor scope compliance?
- **Approach**
  - Using statistical distance measures to evaluate ‘how far’ from our trained context are we currently operating in.
- **Benefits**
  - Maintain safe state by not trusting ML when out of intended context.
- **Limitation**
  - Classification tasks, images and tabular data.



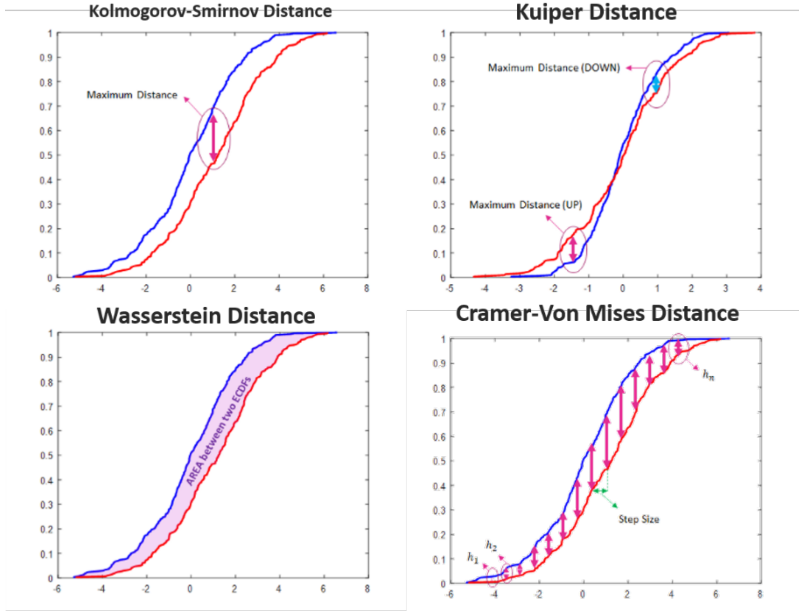


Figure 2: ECDF-based statistical distance measures ([2])

Introduction (Contd..)

- **Our two-fold approach**
  - Measure Statistical Distance Dissimilarity (SDD) *across time series* using ECDF-based distance measures.
  - Estimate ML-system properties like *reliability* and *robustness* by taking SDD into account.

Related Work

Dataset Shift

- **Definition: Dataset shift occurs when the joint distributions of training and test differs, i.e. $P_{tr}(y,x) \neq P_{tst}(y,x)$ [3].**
- **Types of shifts: Covariate shift, Prior Probability shift and Concept shift**
- **Limitation of existing methods**
  - Do not consider statistical distance measures.

Related Work (Contd..)

Reliability

- **Reliability in dependable systems – Continuity of correct service [8]**
  - Measured using Mean Time Between Failures (MTBF)
- **Lack of agreeable reliability definition in machine learning.**
- **Reliability in ML-model driven systems – Continuous service provision + Correctness of the ML model response**
  - Related to vital performance indicator of ML system such as accuracy/inaccuracy, availability, downtime etc. [9]
  - Offers diagnostic information to decision makers.
- **Model-agnostic and model-specific approaches for reliability measurement.**
- **Reliability estimates:**
  - Distribution-based
  - Indicator-based (bounded between [0,1] or arbitrary range of values).

Related Work (Contd..)

Reliability

- **Existing approaches**
  - Regression problems: Sensitivity analysis [10], local cross-validation [11], Confidence estimation on neighbors’ errors [12], bootstrapping [13], a ML-based approach [26].
  - Classification problems: Reliability Assessment Model (RAM) [24].
- **Drawback**
  - Effect of Statistical Distance Dissimilarity (SDD) not taken into consideration.

Related Work (Contd.)

Robustness

- ***Definition*: Property to deliver acceptable behavior in the presence of invalid inputs or stressful environmental conditions [8, 25].**
- **Evaluating ML model's robustness against different vulnerabilities**
  - **Adversarial attacks**:
    - Adversarial robustness [16], Flip probability [23], TRADES [15]
  - **Distributional shift**:
    - Effective robustness [22], Relative accuracy [27], ECDF-based distance measures [17]
- **Limitation**
  - Lack of use of SDD to account for robustness against dataset shift.

Research Questions

1. **SDD-Accuracy Correlation- How does SDD relate to the performance of a model specifically for time series application?**

2. **SDD-based Reliability and Robustness- How can SDD be incorporated into the reliability and robustness measures of a model?**

Chapter 02

Reliability and Robustness measure

Proposed Method

- **Focus on time series applications**
- **Proposed method**
  - Dynamic Time Warping (DTW) distance based K-means clustering amongst training set and hold-out set.
  - Investigate behavior of performance with Statistical Distance Dissimilarity (SDD).
  - Compute reliability and robustness by taking SDD into account.
- **Proposed measures**
  - StaDRe: Statistical Distance based Reliability estimate
  - StaDRo: Statistical Distance based Robustness
- **StaDRe is an extension of *CONFINE* (Confidence estimation based on the Neighbors' Errors) metric presented in work [12].**

$$csCONFI_{NE} = 1 - \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i^2$$

Proposed measures

Reliability metric	Robustness metric
$StaDRe(X^*) = \frac{2 - \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i^2 - \frac{f(c,X^*)}{f(c,O)}}{2}$ ***f***: ECDF-based distance measure; **ε_i**: Error of neighbor *i* ; **X^***: Data instance; ***m***: Nearest neighbors; ***c***: cluster center; ***O***: Origin/Reference	$StaDRo(X^*, P_{min}) = \begin{cases} True & \frac{f(X,X^*)}{d_{P_{min}}} <= 1 \\ False & \frac{f(X,X^*)}{d_{P_{min}}} > 1 \end{cases}$ ***f***: ECDF-based distance measure; **X^***: Data instance; **P_{min}** : Minimum required Performance (or Maximum Error); **d_{Pmin}**: SDD at P_{min}

Experiments

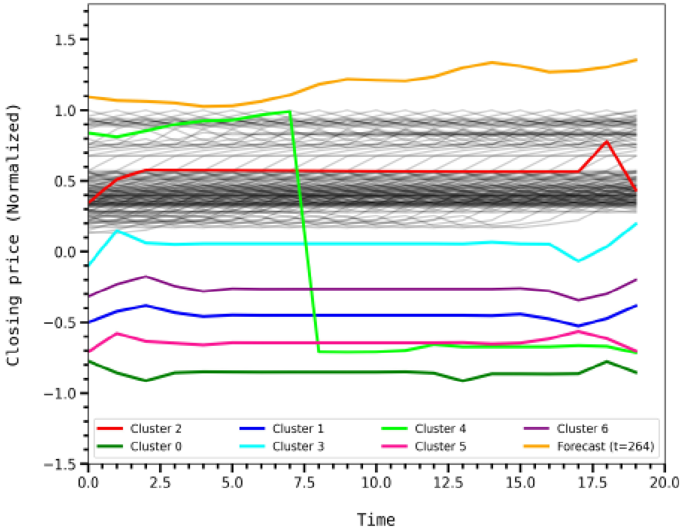
Application – Stock Price Prediction (financial-critical)

- **Datasets**
  - Stock’s closing price data of several companies.
  - **Data acquisition**: National Stock Exchange (NSE) library and Yahoo Finance.
  - The datasets that did not exhibit substantial drift between training and hold-out set were excluded.
  - Datasets were checked for missing values (if any) and cleaned before data preparation.

- **Implementation**
  - **Forecasting models**: LSTM [19] and GRU [21]
  - Training set: 80% ; hold-out set: 20%
  - One-step ahead prediction with window size tuned differently for each dataset.
  - All the data is normalized between -1 and 1.

Experiments (Contd.)

- ## Clustering and StaDRe
  - DTW based K-Means clustering was used to find nearest neighbors for time series data.
  - For MSE, value of m is equal to size of smallest cluster c'.
  - A random sampling for selecting m neighbors is applied when the assigned cluster c is not c'.
  - Origin is a reference time series of all 0s with same shape as c.

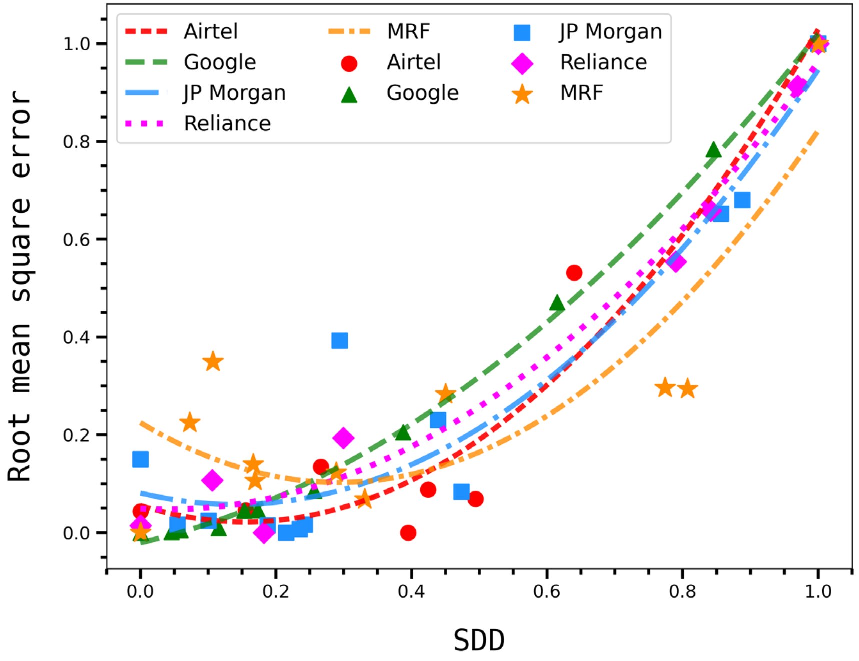


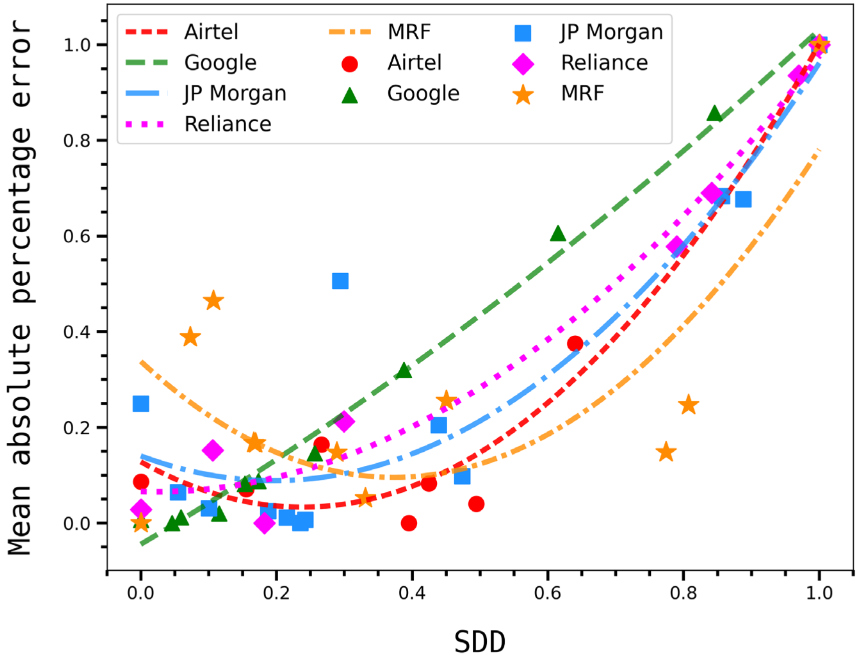
- ## StaDRo
  - Holdout set Y is divided into several subsets $Y^* \subset Y$ of length *l.*
  - The relationship between dataset SDD and the forecasting model's performance is derived by fitting a polynomial curve of degree 2.
  - Value of minimum performance is chosen based on specific requirement for each dataset.
  - $d_{P_{min}}$ is computed by solving for roots of polynomial equations.

Results

Performance vs. Statistical Distance Dissimilarity

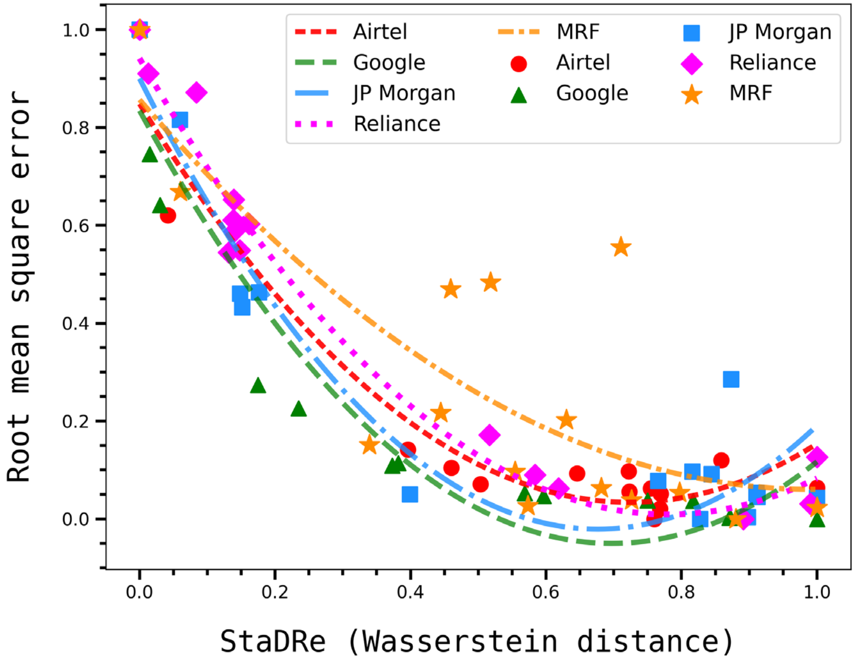
Wasserstein Distance

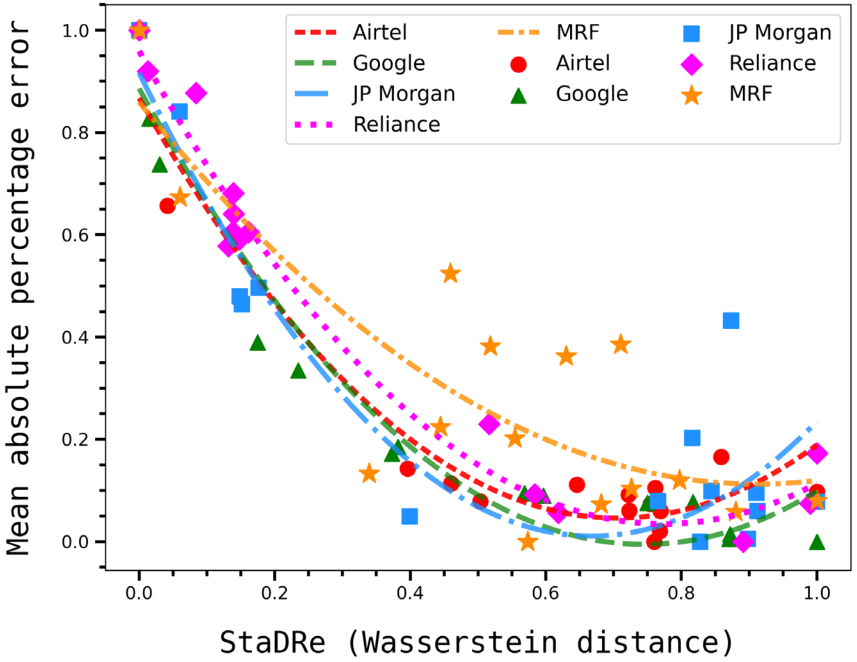




Results (Cont . . .)

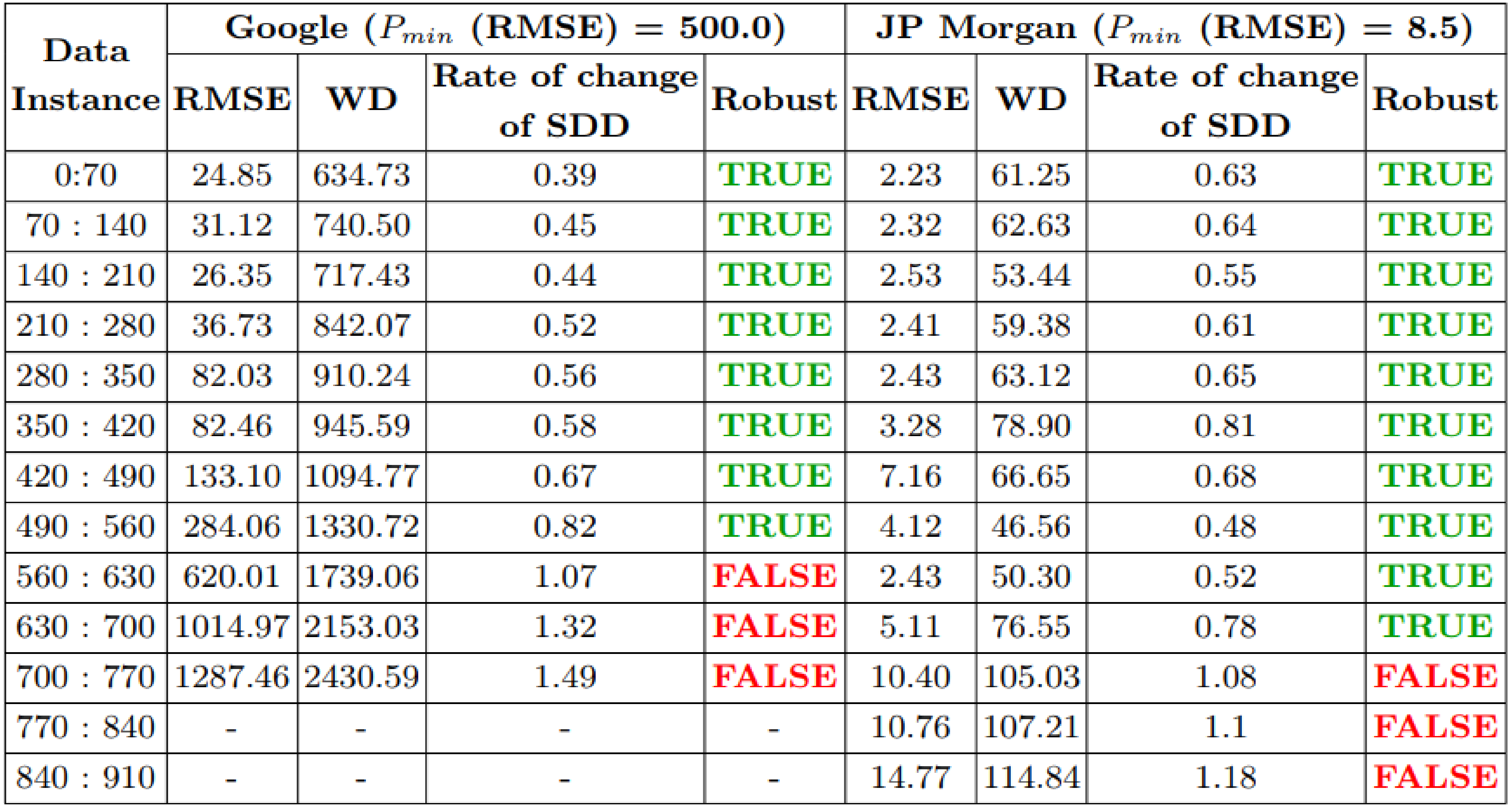
Performance vs. StaDRe (Wasserstein Distance)





Results (Cont . . .)

StaDRo for Google and JP Morgan data



Data Instance	Google (P_{min} (RMSE) = 500.0)				JP Morgan (P_{min} (RMSE) = 8.5)			
	RMSE	WD	Rate of change of SDD	Robust	RMSE	WD	Rate of change of SDD	Robust
0:70	24.85	634.73	0.39	TRUE	2.23	61.25	0.63	TRUE
70 : 140	31.12	740.50	0.45	TRUE	2.32	62.63	0.64	TRUE
140 : 210	26.35	717.43	0.44	TRUE	2.53	53.44	0.55	TRUE
210 : 280	36.73	842.07	0.52	TRUE	2.41	59.38	0.61	TRUE
280 : 350	82.03	910.24	0.56	TRUE	2.43	63.12	0.65	TRUE
350 : 420	82.46	945.59	0.58	TRUE	3.28	78.90	0.81	TRUE
420 : 490	133.10	1094.77	0.67	TRUE	7.16	66.65	0.68	TRUE
490 : 560	284.06	1330.72	0.82	TRUE	4.12	46.56	0.48	TRUE
560 : 630	620.01	1739.06	1.07	FALSE	2.43	50.30	0.52	TRUE
630 : 700	1014.97	2153.03	1.32	FALSE	5.11	76.55	0.78	TRUE
700 : 770	1287.46	2430.59	1.49	FALSE	10.40	105.03	1.08	FALSE
770 : 840	-	-	-	-	10.76	107.21	1.1	FALSE
840 : 910	-	-	-	-	14.77	114.84	1.18	FALSE

Conclusion and Future scope

- **SafeML support for SDD-based compliance assessment of ML models for time series applications.**
- **Inverse correlation between performance and SDD.**
- **Effective metrics (StaDRe and StaDRo) for runtime monitoring of reliability and robustness.**
- **Future Works**
  - Safety-critical application such as ECG-monitoring.
  - Application on Multivariate time series and Vision use case.
  - StaDRo for adversarial examples.
  - Foundation for enhancing explainability and interpretability of ML models for time series applications.

GitHub **https://github.com/n-akram/TimeSeriesSafeML**

Fraunhofer
IESE

Fraunhofer Institute for Experimental Software Engineering IESE

Contact

Mohammed Naveed Akram
Dept. Safety Engineering (SAF)
Tel. +49 631 6800-2253
naveed.akram@iese.fraunhofer.de

Fraunhofer IESE
Fraunhofer Platz 1,
67663 Kaiserslautern
www.iese.fraunhofer.de

Fraunhofer
IESE

Fraunhofer Institute for Experimental
Software Engineering IESE

Thank you for your Attention

Questions ?

Fraunhofer
IESE

Fraunhofer Institute for Experimental
Software Engineering IESE

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Koorosh Aslansefat et al. “Toward Improving Confidence in Autonomous Vehicle Software: A Study on Traffic Sign Recognition Systems”. In: Computer 54 (Aug. 2021), pp. 66–76. doi: 10.1109/MC.2021.3075054

3. Jose Moreno-Torres et al. “A unifying view on dataset shift in classification”. In: Pattern Recognition 45 (Jan. 2012), pp. 521–530. doi: 10.1016/j.patcog.2011.06.019

4. Haider Raza, Girijesh Prasad, and Yuhua Li. “EWMA model based shift-detection methods for detecting covariate shifts in non-stationary environments”. In: Pattern Recognition 48 (Mar. 2015), 659–669. DOI: 10.1016/j.patcog.2014.07.028

5. Rabanser, S., G¨unnemann, S., Lipton, Z.: Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems 32 (2019)

6. Becker, A., Becker, J.: Dataset shift assessment measures in monitoring predictive models. Procedia Computer Science 192, 3391–3402 (2021). https://doi.org/https://doi.org/10.1016/j.procs.2021.09.112,https://www.sciencedirect.com/science/article/pii/S1877050921018512, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021

References

7. Oliveira, G.H., Cavalcante, R.C., Cabral, G.G., Minku, L.L., Oliveira, A.L.: Time series forecasting in the presence of concept drift: A pso-based approach. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI).pp. 239–246. IEEE (2017)

8. A. Avizienis et al. “Basic concepts and taxonomy of dependable and secure computing”. In: *IEEE Transactions on Dependable and Secure Computing* 1.1 (2004), pp. 11–33. doi: 10.1109/TDSC.2004.2.

9. Zoran Bosnic and Igor Kononenko. “An overview of advances in reliability estimation of individual predictions in machine learning”. In:Intell. Data Anal.13 (Apr. 2009),pp. 385–401.doi:10.3233/IDA-2009-0371(cit. on p. XI).

10. Bosni´c, Z., Kononenko, I.: Estimation of individual prediction reliability using the local sensitivity analysis. Applied intelligence 29(3), 187–203 (2008)

11. DEMUT, I.R.: Reliability of predictions in regression models. Doktorandske dny’10 (2010).

12. Briesemeister, S., Rahnenf¨uhrer, J., Kohlbacher, O.: No longer confidential: estimating the confidence of individual regression predictions. PloS one 7(11), e48723 (2012)

13. Efron, B.: Bootstrap methods: another look at the jackknife. In: Breakthroughs in statistics, pp. 569–593. Springer (1992)

References

14. Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. Operations research 58(4-part-1), 902–917 (2010)

15. Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L.E., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 7472–7482. PMLR (09–15Jun 2019), https://proceedings.mlr.press/v97/zhang19p.html

16. Rauber, J., Brendel, W., Bethge, M.: Foolbox: A python toolbox to benchmark the robustness of machine learning models. arXiv preprint arXiv:1707.04131 (2017)

17. Koorosh Aslansefat et al.SafeML: Safety Monitoring of Machine Learning Classifiers through Statistical Difference Measure. 2020. arXiv:2005 . 13166 [cs.LG].

18. Hond, D., Asgari, H., Jeffery, D., Newman, M.: An integrated process for verifying deep learning classifiers using dataset dissimilarity measures. International Journal of Artificial Intelligence and Machine Learning (IJAIML) 11(2), 1–21 (2021)

19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735

20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR abs/1412.6980 (2015)

References

21. Cho, K., van Merri¨enboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder–decoder approaches. In: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. pp. 103–111. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/W14-4012, https://aclanthology.org/W14-4012

22. Rohan Taori et al. Measuring Robustness to Natural Distribution Shifts in Image Classification. 2020. arXiv: 2007.00644 [cs.LG]

23. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.:AugMix: A simple data processing method to improve robustness and uncertainty. Proceedings of the International Conference on Learning Representations (ICLR) (2020)

24. Zhao, X., Huang, W., Banks, A., Cox, V., Flynn, D., Schewe, S., Huang, X.: Assessing the Reliability of Deep Learning Classifiers Through Robustness Evaluation and Operational Profiles. In: AISafety'21 Workshop at IJCAI'21. vol. 2916. ceurws.org (2021)

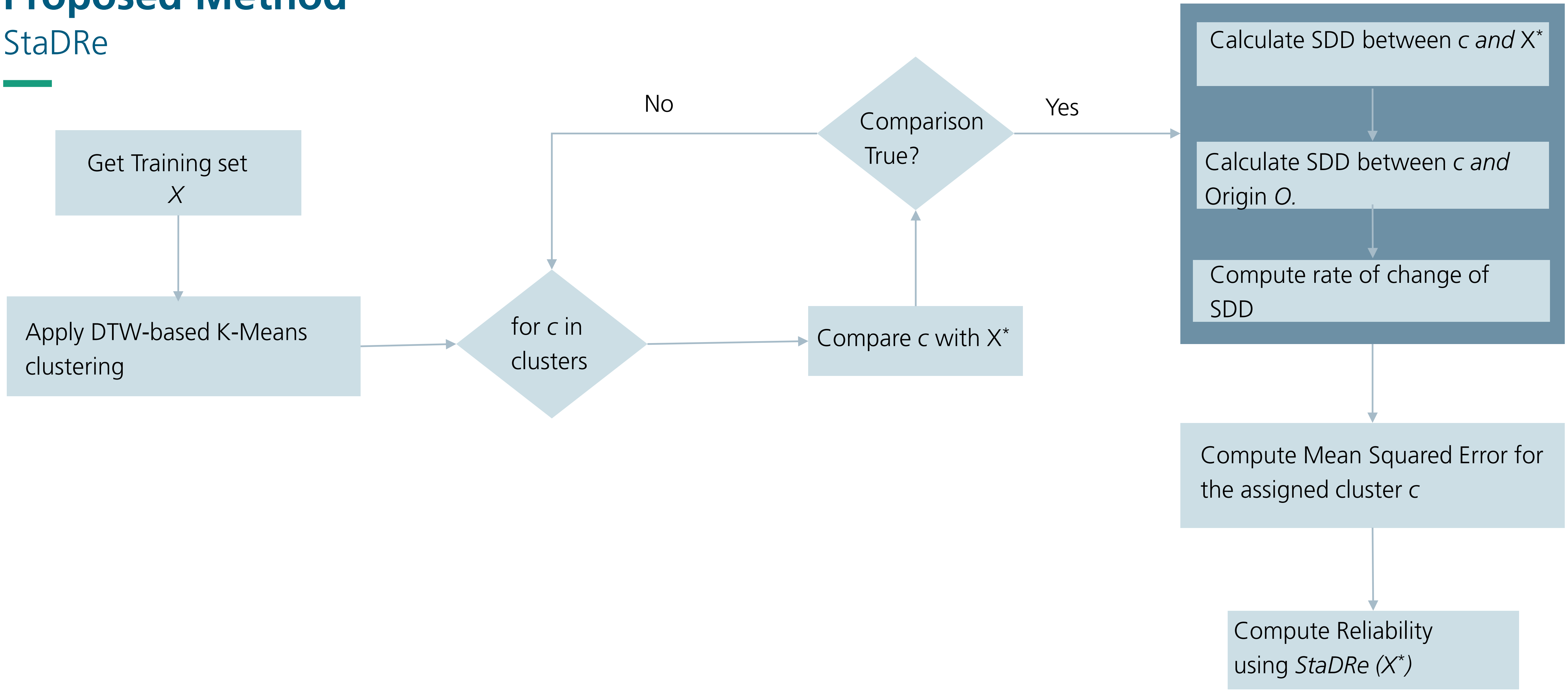
25. IEEE: Standard glossary of software engineering terminology. IEEE Std 610.12-1990 pp. 1–84 (1990). https://doi.org/10.1109/IEEESTD.1990.101064

26. Adomavicius, G., Wang, Y.: Improving reliability estimation for individual numeric predictions: a machine learning approach. INFORMS Journal on Computing (2021)

27. Santurkar, S., Tsipras, D., Madry, A.: Breeds: Benchmarks for subpopulation shift.arXiv preprint arXiv:2008.04859 (2020)

Proposed Method

StaDRe



Get Training set *X*

Apply DTW-based K-Means clustering

for c in clusters

Compare c with X^*

Comparison True?

No

Yes

Calculate SDD between *c and* X^*

Calculate SDD between *c and* Origin *O.*

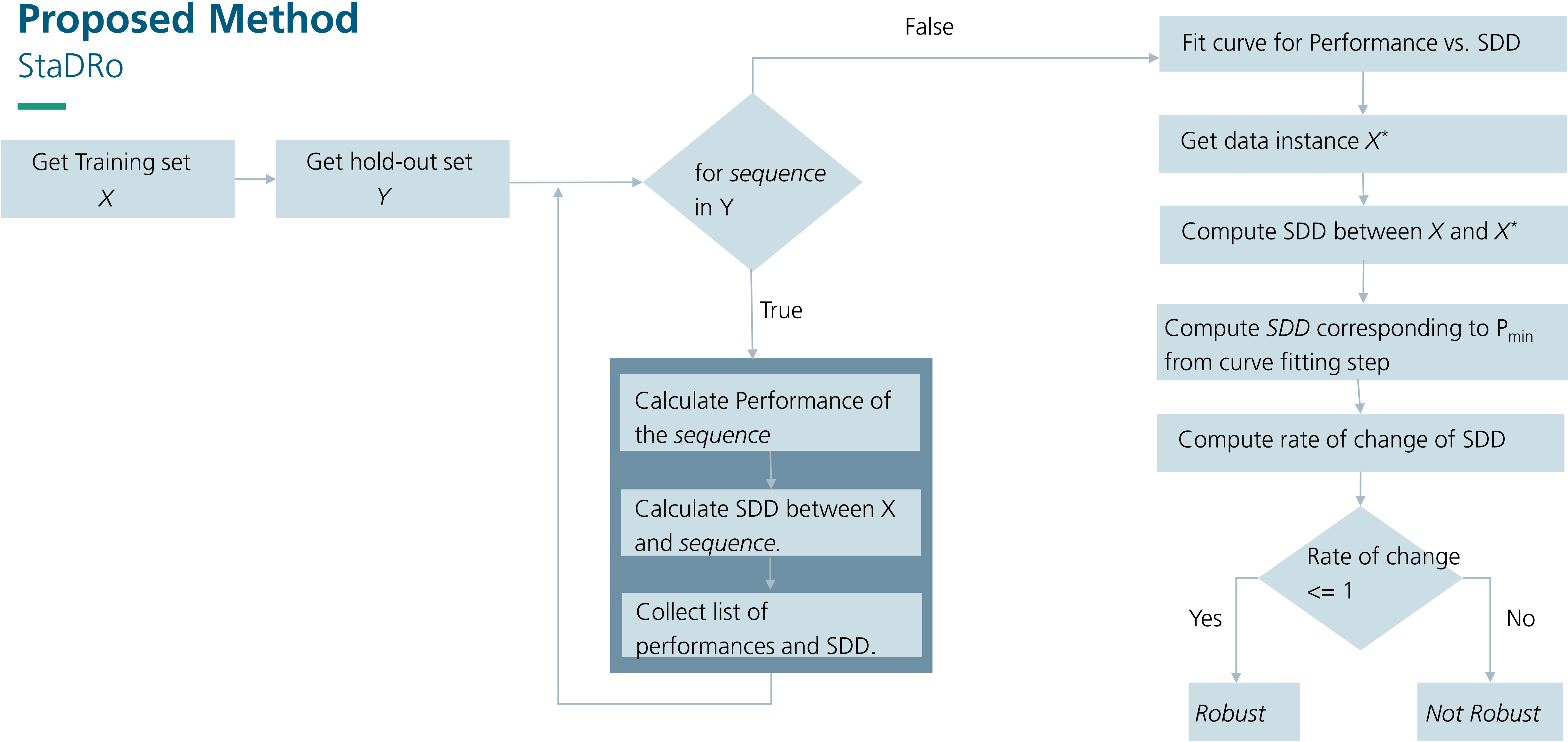
Compute rate of change of SDD

Compute Mean Squared Error for the assigned cluster c

Compute Reliability using *StaDRe* (X^*)

Proposed Method

StaDRo



Get Training set X
Get hold-out set Y
for *sequence* in Y
False
True
Calculate Performance of the *sequence*
Calculate SDD between X and *sequence*.
Collect list of performances and SDD.
Fit curve for Performance vs. SDD
Get data instance X^*
Compute SDD between X and X^*
Compute *SDD* corresponding to P_{min} from curve fitting step
Compute rate of change of SDD
Rate of change <= 1
Yes
No
Robust
Not Robust

Appendix

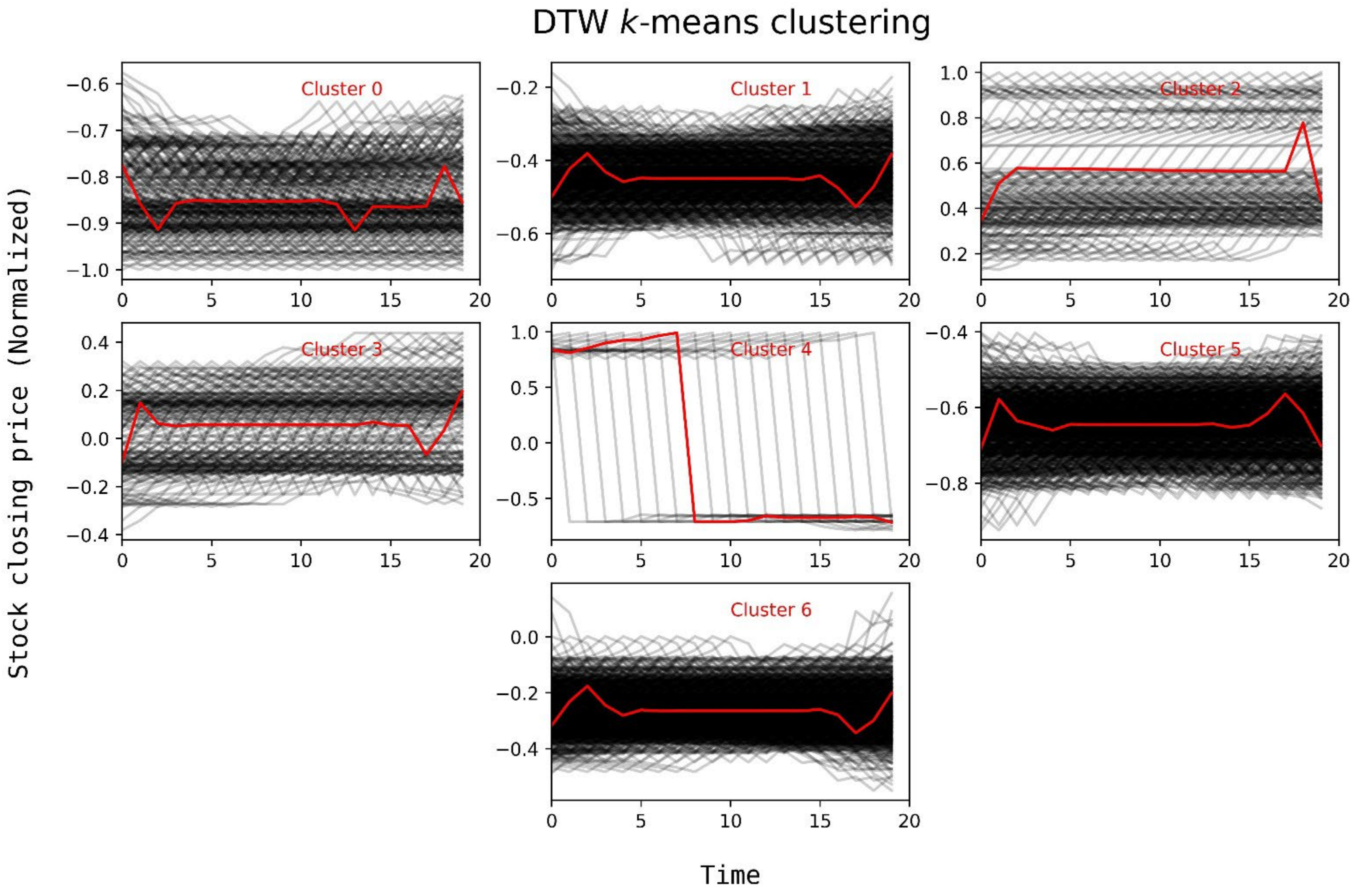
DTW-based K-Means clustering

Algorithm : DTW based K-means clustering

```
Result: Clusters
$X$ = GetTrainingSet();
_ , _ , Silhouette_score = GetDTWBasedClusters(X, 2);
Repeat;
for ClusterNumber in range(min, max, step) do
    cluster, centroid, Score = GetDTWBasedClusters ($X$, ClusterNumber);
        plot(clusters, centroids);
        if ManualCheckOfClusters(clusters, centroids) then
            Silhouette_score = Score;
            break;
        else if Silhouette_score != Score then
            Silhouette_score = Score;
end
return Clusters
```

Appendix

Clustering example for Reliance stock price data



Appendix

Dataset and Implementation details

Characteristic	Dataset				
	Reliance	***Google***	***Airtel***	***JP Morgan***	***MRF***
Start date (dd.mm.yyyy)	04.01.2010	03.01.2005	04.01.2010	02.01.2003	03.01.2005
No. of Data points	2895	4300	3008	4803	4214
Window size	20	50	20	20	20
Number of clusters	7	5	6	5	5
LSTM (# layers; hidden size)	2; 32	2; 40	2; 32	2; 64	2; 32
GRU (# layers; hidden size)	2; 32				