Final report:
Uncertainty Quantification and Concept Drift

Līva Freimane and Haoyu Wei

December 2022

1 Overview

Machine learning models are inherently uncertain. Uncertainty quantification aims to determine how likely certain outcomes are if some aspects of the respective system are not precisely known.

The quantitative characterization and management of uncertainty include the detection of concept drifts, statistical inference, model calibration, and decision-making under uncertainty. UQ not only focuses on foundational ideas in mathematics and statistics but also proposes techniques in applications using computational simulations, such as modeling complex systems in socio-economy and nature science [9].

Concept drift is a phenomenon in which the distribution of the input data changes over time. This may mean changes in the underlying distribution of the data (virtual drift) or changes in the modeled relationship (real concept drift), or changes in both.

In machine learning methods, UQ can be divided into two groups, discriminative methods, and generative methods. The former aims to perform classifications, while the latter concerns data generation or reconstruction. Bayesian inference can be utilized for some discriminative model-agnostic methods. In complex deep learning models, UQ can be used for both neural network calibration and concept drift detection [25].

The topics of uncertainty quantification and concept drift, both deal with the challenges of making predictions and decisions in the presence of uncertainty. In the case of concept drift, the uncertainty arises from the changing nature of the data or environment, while in the case of uncertainty quantification, the uncertainty may come from various sources. Techniques for uncertainty quantification can be used to evaluate the impact of concept drift on a machine learning system and to develop strategies for dealing with it.

In this project, we will focus on uncertainties related to concept drift and on UQ methods usage in cutting-edge machine learning research. In Chapter 2, we present how UQ can enhance trustability in machine learning modeling, and give categorization and examples of state-of-the-art UQ methods. In Chapter 3, we will look at concept drift definition and classification, and give an overview

of concept drift detection methods, not going into details about any specific methods. In Chapter 4, we present several examples in UQ and concept drift. And finally, Chapter 5 consists of a discussion and a conclusion.

2 Uncertainty and uncertainty quantification methods in Machine Learning

Uncertainties are in everyday scenarios in various fields. When we develop models using Deep Learning (DL) and Machine Learning (ML), for making predictions and decisions, it is important to take uncertainties into consideration. It is thus highly desirable to represent uncertainty in a trustworthy manner in any AI-based system and what we need is UQ methods.

There are two main types of uncertainty, aleatoric and epistemic uncertainties. Aleatoric uncertainty refers to irreducible uncertainty, which is not in the model but an inherent property of the data distribution. In contrast, epistemic uncertainty comes from the lack of knowledge. In machine learning and deep learning models, epistemic uncertainty usually refers to a probability distribution over the model parameters [19].

ML/DL models start with a collection of potentially relevant datasets for the decision-making process and they are designed to meet some performance goals via appropriate algorithms or DL architecture. For the cases where we have sufficient data resources, the massive collection of data can be information poor, and some of the data may be incomplete, noisy, or discordant. UQ aims to help understand the reliability and confidence of a model's predictions. In AI research, UQ can be used to pursue two main goals. First, to calibrate models, mostly for neural networks, so that the output confidence reflects the empirical accuracy; and second, to make reliable predictions for out-of-distribution detection, i.e., to detect how a new classification example corresponds to the trained data distribution [25].

In the last 10 years, 2500+ papers focusing on the use of UQ in AI were published [21]. Here we present a comprehensive categorization of research topics and different UQ methods for ML/DL models. As we mentioned in the overview, UQ can be divided into discriminative methods and generative methods, which respectively aim at classifications and data generation. For discriminative models, we can distinguish between three main types of UQ methods in ML [25]:

- **Model-agnostic approaches:** This is a general term for UQ approaches that are independent of the model. For example, better-calibrated DNNs with data augmentation [15].
- **Bayesian methods:** UQ methods that use Bayesian Inference. Bayes' theorem is used to update the probability, and evidence or information becomes available for estimating the uncertainties.
- **Non-Bayesian methods:** These include UQ approaches that are not based on Bayesian inference. Bayesian approaches require modifications

in the training the model, and sometimes Bayesian methods are computationally expensive.

For UQ methods in DL, the Bayesian UQ methods include several subdivision methods to interpret the model parameters robustly; and the non-Bayesian UQ methods are commonly referred to as ensemble techniques. Here we present some uncertainty quantification using Bayesian techniques and ensemble techniques [21]:

- **Bayesian UQ techniques in DL:** To provide information about the reliability of the predictions, Bayesian deep learning (BDL) and Bayesian NNs (BNNs) can be used to interpret the model parameters. BNNs and BDL are robust to overfitting problems and can be used for both small and large datasets. Bayesian methods can be used for approximate inference to learn the posterior distribution of the parameters. The generally used Bayesian approximate methods include (we don't present the technical details of the following methods, [1, 20, 2] are related papers and it is worthwhile to read them):
  - Markov chain Monte Carlo (MCMC)
  - Variational Inference (VI)
  - Laplacian approximation

  Bayesian UQ methods also can combine the dropout method to avoid overfitting problems in DL. Monte Carlo (MC) dropout is an effective method, [13] is an example of MC dropout with a bootstrap ensembling-based method for the task of vehicle control.

  Another common way to utilize Bayesian UQ methods is to combine them in DL architecture. An example is Variational autoencoders (VAEs), which is a variant of DL consisting of an encoder and a decoder. The encoder aims to map the high-dimensional inputs to low-dimensional latent variables, and the decoder reconstructs inputs using variables. The latent variables conform to a given prior distribution, and VAE is an effective method to model the posterior using variational inference [5].

  With the continuous development of ML, Bayesian UQ methods have been applied in various ML frameworks. Active learning aims to learn from unlabeled samples involving actively most informative data points, and it can utilize Bayesian UQ methods to represent uncertainty among massive data, e.g. deep Bayesian active learning methods [10]. Reinforcement learning frameworks can use Bayesian UQ methods to estimate the performance and robustness of the model in various fields, and Bayesian UQ methods' usage has been widely-investigated in the literature [16].

- **Ensemble techniques in DL:** Ensemble techniques are another large category of UQ methods, mainly for DL. Instead of estimating uncertainties via Bayesian inference, ensemble techniques target to use of an

ensemble of models as a more powerful model to enhance predictive performance.

One common way to use ensemble techniques for uncertainty quantification is to create a diverse set of models trained on different subsets of the data. This can help to capture a wider range of possible relationships and features in the data, and it can also lead to more robust predictions. For example, [18] proposed test-time augmentation to improve the performance of different ensemble learning techniques, and demonstrated how an ensemble of several trained deep neural networks can be as good as to many refined neural network methods with respect to test performance.

One important limitation of the ensemble method is the weights of different ensembles in compositions are usually uninformed rather than based on the single model's reliability. [6] employed Bayesian model averaging, where both the reliability and uncertainty of every single model were considered in estimating the weights in ensemble learning. The deep ensemble approaches combining Bayesian methods are named as deep Bayesian ensemble, attracting more and more attention these years.

Overall, ensemble techniques can be a useful tool for uncertainty quantification because they can build powerful models and provide a more comprehensive view of the data. Ensemble techniques can help to reduce the risk of overfitting or making overly confident predictions.

To sum up, uncertainties are an inherent part of any prediction or decision-making process, and it is important to understand and quantify them for more informed, reliable, and robust results.

In a virtual laboratory and digital twins, uncertainty quantification is important for a number of reasons. In virtual experiments, data is often generated from simulations and emulations, which can introduce uncertainties due to the approximations and assumptions in the simulating. Additionally, virtual experiments may be subject to measurement errors or other sources of uncertainty, such as the inherent variability of the system. Quantifying these uncertainties can help better understand the limitations of their virtual experiments and the reliability of the results [19]. UQ can help to improve the accuracy and reproducibility of the results, as well as the overall credibility and usefulness of the virtual laboratory.

3 Concept drift

Most machine learning models are static, but our world is dynamic. As a result of this, the model's accuracy can decay with time. This may happen due to a phenomenon called concept drift. Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time [12].

Concept drift is a change in the joint distribution between two time instances t and $t + w$ where t could be a particular time point or time interval, and w denotes the time window when the distribution change is being checked at.

Formally, concept drift occurs, if

$$\exists t : P_t(X, y) \neq P_{t+w}(X, y). \tag{1}$$

The joint distribution $P_t(X, y)$ can be decomposed and rewritten as

$$P_t(X, y) = P_t(X) \times P_t(y|X), \tag{2}$$

where $P_t(X)$ is the input data probability distribution and $P_t(y|X)$ - posterior probability distribution of the target labels, i.e., the decision boundary [26]. This concept drift definition is crucial for the classification of concept drift as changes in the data stream can be characterized by changes in the components of the equation (2). Different papers have different classifications of concept drift, we use the classification given in [12]. The classification distincts between three sources, namely:

- Input data probability distribution changes, but the decision boundary remains the same. This kind of change is called *feature drift, covariate shift* or *virtual drift.*
- Decision boundary changes, but the input data probability distribution remains the same. This is referred to as real *concept drift, concept drift* or *actual drift.*
- Both decision boundary and input data probability distribution change.

[26] gives an overview of the terminology used based on the probabilistic source of change. A further classification is made by the authors of [7] based on the rate at which the concept evolves. A drift may happen:

- suddenly/abruptly, by switching from one concept to another (e.g., replacement of a sensor with another sensor that has a different calibration in a chemical plant)
- incrementally, consisting of many intermediate concepts in between (e.g., a sensor slowly wears off and becomes less accurate)
- gradually (e.g., when a new road is constructed which is a shortcut going from A to B, people tend to use the old road along with the new due to their habits but eventually use the new afterward).
- There are also recurring concept drifts, which are patterns or trends that tend to repeat themselves at intervals, and are commonly found in seasonal data.

One of the challenges for concept drift handling algorithms is not to mix the true drift with an outlier or noise, which refers to a once-off random deviation or anomaly. No adaptivity is needed in the latter case [7].

Drift detection refers to the techniques and mechanisms that characterize and quantify concept drift via identifying change points or change time intervals. Concept drift detection methods can be divided into supervised (requiring

ground truth values) and unsupervised approaches [26]. The authors [26] classify concept drift methods into four categories:

- **Data distribution based detection** - these methods use distance measures to estimate the similarity between the data distributions at two different time windows. Concept drift is then detected if the two distributions are significantly distant. The main advantage of these approaches is that they can be applied to both labeled and unlabeled data sets since the methods only consider the distribution of data points. However, changes in the data distribution don't always lead to changes is performance. An example of a data distribution based detection algorithm is the Equal Intensity k-Means Space Partitioning algorithm in [23].
- **Performance-based detection** - these methods typically trace the deviations in the model's output error to detect changes. These approaches only handle the change when the model's performance is affected. The majority of these methods require quick arrival of ground truth values for the predictions, which may not be easily available. An classical example of performance based drift detection is the Drift Detection Method [3].
- **Multiple hypothesis-based detection** - these approaches apply several detection methods and aggregate their results in parallel or in hierarchy. [11] proposes a concept drift detector called Hierarchical Linear Four Rates detector that is an example of this subsection of detectors.
- **Contextual-based detectors** - use context information available from the system and data to detect the drift. A framework for context-aware drift detection algorithm can be found in [27]

4 Relevance and application examples

4.1 UQ application examples in simulator and virtual experiments

UQ methods play a significant role in reducing the impact of uncertainties during optimization and decision-making processes. And UQ methods have been widely applied in various methods, and [21] provides a comprehensive review of UQ methods' application and developments in Deep learning.

In virtual experiments, UQ can help to analyze errors or other sources of uncertainty during data collections and simulations and to understand the reliability of the results. Here we consider an example of UQ application in natural science and virtual experiments, in the domain of Quantum molecular dynamics simulations and chemistry experimentation.

- **Bayesian machine learning for quantum molecular dynamics, based on [14]**

In quantum dynamics calculations in theoretical chemistry, solving the Schrödinger equation can help to predict the outcome or understand the mechanisms of the microscopic interactions of molecules with other molecules.

The system of differential equations can generally be written as:

$$[\hat{D}_r I + U(V)]\Psi(r) = 0 \tag{3}$$

where $\hat{D}_r$ is a differential operator acting on functions of r, $W(r)$ is a vector of N_b basis set expansion coefficients, each depending on r, U is an $N_b * N_b$ Hermitian matrix that depends on V, and I is the $N_b * N_b$ identity matrix. With the given $V(r)$, a set of N-dimensional potential energy surfaces, the equation can be solved numerically in appropriate boundary conditions to compute observables such as the bound state energies, the probabilities of molecular collision outcomes, or the chemical reaction rates.

However, there are two major challenges in solving this equation in a traditional way. Firstly, the time complexity can scale up as quickly as $O({N_b}^3)$, and secondly, the matrix U is parametrized by the N-dimensional potential energy surfaces $V(r)$.

To tackle those challenges, [14] combined the Bayesian statistics, and developed a Bayesian machine learning based simulator into the equation, to obtain not only the quantum predictions but also the error bars of the dynamical results on uncertainties from inputs. Bayesian machine learning makes the targeted equation with a non-parametric distribution of potential energy surfaces, conditioned by the desired dynamical properties.

With Gaussian processes (GP), a non-parametric Bayesian machine learning method, the problem can be formulated as a model $y(x)$, y represents the output of the equation (in [14] is the reaction probability), and x is a vector of all parameters defining the equation. With a random combination of x and the corresponding numerical results y as the training dataset, the GP model can be used as a simulator to predict the the reaction probabilities at any combination of the parameters x. The GP model can give the prediction of the equation output with the dynamical calculations and the relative error is only around 4%. The GP model can also be used to estimate the error bars of the uncertainties, from inherent error in quantum chemistry and calculation to get potential energy surfaces $V(r)$. Figure 1 gives an example of uncertainty approximation.

4.2 Concept drift examples

The phenomenon of concept drift is usually considered in the context of data streams. Real-world data streams pose the challenge of concept drift to the implementation of machine learning models and data analysis [24]. The presence of concept drift can make prediction results inaccurate and therefore can lead to sub-optimal decisions. Thus there is a need to enhance intelligent systems operating on real-world data streams with concept drift-aware learning machine

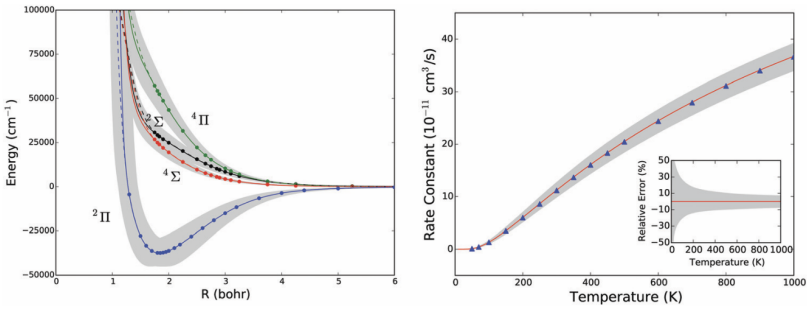

Figure 1: Left panel: The adiabatic interaction potentials, and the grey regions show the estimated uncertainty of the potentials. Right panel: The results of rigorous quantum scattering calculations, with the grey area showing the uncertainty of the collision rate by the GP model.

learning frameworks, that would ensure the validity of the model predictions [4, 8].

We will consider two cases where concept drift can occur - one in real-world data of fossils and the other - the importance of concept drift in virtual laboratories such as Destination Earth.

- **Concept drift in fossil data, based on [17]**

  Fossils are the remains of organisms from earlier geological periods preserved in rock. One of the major directions in the computational analysis of fossil data is to reconstruct environmental conditions and track climate changes over millions of years. The distribution of fossil animals in space and time makes informative features for such modeling. Labeled data linking organisms to climate is available only for the present day, where climatic conditions can be measured. The approach is to train models in the present day and use them to predict climatic conditions over the past.

  As species continuously go extinct and new species originate, animal communities today are different from the communities of the past, and the communities at different times in the past are different from each other.

  One of the main challenges for such modeling is that present species are not the same as species in the past. The further to the past, the less overlap between the species lists at present and in the past is expected due to continuously ongoing evolution, the origination of new species, and the extinction of past species and immense body of evolutionary theory thereafter. Thus, from the data perspective, there is a continuously ongoing concept drift.

  The closer to the present day (left side of the timeline in Figure 2), the larger the overlap with the genera that are alive today. No genera are the same as genera observed eight million years ago. If we were to use presence

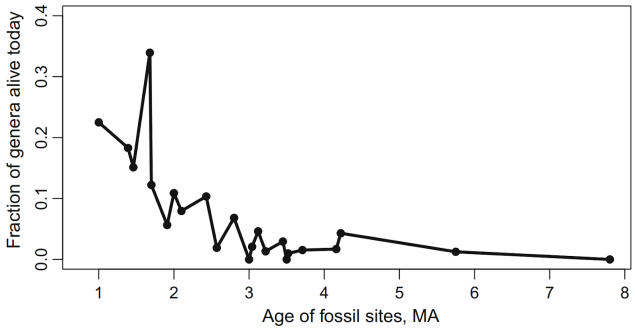

Figure 2: Summary of concept drift in the Turkana Basin fossil data [17]

or absence of species as features for predictive model, there would be a gradual, but catastrophic drift in the feature space.

- **Concept drift relevance in Virtual laboratories (Destination earth), based on [22]**.

  Destination Earth is a major initiative of the European Commission. It aims to develop a very high-precision digital model of the Earth (a Digital Twin) to monitor and predict environmental change and human impact to support sustainable development.

  Destination Earth aims to use Digital Twins which will ensure live coupling between the physical asset and its digital twin via multiple streaming data sources originating from live sensing of the physical process. Digital Twins will combine data from real-time observations and simulations.

  Digital Twins are a core part of virtual laboratories. This kind of workflow (live coupling with data from sensors) can have unknown dynamics in the steaming data, which is why concept drift might be an issue, therefore concept-drift-aware machine learning systems are necessary.

5 Discussion and Conclusions

Given a prediction, it is necessary to understand its reliability. The ability to make informed decisions under uncertainty is crucial for the reliable deployment of machine learning systems. Concept drift has been recognized as the root cause of decreased effectiveness in many data-driven information systems [12], which makes it an important aspect of virtual laboratories in the context of digital twins and emulators.

Interpretable machine learning methods and UQ are related - the methods of Explainable AI try to show the way to the decision, while the methods of UQ try to give a realistic evaluation regarding the reliability of the decision. State-of-the-art

In [26] it is noted that drift detection for classification tasks is the main scope of detecting concept drift in machine learning tasks while little attention is paid to a regression setting. There is no single drift detector that works better than all the others in all scenarios.

In [12] the authors note that regarding concept drift understanding, all drift detection methods can answer "When" (as in when and for how long the drift occurs), but very few methods have the ability to answer "How" (how severe) and "Where" (where the drift region is). A method that would answer these questions might advance drift adaptation.

The main challenge of UQ methods in machine learning is to make reliable predictions in new deep learning architecture. In [25], the author proposes the combination of explainable AI and UQ as a research direction.

In this project, we considered uncertainty quantification and concept drift - its definition, classification by drift source, and by the rate at which the concept evolves. We gave examples that link uncertainty quantification and concept drift to real-life data, and data streams which are crucial in building digital twins for virtual laboratories.

Follow-up work related to concept drift could include an in-depth review of concept drift adaptation methods, as well as the classification of such methods. Additionally, one could research and create a workflow for concept drift-aware machine learning system. For uncertainty quantification in virtual laboratories, further study could focus on a more comprehensive review of UQ methods in the whole digital twins building process, from data collection, simulation to maintenance. For more general machine learning challenges, the efficient and high-performance UQ methods could also be a domain to be dug deeper into. Finally, the more direct relationship and connections of virtual laboratories and UQ/concept drift could be a topic for further studies as well.

References

[1] Matthew A Kupinski et al. "Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques". In: *JOSA A* 20.3 (2003), pp. 430–438.

[2] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] João Gama et al. "Learning with Drift Detection". In: vol. 8. Sept. 2004, pp. 286–295. ISBN: 978-3-540-23237-7. DOI: 10.1007/978-3-540-28645-5_29.

[4] Gregory Ditzler and Robi Polikar. "Incremental Learning of Concept Drift from Streaming Imbalanced Data". In: *IEEE Transactions on Knowledge and Data Engineering* 25.10 (2013), pp. 2283–2301. DOI: 10.1109/TKDE.2012.136.

[5] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[6] Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. "Sentiment analysis: Bayesian ensemble learning". In: *Decision support systems* 68 (2014), pp. 26–38.

[7] João Gama et al. "A Survey on Concept Drift Adaptation". In: *ACM Computing Surveys (CSUR)* 46 (Apr. 2014). DOI: 10.1145/2523813.

[8] Persson et al. Beyene Welemariam. "Improved concept drift handling in surgery prediction and other applications". In: *Knowl Inf Syst* 44 (2015), pp. 177–196. DOI: 10.1007/s10115-014-0756-9.

[9] Y. M. Marzouk and K. E. Willcox. "Uncertainty Quantification". In: *The Princeton Companion to Applied Mathematics*. Ed. by N. Higham et al. Vol. II. Princeton University Press, 2015. Chap. 34, pp. 131–134.

[10] HM Sajjad Hossain, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. "Active learning enabled activity recognition". In: *Pervasive and Mobile Computing* 38 (2017), pp. 312–330.

[11] Shujian Yu and Zubin Abraham. "Concept Drift Detection with Hierarchical Hypothesis Testing". In: June 2017, pp. 768–776. ISBN: 978-1-61197-497-3. DOI: 10.1137/1.9781611974973.86.

[12] Jie Lu et al. "Learning under Concept Drift: A Review". In: *IEEE Transactions on Knowledge and Data Engineering* PP (Oct. 2018), pp. 1–1. DOI: 10.1109/TKDE.2018.2876857.

[13] Christian Hubschneider, Robin Hutmacher, and J Marius Zöllner. "Calibrating uncertainty models for steering angle estimation". In: *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE. 2019, pp. 1511–1518.

[14] R. V. Krems. "Bayesian machine learning for quantum molecular dynamics". In: *Phys. Chem. Chem. Phys.* 21 (25 2019), pp. 13392–13410. DOI: 10.1039/C9CP01883B. URL: http://dx.doi.org/10.1039/C9CP01883B.

[15] Sunil Thulasidasan et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks". In: *Advances in Neural Information Processing Systems* 32 (2019).

[16] Xujiang Zhao et al. "Uncertainty-based decision making using deep reinforcement learning". In: *2019 22th International Conference on Information Fusion (FUSION)*. IEEE. 2019, pp. 1–8.

[17] I Žliobaitė. "Concept drift over geological times: predictive modeling baselines for analyzing the mammalian fossil record". In: *Data Min Knowl Disc* 33 (2019), pp. 773–803. DOI: 10.1007/s10618-018-0606-6.

[18] Arsenii Ashukha et al. "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning". In: *arXiv preprint arXiv:2002.06470* (2020).

[19] José Rıos et al. “Uncertainty of data and the digital twin: a review”. In: *International Journal of Product Lifecycle Management* 12.4 (2020), pp. 329–358.

[20] Jakub Swiatkowski et al. “The k-tied normal distribution: A compact parameterization of Gaussian mean field posteriors in Bayesian neural networks”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9289–9299.

[21] Moloud Abdar et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. In: *Information Fusion* 76 (2021), pp. 243–297.

[22] “Destination Earth Brochure”. In: (2021). URL: https://digital-strategy.ec.europa.eu/en/library/destination-earth.

[23] Anjin Liu, Jie Lu, and Guangquan Zhang. “Concept Drift Detection via Equal Intensity k-Means Space Partitioning”. In: *IEEE Transactions on Cybernetics* 51.6 (2021), pp. 3198–3211. DOI: 10.1109/TCYB.2020.2983962.

[24] Hassan Mehmood et al. “Concept Drift Adaptation Techniques in Distributed Environment for Real-World Data Streams”. In: *Smart Cities* 4 (Mar. 2021), pp. 349–371. DOI: 10.3390/smartcities4010021.

[25] Dominik Seuß. “Bridging the gap between explainable AI and uncertainty quantification to enhance trustability”. In: *arXiv preprint arXiv:2105.11828* (2021).

[26] Firas Bayram, Bestoun S. Ahmed, and Andreas Kassler. “From concept drift to model degradation: An overview on performance-aware drift detectors”. In: *Knowledge-Based Systems* 245 (2022), p. 108632. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.108632.

[27] Oliver Cobb and Arnaud Van Looveren. “Context-Aware Drift Detection”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 4087–4111.