Quantifying the dynamic predictability of train delay with uncertainty-aware neural networks

Journal Article

Author(s):
Spanninger, Thomas; Wiedemann, Nina; Corman, Francesco

Publication date:
2024-05

Permanent link:
https://doi.org/10.3929/ethz-b-000668642

Originally published in:
Transportation Research Part C: Emerging Technologies 162, https://doi.org/10.1016/j.trc.2024.104563

Funding acknowledgement:
181210 - DADA - Dynamic data driven Approaches for stochastic Delay propagation Avoidance in railways (SNF)

Transportation Research Part C 162 (2024) 104563

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Quantifying the dynamic predictability of train delay with uncertainty-aware neural networks

Thomas Spanninger [a,*,1], Nina Wiedemann [b,1], Francesco Corman [a]

[a] Institute of Transport Planning and Systems, ETH Zurich, Switzerland
[b] Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

ARTICLE INFO

Keywords:
Train delay prediction
Uncertainty estimation
Stochastic models
Predictability

ABSTRACT

The digital transformation of railway systems has sparked research in train delay prediction. While efforts have predominantly set on maximizing prediction accuracy, there remains a significant need to explore a deeper understanding of the prediction-associated uncertainty. This study proposes uncertainty-aware neural networks, extended with test-time-dropout and loss attenuation, to predict train delays and also estimate the level of associated confidence. Our approach outperforms commonly-used stochastic methods in terms of accuracy and precision. We further introduce a dynamic prediction horizon framework (DPHF) to systematically compare and validate uncertainty-enhanced predictions over time. We suggest the likeliness of realization (LoR) to evaluate predictions with confidence estimates and to quantify dynamic predictability, which we find to be best described by an exponential decay for an increasing prediction horizon. While the model-driven (epistemic) uncertainty remains relatively small and constant as the prediction horizon increases, the data-inherent (aleatoric) uncertainty is substantially larger and grows significantly. This indicates that the observed decay in predictability is not an artefact of the modelling process but indeed an inherent property of train delays. This study thus provides new insights that can be used to increase the robustness and reliability of railway operations, emphasizing innovative modelling and the utilization of emerging data sources.

1. Introduction

Train delays have significant economic and social impacts, affecting millions of people every day worldwide. Delayed train rides cause stress (van Hooff, 2015; Lyons and Chatterjee, 2008) and prevent people from using public transport (Martin et al., 2021; Lijesen, 2014), despite its sustainability and many further advantages. Predictions of train delays can help transportation operators and passengers better plan their journeys (Leng and Corman, 2020), reduce congestion and overcrowding, and improve overall efficiency and safety. Emerging computational technologies detecting patterns in large datasets have become available over the last twenty years to predict train delays based on historical data (offline) and information about the current situation (online).

In their review on train delay predictions, Spanninger et al. (2022) highlight that most contributions aim to generate accurate point estimates, which do not inform about their associated level of confidence. This is a problem because predictions are not always correct, and prediction uncertainty is important for both traffic controllers and users. Traffic controllers benefit from uncertainty-aware predictions by enhancing their risk-based decisions of traffic management to minimize the cascading consequences of train delays, including the re-routing of trains, the cancellation or keeping of planned connections and (partial) cancellations of services.

[*] Corresponding author.
E-mail address: tspanninger@ethz.ch (T. Spanninger).
[1] Authors contributed equally.

https://doi.org/10.1016/j.trc.2024.104563
Received 14 September 2023; Received in revised form 31 January 2024; Accepted 5 March 2024
Available online 11 April 2024

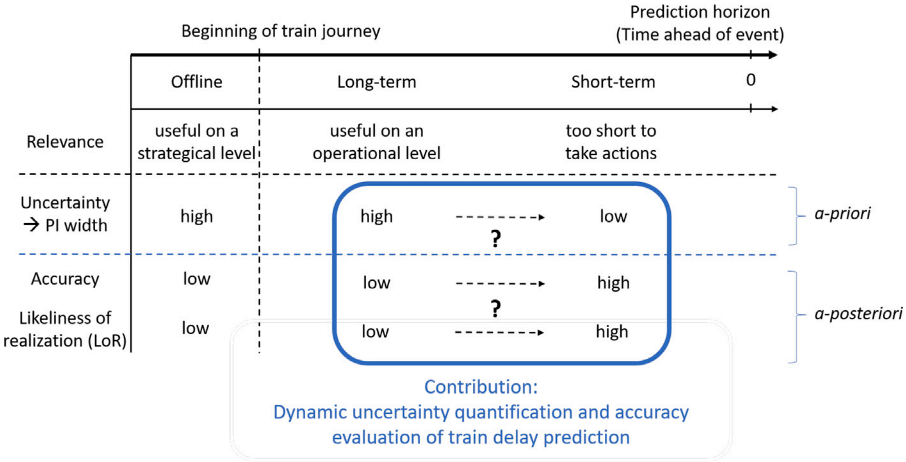

Beginning of train journey
Prediction horizon
(Time ahead of event)
Offline
Long-term
Short-term
0
Relevance
useful on a strategical level
useful on an operational level
too short to take actions
Uncertainty → PI width
high
high
low
?
a-priori
Accuracy
low
low
high
Likeliness of realization (LoR)
low
low
high
?
a-posteriori
Contribution:
Dynamic uncertainty quantification and accuracy evaluation of train delay prediction

Fig. 1. Relevance, accuracy and uncertainty of train delay predictions in a dynamic perspective.

Similarly, users need to know how likely a train will be delayed in order to make informed behavioural decisions (e.g. planning activities after their trip or when to arrive at the platform to avoid missing a train).

Another practical limitation of recent train delay prediction studies is their focus on short-term prediction horizons. Tiong et al. (2023) found that a majority of studies concentrate on predicting delays for the next station, which has ”limited applicability in practice” (Tiong et al., 2023, p. 4). Although short-term prediction accuracy is typically high, railway operators often face challenges in taking full advantage of this accuracy due to the limited time available for implementing effective traffic control actions. On the other hand, long-term predictions are less accurate but could potentially be much more useful (Nabian et al., 2019). A way to achieve this goal is to consider the prediction not as an (inaccurate) point-estimate but as a probability distribution of confidence that some outcomes are most likely. Consequently, it is of particular interest how the quality of the confidence of the delay prediction dynamically evolves for longer prediction horizons which fits in the concept of predictability, formally defined later in this paper.

To address these shortcomings, we

1. leverage an uncertainty-aware neural network extended with test-time-dropout and loss attenuation to predict train delays and also estimate the level of associated confidence.
2. propose a dynamic prediction horizon framework (DPHF) to quantify and study the dynamic predictability of train delay.

In accordance with existing literature, we use the terms uncertainty quantification or confidence estimation to describe the *a-priori* (i.e., prospective) assessment of the confidence or “sureness” of a prediction. Single-value predictions do not provide estimations for their associated confidence and can only be evaluated *a-posteriori* (i.e., retrospective), in terms of prediction accuracy. The proposed neural network outputs a best-estimate single value for the targeted delay and, additionally, a single value confidence estimation for the associated uncertainty of the prediction. Together, the output can be interpreted as a prediction interval (PI) or as a parametric probability distribution under certain assumptions. Therefore, the size of the PI (PI width) serves prospectively as an interpretable level of confidence, and the resulting prediction is effectively uncertainty-aware. We suggest assessing the quality of the uncertainty-aware predictions *a-posteriori* via the PI coverage (i.e., the fraction of realizations covered by the predicted interval), and the Likeliness of Realization (LoR) within the predicted probability distribution.

Furthermore, we study the evolution of the prediction quality over time with a dynamic prediction horizon framework (DPHF). While some prediction inputs are known well in advance of the event (offline), others, like the current delay, emerge during the train’s journey (online). As shown in Fig. 1 top horizontal axis, predictions made further ahead of an event are inherently more valuable, as they allow operators greater flexibility in response actions at strategical and operational levels, such as potential reroutings or changes of the timetable. At the same time, long-term predictions tend to be less accurate. However, there is a lack of understanding of how the quality of the prediction of train delay evolves from long-term to short-term (symbolized by the question marks in the blue box of Fig. 1). Assuming that the long-term prediction accuracy cannot be increased significantly due to the inherent uncertainty of the process and measurements, we propose to quantify the associated uncertainty dynamically (i.e., over time) and aim to understand how the performance of the uncertainty-aware prediction evolves from long-term to short-term predictions. While the accuracy of a prediction model is a reasonable measure for a promised prediction quality, we are interested in the changes in the uncertainty bounds over time arising from the algorithm itself. Therefore, we study the effects of influencing factors on the temporal dynamic prediction uncertainty. These contributions provide valuable insights for traffic controllers basing their traffic control actions on predictions and also for passengers deciding on actions during and after their trip.

Our results demonstrate that the proposed uncertainty-aware neural network outperforms commonly-used inherently stochastic models in predicting train delays, achieving superiority not only in terms of prediction accuracy, but also in the quantified associated confidence. Furthermore, we discover within the DPHF that the dynamic predictability of train delay decreases *exponentially* as the prediction horizon increases. This finding aligns with a recent similar study on bus delay predictability conducted by Büchel and Corman (2022).

Our study further uncovers that the prediction of train delays inherently possesses a significant degree of aleatoric uncertainty, highlighting the intrinsic randomness associated with the phenomenon. We also find that the predictability of the final delay of a train in our empirical study significantly depends on the train's offline characteristics, in particular its category (regional or inter-city train).

The remainder of this study is structured as follows. Section 2 provides a review of stochastic train delay prediction approaches, uncertainty-aware machine learning prediction methods and predictability analytics within the field of transportation. In Section 3, we describe the proposed uncertainty-aware neural network and the commonly-used benchmark prediction models. Furthermore, we clarify how to use the theory of entropy to estimate the aleatoric and epistemic prediction uncertainty. Section 4 reports the results of our case study of the DPHF within the Swiss railway network, and Section 5 discusses the limitations of our models as well as data before we conclude in Section 6.

2. Literature review

Train delay prediction has been an active area of research for some decades already (Wen et al., 2020). Therefore, a considerable body of literature on predicting train delays exists with a variety of applied models and used input data. A general overview of train delay predictions is provided by Spanninger et al. (2022), who find a trend towards data-driven approaches, relying heavily on historical train operation data. Those approaches have been reviewed more in detail by Tiong et al. (2023), who notice that prediction approaches based on neural networks are most commonly used on the data-driven side and that a majority of data-driven train delay prediction approaches focus on short-term prediction accuracy.

Early on, Peters et al. (2005) apply a classic feedforward neural network model with 63 neurons to predict train delays and already conclude that complex neural network models are highly data-hungry, meaning that a lot of training data is necessary to achieve a good prediction performance. Furthermore, Yaghini et al. (2013) points out that a great challenge of applying neural networks to predict train delays is to design a network architecture. They investigate three methods of setting up a neural network structure. A quick method, where a simple structure is assumed straight from the beginning; a dynamic method, where the amount of neurons is increased during training as long as it enhances the prediction quality; and the multiple method, where multiple neural network structures are trained in parallel — without concluding that any of the three is clearly the best. Liu et al. (2017) deal with large dwell time fluctuations and show that in their case study of real-time train delay predictions that the back-propagation neural network approach with the sigmoid activation function, optimized by genetic algorithm, outperforms the wavelet neural network with the Morlet wavelet activation function. Within the perspective of big data analytics, Oneto et al. (2018) present a deep extreme learning neural network approach with multiple hidden layers to improve the feature learning. They develop a large-scale train delay prediction system taking into account the current day of the week, whether it is a holiday or a working day, and the past train delays of trains running within the system. Wen et al. (2020) present that a higher prediction accuracy can be achieved by long short-term memory neural networks, which inherit "memory cells" within their structure to maintain information in memory for long periods. This is especially useful for sequential data, like time series, but has its drawback in terms of being more complex and computationally intense compared to traditional and deep neural networks. Particle swarm optimization is proposed by Bao et al. (2021) to optimize the network architecture of an extreme learning machine neural network to predict train delays. Also recently, Huang et al. (2020) presented a combined approach of a fully-connected neural network and two long short-term memory components to separately process operational and non-operational features. They show that their proposed hybrid network architecture outperforms classical neural networks.

All train delay prediction approaches based on neural networks aim to achieve the highest prediction accuracy (i.e., the lowest prediction error), measured with the mean absolute error (MAE) or root mean squared error (RMSE). None of the above-summarized approaches quantifies the uncertainty associated with the generated predictions. Furthermore, only Oneto et al. (2018) present a temporal dynamic prediction quality analysis for different prediction horizons.

Railway operations have a considerable amount of inherent variability. Therefore, stochastic models, designed to exploit this variability, are gaining popularity (Artan and Şahin, 2022). Early studies of stochastic event-based approaches (Spanninger et al., 2022) employed general graph models (e.g., Carey and Kwieciński, 1994; Hallowell and Harker, 1996; Yuan and Hansen, 2007). Carey and Kwieciński (1994) determine a simple formula to approximate the stochastic effect of headways to trip link times. Yuan and Hansen (2007) propose a detailed stochastic model to predict the delay propagation of trains in stations, explicitly capturing route conflicts and late transfer connections. Berger et al. (2011) propose a stochastic model to predict train arrival and departure delays including waiting policies, driving time profiles, and catch-up potential. Büker and Seybold (2012) present a large-scale activity graph model to analytically describe the propagation of train delay on a network level. Their approach is based on an expansion of exponential polynomials as flexible (i.e., easy to convolute) distribution functions to avoid heuristics like the commonly used Monte Carlo sampling. In a similar stream of research, Meester and Muns (2007) propose the usage of phase-type approximation for the stochastic analysis of delay propagation in large railway networks.

Keyhani et al. (2012) use estimated probability distributions for train arrival and departure delays to predict the feasibility of connections. Also Lemnian et al. (2014) predict the feasibility (categorically: safe, uncertain, critical, break) of connections based on a stochastic event-activity graph model and evaluate the prediction error rate for up to three hours ahead of the planned connection.

Markov chain models have gained prominence in recent years for predicting train delays, with various implementations and extensions to enhance the prediction performance (Şahin, 2017; Barta et al., 2012; Kecman et al., 2015; Gaurav and Srivastava, 2018; Spanninger et al., 2021; Büchel et al., 2021; Şahin, 2022; Artan and Şahin, 2022, 2023). In essence, markov chain approaches are based on the interpretation of the evolution of train delay along a journey as a stochastic process and the assumption that the probability distribution of the delay at the next stop only depends on the most recent delay. Delay is typically classified into states and in this way, transition probability matrices can be calibrated based on historical train movement data and used to predict train delays. Artan and Şahin (2023) propose a relaxation of the commonly-used homogeneity assumption (i.e., the delay transition probability matrix is the same for all parts of the railway network) and Spanninger et al. (2023) present further flexibilizations for the state space definition to enhance the prediction quality. However, markov chain approaches take into account inter-train effects only implicitly in the form of transition probabilities. Bayesian networks can be seen as extensions of markov chain approaches in this sense and have been proposed recently by Corman and Kecman (2018), Lessan et al. (2019) and Huang et al. (2022). The allowance of inter-train dependencies, however, has two drawbacks. First, a dependency structure must be defined for the prediction horizon — which is a prediction of the sequence of events itself. This problem has been solved by assuming the timetable holds for the future by Corman and Kecman (2018) and by data-driven machine learning techniques by Lessan et al. (2019). The latter study, however, proposes to use only the two most recent delays of the same train for the prediction model. The second drawback of allowing inter-train dependencies is a more complex theory of convoluting probability distributions for future events. So far, the proposed approaches are based on Gaussian assumptions for the delay at events and linear dependencies in between. Huang et al. (2022) propose a dependency structure for the bayesian network based on a context-driven delay evolution clustering assuming delays of multiple trains as parents. They show an improvement of the stochastic prediction quality by evaluating the cumulative probability distribution function of the predictive errors.

Despite the existence of numerous stochastic approaches for train delay prediction, only a few studies investigate the uncertainty associated with their generated predictions, especially in the context of a temporal dynamic perspective. Berger et al. (2011) evaluate the change of the predicted probability distribution over time, only exemplary for a single train journey, however. Similar to deterministic approaches, stochastic approaches are also typically evaluated by the mean absolute error (MAE) and root mean squared error (RMSE). Wang et al. (2022) evaluate the model accuracy of train delay predictions obtained by a proposed support vector machine model for different values of allowed errors, which can be interpreted as model confidence bounds. Keyhani et al. (2012) introduce a reliability rating and thereby stochastically quantify the feasibility of a train connection with a discrete probability distribution. Corman and Kecman (2018) present the evolution of the predicted delay distributions for a decreasing prediction horizon and evaluate the standard deviation of delay predictions for more than five minutes. However, they also do not discuss whether or not this predicted uncertainty is good or bad (i.e., close or far from reality) in any way.

While, to the best of our knowledge, there has not been a systematic investigation of the predictability of train delays *over time*, such analysis is available in related fields. Bus arrival time prediction was already formulated as a dynamic problem in 2002 by Chien et al. (2002). Büchel and Corman (2022) describe an exponential decay of predictability with the horizon, following Cats and Loutos (2016) who analyse the effect on waiting-time predictions. In contrast, Sun et al. (2007) suggest a rather linear relation between the time horizon and the bus arrival prediction error. Similar analysis can be found in other applications involving spatio-temporal data; for example, trajectory-based next place prediction in individual human mobility (Tenzer et al., 2022; Tsiligkaridis et al., 2022). In next place prediction, the predictability increases strongly with the first few observations and converges for shorter horizons, as opposed to the exponential decay with *longer* horizon observed for bus delay.

In this work, we aim to study the dynamic predictability of train delays with uncertainty-aware neural networks. Commonly employed NNs, including the models that were applied for train delay prediction, usually provide point estimates (i.e., single values) without any notion of associated uncertainty. Guo et al. (2017) further show that neural networks are, in general, poorly calibrated, where calibration refers to the alignment between prediction confidence and the actual sample accuracy. To tackle this issue, there has been extensive work on uncertainty quantification of NNs. Gawlikowski et al. (2021) survey uncertainty quantification methods for deep networks and define four groups, namely predicting the uncertainty with deterministic neural networks, employing bayesian neural networks, ensemble approaches and test-time data augmentation. While some methods, such as data augmentation, are mainly suitable for image data, ensemble methods (Lakshminarayanan et al., 2017) or test-time dropout (Gal and Ghahramani, 2016) can be applied for any task. In these approaches, the variance over the outputs of each forward pass (per model or with random dropout of neurons) is considered as the prediction uncertainty (Lakshminarayanan et al., 2017; Mentch and Hooker, 2016), and was shown to approximate bayesian marginalization (Wilson and Izmailov, 2020). Mazloumi et al. (2011), for example, estimate prediction intervals for bus travel time prediction using model ensembles.

Putting the above considerations together, we see the following shortcomings in the existing literature on train delay predictions: Stochastic models, despite their potential, are often underutilized in assessing the confidence levels within train delay predictions. Additionally, commonly-used approaches employing machine learning techniques mainly focus on maximizing the accuracy of their single value prediction without adequately quantifying the associated (un-)certainty. Furthermore, the analysis of predictability has only seen little attention, and the quantification of its dynamic evolution for train delays has remained unexplored so far.

3. Methodology

3.1. Problem definition

In this study, we aim to predict the arrival delay of trains. Therefore, we interpret the arrival delay as a continuous random variable Y and utilize independent variables summarized in the vector $\mathbf{X}$ as input. Consequently, the prediction task corresponds to

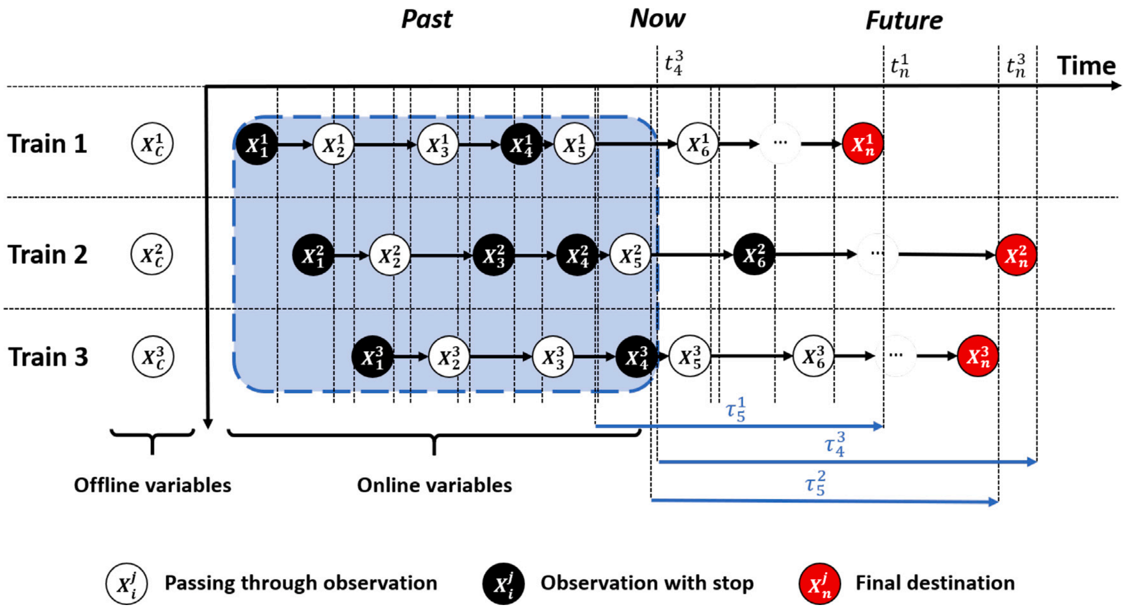

Past
Now
Future
Time
t_4^3
t_n^1
t_n^3
Train 1
Train 2
Train 3
X_C^1
X_C^2
X_C^3
X_1^1 X_2^1 X_3^1 X_4^1 X_5^1 X_6^1 ... X_n^1
X_1^2 X_2^2 X_3^2 X_4^2 X_5^2 X_6^2 ... X_n^2
X_1^3 X_2^3 X_3^3 X_4^3 X_5^3 X_6^3 ... X_n^3
τ_5^1
τ_4^3
τ_5^2
Offline variables
Online variables
X_i^j Passing through observation
X_i^j Observation with stop
X_n^j Final destination

Fig. 2. Visualization of the dynamic prediction horizon framework (DPHF).

defining a model $\mathcal{M}(\Theta) : \mathbf{X} \to Y$ based on the set of parameters Θ that estimates the conditional probability distribution $P(Y|\mathbf{X})$, which identifies the unknown true mapping $C : \mathbf{X} \to Y$.

We use a set of historical observations $D_N = \{(\mathbf{x_1}, y_1), \dots, (\mathbf{x_N}, y_N)\}$ to learn $\mathcal{M}(\Theta)$. The input space $\mathbf{X}$ in our study consists of the

- train ID
- train position
- current delay
- remaining distance to the final destination
- remaining planned running time
- day of the week
- time of the day
- train journey characteristics: direction and category.

While the day of the week, direction and train category are offline variables, known in advance, we refer to the other input variables as online, emphasizing their availability only in real-time. For implementation details, we refer to our source code available at https://github.com/mie-lab/train_delay.

3.2. Dynamic prediction horizon framework (DPHF)

To study the evolution of the dynamic predictability of train delay, we construct a dynamic prediction horizon framework (DPHF). Therefore, we partition the journey of train j into intervals, separated by operational points (where event time is measured and can be compared to the plan, see Section 4). When train j passes the operational point k at time t_k^j, an update of the current delay of train j becomes available. Based on this new online information at time t_k^j for $k = 1, \dots, n$, we predict the arrival delay of train j at a specific station. The concept of the DPHF is visualized in Fig. 2, which is pictured at the moment t_4^3. The x-axis shows the evolution of time, while the y-axis represents different trains. Before starting their journey, the offline information summarized in X_C^j is already available according to the schedule. During the journey, train j passes operational points and online information becomes available. Operation points can be with and without scheduled stops. At scheduled stops, trains have to perform onboarding/alighting of passengers and should depart after the planned departure time. We define τ_k^j as the prediction horizon of the delay prediction of train j at time t_k^j (i.e., the time between the prediction is made and the final arrival time, visualized in Fig. 2 in blue).

In Fig. 2, train 3 has just passed its fourth operational point, and therefore, its final delay prediction is updated (with prediction horizon denoted as τ_4^3). At the final operational point n of train j, passed at time t_n^j, the current delay equals the targeted final arrival delay.

3.3. Uncertainty-aware neural network

We, for the first time, propose to extend a neural network to an uncertainty-aware prediction approach for train delays. This choice is motivated by (1) previous literature reporting successful use of NNs for this problem (Shi et al., 2020; Kecman and Goverde, 2015; Malavasi and Ricci, 2001; Oneto et al., 2018; Boateng and Yang, 2023), (2) the potential to capture aleatoric and epistemic uncertainties, as explained in the following, and (3) our empirical findings of NNs outperforming stochastic machine learning models. We will first give a brief introduction to neural networks and then explain the uncertainty quantification extension.

3.3.1. Network architecture

Neural networks have revolutionized computer science in recent years, due to their power to approximate arbitrarily complex functions. A simple neural network consists of l layers of a certain number of neurons that are *fully connected*, i.e., each neuron receives a weighted sum of the outputs of all neurons in the previous layer. Let y_i^l be the output of the ith neuron of the lth layer. It is computed as $y_i^l = g(\sum_j w_{ij}^l y_k^{l-1} + b_i)$, where W^l is the adjustable weight matrix in layer l and b is the bias vector. Non-linearity is added via an activation function g applied to each neuron. W and b are trained via stochastic gradient descent; i.e., the weights are adjusted by backpropagating the gradient of a loss function that measures the error of the outputs of the final layer with respect to the ground truth label. For further explanation, we refer to LeCun et al. (2015).

In this study, we implement a simple fully-connected network with two hidden layers (128 neurons each), with ReLU (Agarap, 2018) activation in both layers and a Sigmoid activation in the output layer (one neuron). In both layers, Dropout (Srivastava et al., 2014) with a keep probability of 0.5 is applied to prevent overfitting and to enable uncertainty estimation (see following section). The Sigmoid outputs ($\in (0, 1)$) are then scaled to yield predictions in form of seconds of delay. The model is implemented in Pytorch (Paszke et al., 2019), with a batch size of 8, using AdamOptimizer and a learning rate of 10^{-5}. The model is trained on the complete training data, instead of training one model per train, since the latter shows inferior predictive performance (see Appendix H). A sensitivity analysis of the performance with respect to the network hyperparameters (number of layers, number of neurons, and learning rate) can be found in Appendix G.

3.3.2. Uncertainty quantification

The proposed technique of uncertainty-aware machine learning enables the division of uncertainty into two components: aleatoric and epistemic uncertainties. These correspond respectively to uncertainties arising from inherent data variability and from imperfection in the model (Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017). The differentiation between aleatoric and epistemic uncertainty can be theoretically motivated from an information-theory perspective, through a decomposition of the Shannon entropy. The entropy H of the outputs Y is reformulated as a sum of conditional entropy $H(Y|\theta)$ and mutual information $I(Y, \theta)$ (Ash, 1965):

$$H(Y) = H(Y|\theta) + I(Y, \theta) \tag{1}$$

where $\theta \sim \Theta$ is a random variable and here denotes the parameters of the model. Since $H(Y|\theta)$ is the conditional entropy of the outputs given the model, it is considered to represent the aleatoric uncertainty. The second part can be rewritten in terms of Kullback–Leibler-divergence D_{KL}:

$$I(Y, \theta) = \mathbb{E}_{\Theta}\left[D_{KL}\left(p(Y|\Theta) \parallel p(Y)\right)\right]$$

It describes the epistemic uncertainty, as it corresponds to the divergence that is due to a specific model (defined by Θ). The epistemic uncertainty can be reduced with an appropriate choice of Θ, or, in other words, with more expressive or more suitable models.

These two components are, however, intractable to compute, due to the integral in the entropy computation ($H(Y) = -\int_Y p(y) \ln p(y) dy$). Various methods have been proposed to approximate the epistemic and aleatoric uncertainty, including bayesian approaches and ensemble methods.

In this study, we follow Kendall and Gal (2017) to estimate epistemic and aleatoric uncertainties via test-time dropout (Gal and Ghahramani, 2016) and loss attenuation respectively. Dropout refers to randomly deactivating a subset of neurons during each forward pass, and was originally proposed as a regularization method for model training (Srivastava et al., 2014). Activating dropout at test time imposes randomness in the outputs (due to the random choice of deactivated neurons), and was shown to resemble a bayesian approximation (Gal and Ghahramani, 2016). Specifically, Gal and Ghahramani (2016) show that "the dropout objective minimizes the Kullback–Leibler divergence between an approximate distribution and the posterior of a deep Gaussian process" (Gal and Ghahramani, 2016). Intuitively, random dropout of neurons induces a probability distribution over models and can thus recover the posterior via Monte-Carlo sampling, i.e., with several stochastic forward passes.

To estimate the aleatoric uncertainty, on the other hand, we apply loss attenuation. In loss attenuation, the network is trained to output the uncertainty directly within the prediction step via a tailored loss function. As proposed by Nix and Weigend (1994) we use a negative log-likelihood loss function. Let θ denote the parameters of the neural network, let x be the input sample (i.e., current delay, train identification information and context features) and y the ground truth output (i.e., here the final train delay), then

$$\mathcal{L}_{NLL}(\theta) = \mathbb{E}_{x,y}\left[\frac{1}{2}\log \hat{\sigma}^2(x) + \frac{(y - \hat{y}(x))^2}{2\hat{\sigma}^2(x)}\right] \tag{2}$$

where $\hat{y}$ and $\hat{\sigma}$ are the outputs from the last layer and estimate the expected value of the realization and the associated uncertainty respectively, assuming that the errors are approximately normally distributed (see Appendix A).

The total estimated uncertainty for one sample is obtained by summing the aleatoric uncertainty (i.e., $\hat{\sigma}$) with the epistemic uncertainty estimated via test-time-dropout.

3.4. Benchmark models

We compare the prediction accuracy and quality of uncertainty estimation from the proposed uncertainty-aware neural network model to six commonly-used machine learning and stochastic models. Stochastic models are based on the assumption that all process times within the dynamics of railway traffic follow underlying but unknown probability distributions. By using historical data, these models estimate parameters for assumed underlying probability distributions to generate predictions for train delays. In the context of train delay prediction, stochastic approaches often employ event-activity graphs, which provide a visual representation of arrival, pass-through, and departure events, as well as alternating running and dwelling processes for multiple trains (Spanninger et al., 2022). Recent advancements in stochastic modelling techniques include non-stationary markov chain models and bayesian networks with hybrid dependency structures.

3.4.1. Markov chain (MC) models

In the context of markov chain models, the delay evolution of a train along its journey is interpreted as a discrete stochastic process. These models assume that the probability of a train being delayed at a given point in time only depends on its delay status at the previous point in time. This assumption is known as the markov property. Markov chain models offer a powerful descriptive capability by capturing the transitions in-between n delay states through a transition probability matrix (TPM) of dimension n^2.

A variety of markov chain based-approaches have been proposed for predicting train delays. Within this study, we implement and assess the latest MC contributions to the field (Artan and Şahin, 2022; Şahin, 2022; Spanninger et al., 2023), namely

- Markov chain multi-step (*MC-multi*)
  Building upon the proposed inhomogeneous markov chain model by Artan and Şahin (2022), we partition a train journey into a series of consecutive processes. For each process, we estimate a transition probability matrix (TPM) using historical delay observations and generate predictions by multiplying these TPMs.
- Markov chain 2-step event-based (*MC-2step-E*)
  For the MC-2step-E approach, we partition the train journey into two parts: From the beginning to the current location of the train and from this point to the train's final destination. We then similarly calibrate TPMs for delay state transitions in-between the initial departure, the intermediate location and the final destination. In this way, we consolidate the aggregated variability along those sections within the TPMs and can generate predictions for the final arrival delay.
- Markov chain 2-step process-based (*MC-2step-P*)
  Similarly to the previous approach, the train journey is partitioned into two parts. Following the idea in Spanninger et al. (2023), we, however, use process time deviation instead of the absolute delay as the underlying variable of the stochastic process.

For all markov chain models, we use a state space Ω of dimension $n = 20$ and define the state boundaries according to the respective quantiles of the empirical delay distributions (or process time distributions in context of the MC-2step-P model).

3.4.2. Bayesian networks (BN)

Bayesian networks are another commonly-used model to generate stochastic predictions for train delays. They represent a system of variables and their conditional dependencies using a directed acyclic graph (DAG). For details on BN theory, we refer to Scutari and Denis (2021). Following the applications of Corman and Kecman (2018), Lessan et al. (2019) and Huang et al. (2022), we compare the predictions of our proposed uncertainty-aware neural network approach to a bayesian network assuming Gaussian distributions for all variables and linear dependencies with a hybrid dependency structure following Lessan et al. (2019) and divide train journeys into intervals of 3 minutes for the implementation, similarly to Huang et al. (2022).

3.4.3. Boosted-trees as probabilistic predictors

Gradient boosting decision tree methods such as random forests were repeatedly shown to outperform neural networks on simple and tabular data (Grinsztajn et al., 2022). Since random forests in its original form do not provide uncertainty estimates, we employ the NGBoost (Duan et al., 2020) approach that was proposed as a probabilistic boosting method, predicting the parameters of a probability distribution to derive the expected value and uncertainty. In this context, we experiment with the normal and lognormal distribution to model the error distribution. The same input features as for the NN are used.

3.5. Evaluation methods

The nature of uncertainty or confidence of predictions in our study are expressed by (1) a prediction interval (potentially constructed from single-value prediction and uncertainty estimate), (2) a discrete probability distribution or (3) a continuous probability distribution, as visualized in Fig. 3. We propose to compare models based on their accuracy (MAE and RMSE) and likeliness of realization (LoR), and to analyse the uncertainties via the conformal prediction framework, i.e., prediction intervals (PIs).

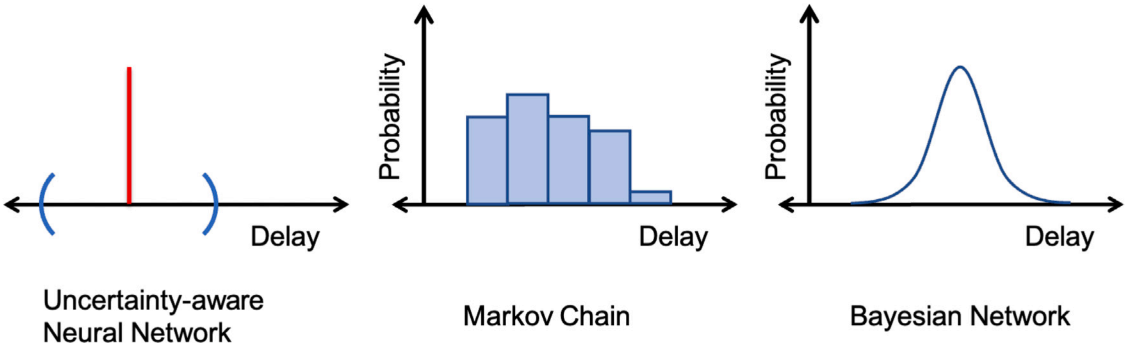


Fig. 3. Visualization of prediction outputs for uncertainty-aware neural networks, markov chain models and bayesian network models.

3.5.1. Accuracy

Following Taylor and Thompson (1982), accuracy measures the proximity of a set of single-value predictions to the actual realized values. The mean absolute error (MAE) and the root mean squared error (RMSE) serve as the most commonly used performance indicators in the context of train delay prediction. This is primarily due to the majority of contributions focusing on single-value predictions. However, for stochastic predictions, the direct application of MAE and RMSE necessitates the reduction of the predicted probability distributions to single-value representatives. Unfortunately, this process results in the loss of the inherent richness and depth of probabilistic predictions, which we argue offers a more comprehensive understanding of the future compared to single-value predictions.

3.5.2. Likeliness of realization (LoR)

To overcome these limitations, we introduce the likeliness of realization (LoR) as a meaningful measure to assess the quality of an uncertainty-aware prediction for a given prediction horizon. Unlike mean absolute error (MAE) and root mean squared error (RMSE), which are tailored to single-value predictions, the LoR allows for the evaluation of stochastic predictions, capturing the degree to which predicted probability distributions align with actual outcomes. This metric offers a more nuanced understanding of prediction quality, preserving the inherent richness of probabilistic predictions and connecting them to the real-world phenomena they aim to represent

The LoR is designed to measure the concentration of the predicted probability distribution near the realized value of delay y_i, observed a-posteriori. To do so, we use a range Δ, which forms the basis of a symmetric interval. This interval, $\mathbb{I}_\Delta(y_i) = [y_i - \Delta, y_i + \Delta]$, is centred around the realized delay observation y_i. The LoR$_\Delta$ for multiple predictions $i = 1, \dots, N$ is then formally defined as the average of the integrals of the predicted delay distributions $f_i(x)$:

$$LoR_\Delta = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathbb{I}_\Delta(y_i)} f_i(x) \quad dx \tag{3}$$

The LoR$_\Delta$ is a probability taking values from 0% to 100%. It is important to note that the LoR does not provide additional insights for deterministic predictions.[2]

3.5.3. Analysing uncertainty via prediction intervals

Following the conformal prediction framework (Shafer and Vovk, 2008; Vovk et al., 2005; Angelopoulos and Bates, 2021), we evaluate the uncertainty-enhanced predictions by the width of their associated *prediction interval* (PI). Conformal prediction allows transforming a heuristic uncertainty (per sample) into a rigorous and interpretable uncertainty in the form of a PI (Angelopoulos and Bates, 2021). This involves scaling the uncertainty estimates of a model by a factor determined on a validation dataset to achieve a certain coverage rate (or realizations being within the boundaries of the PI), denoted as ϕ. For stochastic models that produce probability distribution outputs, we define the prediction interval (PI) as the range around the expected value that encompasses $\phi \cdot 100\%$ of the probability mass. The PI width can be seen as a measure for *precision*. Narrower PIs indicate higher precision, reflecting less uncertainty, while wider PIs signify lower precision and more uncertainty. Let $\hat{y}_i$ be the predicted final delay for the ith sample ($i \in [1..n]$), and let $\hat{\sigma}_i$ be the estimated uncertainty. The PI is defined as a symmetric interval $\mathcal{I}_i$ around $\hat{y}_i$ that is constructed such that a specific fraction $\phi \in [0, 1]$ of the ground truth values y_i lie within the interval bounds; formally $\mathbb{P}(y_i \in \mathcal{I}_i) \geq \phi$ for some user-defined ϕ. To satisfy this so-called *coverage*, the uncertainty is scaled by a factor $\hat{q}$ which is calibrated on a validation dataset. Vovk et al. (2005) show that setting $\hat{q}$ to the $\frac{\lceil \phi(n+1) \rceil}{n}$ quantile of the calibration scores guarantees that the interval $\mathcal{I}_i = [\hat{y}_i - \hat{\sigma}_i \hat{q}, \hat{y}_i + \hat{\sigma}_i \hat{q}]$ achieves the required coverage of ϕ on the calibration data (Angelopoulos and Bates, 2021). If the test data is sufficiently similar, the *empirical coverage* achieved on the test data should be close to ϕ as well.

[2] A deterministic prediction can be interpreted as a Dirac distribution. Consequently, the LoR would either be 100% or 0% depending on whether the predicted value falls within the realization surrounding interval.

Based on this framework, we can evaluate the precision of uncertainty-aware predictions by measuring the mean PI width at a fixed coverage ϕ, denoted as $W(\phi)$. The mean PI width is the average size of the interval $\mathcal{I}_i$, corresponding to

$$\mathcal{W}(\phi) = \frac{1}{n}\sum_{i=1}^{n} 2\hat{\sigma}_i(x)\hat{q}\ .$$

3.6. Predictability of train delay

Understanding the inherent predictability of the phenomenon *train delay* is of immense practical relevance for railway operators and passengers. Rather than emphasizing a model's predictive efficacy, the predictability of a phenomenon refers to the inherent characteristics or structure within the phenomenon that make it possible to anticipate or predict. For instance, if train delays consistently occur at the same time every day due to scheduled maintenance, the phenomenon features high predictability.

We propose to approximate the intrinsic predictability of train delays by analysing the LoR of the (empirically) best prediction model, assuming that it captures discernible patterns and abstracts from inherent randomness in its uncertainty estimation. Specifically, we characterize the dynamic evolution of the LoR across increasing prediction horizons, denoted as τ. The LoR encapsulates the quality of the predictions with estimated levels of confidence, offering a comprehensive view that combines both point estimates and their associated uncertainties.

The proposed DPHF is designed to study the predictability of train delay. To simplify the quantification of predictability, we assume the LoR decays (i.e., reduces over time) proportionally to its value, as done in a similar delay analysis of road-based bus systems (Büchel and Corman, 2022). This corresponds to an exponential decay and allows quantifying the approximation of the unknown predictability with only three parameters (shift parameters α and γ for the short- and long-term predictability and a decay parameter λ).

Let $Q(\tau)$ be the LoR for the generated prediction by our proposed uncertainty-aware neural network predictions for prediction horizon τ. Exponential decay is defined by a functional form:

$$\frac{dQ(\tau)}{d\tau} = -\lambda Q(\tau) \tag{4}$$

Rearranging and integrating Eq. (4) leads to

$$\begin{aligned}\frac{dQ(\tau)}{Q(\tau)} &= -\lambda \cdot d\tau\\ \ln(Q(\tau)) &= -\lambda\tau + c\\ Q(\tau) &= e^c \cdot e^{-\lambda\tau}\end{aligned} \tag{5}$$

However, unlike in standard exponential decay (Eq. (5)), in our case the extreme value of the LoR for $\tau \to \infty$ is not 0. Instead, the LoR converges to a stable long-term prediction level γ, corresponding to the LoR of a prediction solely based on historical data; without considering real-time information. Furthermore, we rewrite $e^c := \alpha$ which leads to the following parameterization of predictability $\mathcal{P}(\tau)$:

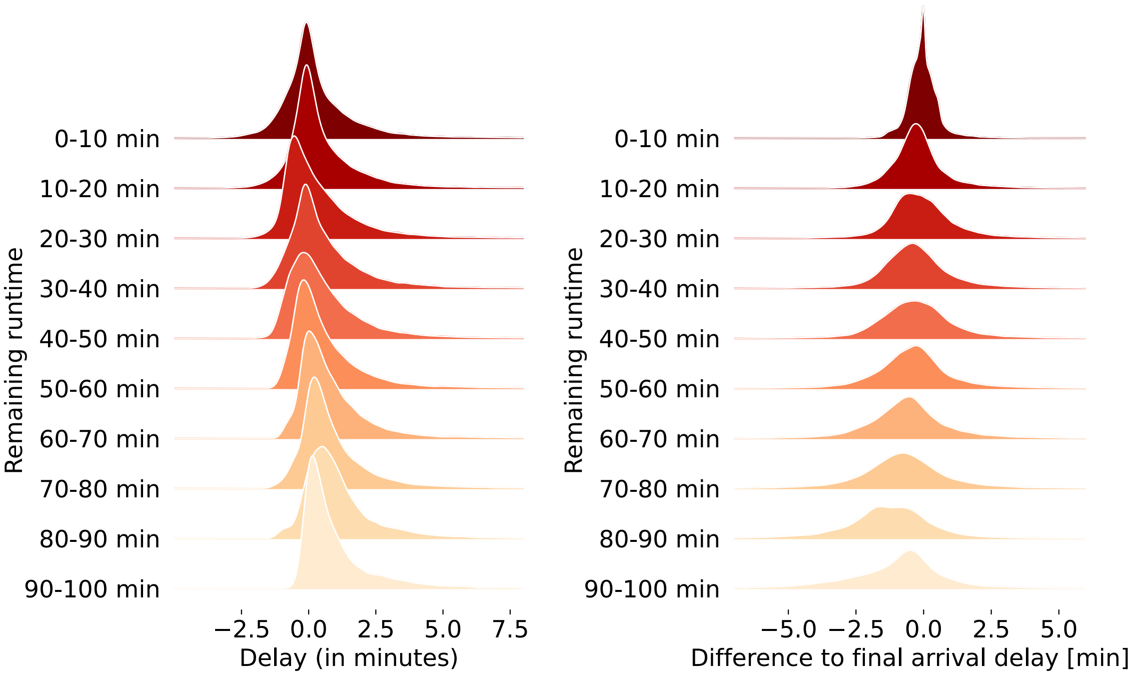
$$\mathcal{P}(\tau) = \alpha \cdot e^{-\lambda\tau} + \gamma \tag{6}$$

The fitted value of γ represents the long-term predictability convergence level $\lim_{\tau\to\infty}\mathcal{P}(\tau)$. It can be interpreted as the inherent uncertainty in the phenomenon of train delay without any online information. In other words, the level of predictability γ can be achieved without real-time information already before the train has started its journey. In contrast, α describes the gain of prediction quality in the very short-term, on top of γ, since $\lim_{\tau\longrightarrow 0}\mathcal{P}(\tau) = \alpha \cdot 1 + \gamma$. Therefore, $\alpha + \gamma$ can be interpreted as the limit of best possible predictability shortly before the target event takes place, taking into account online information. Finally, the decay parameter λ describes how fast the prediction quality decreases with increasing prediction horizon (i.e., how fast the phenomenon-inherent randomness suppresses the value of real-time information). A large gamma means that the phenomenon becomes difficult to predict fast already for small prediction horizons τ. In this sense, the fitted value of λ provides insights into the dynamic value of real-time information for train delay prediction.

4. Results

4.1. Data and experimental setup

We employ the proposed models to predict the delay of trains running on the busy corridor Zurich-Chur within the Swiss railway network. Spanning approximately 100 kilometres in length and predominantly composed of double-track sections, this corridor hosts about 80 passenger train services scheduled per day in both directions (down: Zurich-Chur, up: Chur-Zurich). For the sake of simplicity, our main experiment studies the predictability of the *final delay* in Zurich or Chur; however, our framework is generally applicable for predicting the delay at any intermediate stations or observational points (see Appendix J). These services typically adhere to one of two established stopping patterns, either making two stops or nine stops along the way. This dichotomy in stopping patterns provides us with an ideal opportunity to investigate and assess the level of confidence associated with the respective dynamic predictions of train delay.



(a) Evolution of the distribution of the current train delay along the journey

(b) Evolution of the difference between the current and the final delay over time

Fig. 4. Evolution of the current delay and the difference of the current delay to the final delay along the trains' journey.

We use an extensive dataset of historical train movement information provided by the Swiss Federal Railways (SBB). This rich database encompasses planned and realized passing times at various signals and other significant operational points (ops). During a typical journey along the Zurich-Chur corridor, a train traverses 130 to 140 such ops (depending on the exact track routing), enabling the computation of delay as a simple deviation between planned and actual passing times.

We utilize data of historical train observations from January to June 2019 for all passenger trains operating along the entire corridor Zurich-Chur, encompassing both directions. For trains travelling direction up, the final delay refers to the arrival delay at Zurich main station, while for trains travelling direction down, it corresponds to the final arrival delay in Chur.

In total, the dataset consists of 14,969 train journeys. The overall data quality is exceptionally high, with minimal data cleaning required due to missing timestamps. Only 15 train journey observations (0.09%) were removed due to a significant number of missing values. Our finally used database for testing our proposed model consists of 14,954 train journeys, almost evenly distributed between the two directions, with 7533 journeys in one direction and 7421 journeys in the other.

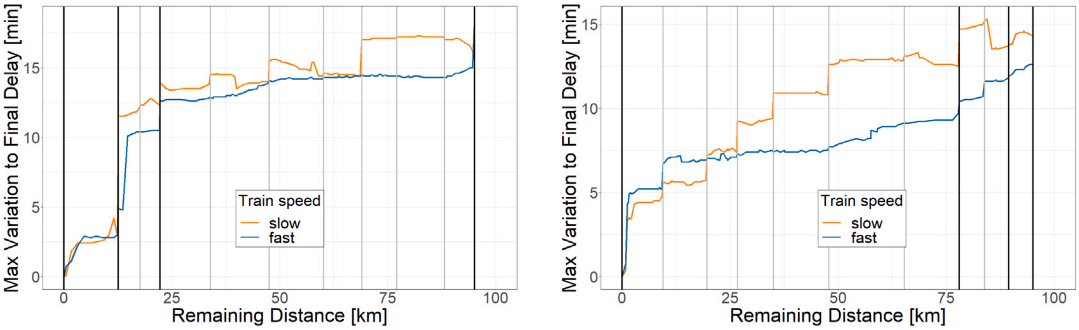
The data is split chronologically into training (75%), validation (10%), and test set (15%). The validation set is used to tune the parameters and to calibrate the PI width. Therefore, we dynamically predict and update the predictions of 2448 train runs of the test dataset for prediction horizons ranging from 6 s to 95 min (0.07 km to 96.24 km), corresponding to a total of 332,168 samples for final arrival delays. All results reported in the following were obtained on the test set.

Typically, the train services between Zurich and Chur are scheduled with two or nine intermediate stops. However, as railway operations are influenced by a lot of abnormal conditions, we can also observe 1042 train journeys (7%) with one, three to eight, or eleven to twelve intermediate stops.

We categorize trains with a planned running time of less than 85 minutes as fast trains and trains with a planned running time of equal or more than 85 minutes as slow trains. This corresponds to fast inter-city trains with typically two stops along the way and slower inter-regional trains with typically nine intermediate stops. In this way, 56% of the observed trains are categorized as fast trains (8377), while 44% of the observed trains (6577) are categorized as slow trains.

4.1.1. *Evolution of the variability of train delays*

As a basis for the evaluation of dynamic predictions in the DPHF, we first investigate the variability of the train delay and its relation to the final delay over time. The evolution of delay along the trains' journeys from Zurich to Chur is visualized in Fig. 4(a). The visualization reveals that the empirical distribution of the final arrival delay (top) shows a higher variance than at initial departure (bottom). In Switzerland, trains with less than three minutes arrival delay are, in general, considered to be punctual.



(a) Evolution of Delay Variation (”Current Delay - Final Delay”) for direction Zurich-Chur.

(b) Evolution of Delay Variation (”Current Delay - Final Delay”) for direction Chur-Zurich.

Fig. 5. Evolution of maximal delay variation (”Current Delay - Final Delay”) for both corridor directions. Black and bold vertical lines indicate stops of fast and slow trains, grey and thin lines additional stops of slow trains.

Since we especially analyse the influence of online real-time information on the prediction of the final arrival delay, we are mostly interested in the spread of the variability between the delay at a specific point/time along the route and the final arrival delay. This information is visualized in Fig. 4(b), where the x-axis represents the difference of the final delay of a train and its current delay at this point in time (final delay–current delay). We can observe that the spread of differences decreases significantly for a shorter remaining running time. The amount and variability of the difference between the current and the final delay is intuitively related to the inherently planned supplement time. Empirical analyses have shown that supplement time in our case study is distributed linearly along the route of the trains, and slow trains have more allocated supplement time than fast trains. The decrease of the spread between the current and the final delay connects with the common benchmark of using the current delay as an intuitive prediction.

4.1.2. Variability of delay with respect to train category and direction

Fig. 5 depicts the maximal variation between current delay and final delay for slow and fast trains in both directions. Black vertical lines indicate the planned stops of fast (and slow) trains, while the grey vertical lines represent the additional stops of slow trains. Notably, the intermediate stops of the fast trains in downwards direction are significantly closer to the terminal station. Consequently, Fig. 5 highlights substantial differences in the maximal variation per direction for corresponding remaining distances.

For trains from Zurich to Chur, the maximal variation remains relatively high at 10–15 minutes from 96 to 20 kilometres remaining distance. However, for the other direction, the maximal variation decreases significantly after the second intermediate stop at approximately 75 kilometres remaining distance. Based on these observations, we anticipate that the number and location of stops substantially influence predictability.

4.2. Prediction accuracy results

To evaluate the prediction accuracy of the proposed uncertainty-aware neural network and benchmark models, Table 1 presents the results in terms of MAE, RMSE, and LoR — averaged for all predictions for different directions, train categories and prediction horizons. For the stochastic models, the expected values of the predicted distributions are used as representative single-value predictions to evaluate the MAE and RMSE. For the sake of comparability, Table 1 includes all samples that can be predicted by all models (80%). The MC-2stepP model, thereby, is the limiting model, as this model needs a minimal number of observations from the initial point of the train journey to the current point of operation (where the prediction is performed) and from this point to the final destination to calibrate transition probability matrices. We report the result evaluation with all possible samples as well as the evaluation removing exceptionally large delays of more than 15 minutes in Appendix K, which shows similar model ranks.

The comparison shows that the neural network performs best in all criteria. Remarkably it also performs best in terms of LoR, despite the advantage of stochastic models estimating an entire probability distribution. We can also note that all models outperform the intuitive benchmark models of the simple median and the current delay significantly. The simple median corresponds to the median of the arrival delays of the same train within the training dataset and the current delay equals the latest online information of the train’s delay at the time of the prediction.

Among the stochastic models (i.e., markov chain and bayesian network), Table 1 illustrates that the MC-2stepE model achieves superior prediction quality in terms of MAE and LoR overall. In contrast, the bayesian network approach demonstrates relatively inferior values for MAE and LoR, implying that the initial assumption of a Gaussian distribution may result in a degradation of prediction quality compared to the markov chain models with less stringent prior assumptions.

To analyse the dependence of the NN performance on the chosen NN architecture and input features, we conduct a sensitivity analysis in Appendix G. Surprisingly, supplying the NN with additional information such as supplement/buffer time or weather data did not improve the results. Apparently, the supplement times can be inferred from the data, or they are not indicative for the final delay by themselves. Similarly, the results are robust with respect to the network architecture, affecting the RMSE or LoR by less than 2%.

Table 1
Comparison of the average prediction accuracy in terms of MAE and RMSE and the average quality of predicted confidence (LoR) for the proposed models.

Model	MAE [s]	% to NN	RMSE [s]	% to NN	$LoR_{\Delta=30s}$ [%]	% to NN
NN	46.1	0.0%	88.28	0.0%	46.92	0.0%
NGB-LN	48.84	+5.9%	92.06	+4.3%	33.17	−29.3%
NGB-N	48.92	+6.1%	91.7	+3.9%	41.55	−11.4%
MC-2stepE	52.59	+14.1%	107.29	+21.5%	40.53	−13.6%
MC-2stepP	57.73	+25.2%	94.74	+7.3%	36.83	−21.5%
MC-multi	57.05	+23.8%	109.8	+24.4%	36.49	−22.2%
Simple median	60.2	+30.6%	121.94	+38.1%	38.22	−18.5%
Current delay	68.46	+48.5%	108.32	+22.7%	28.9	−38.4%

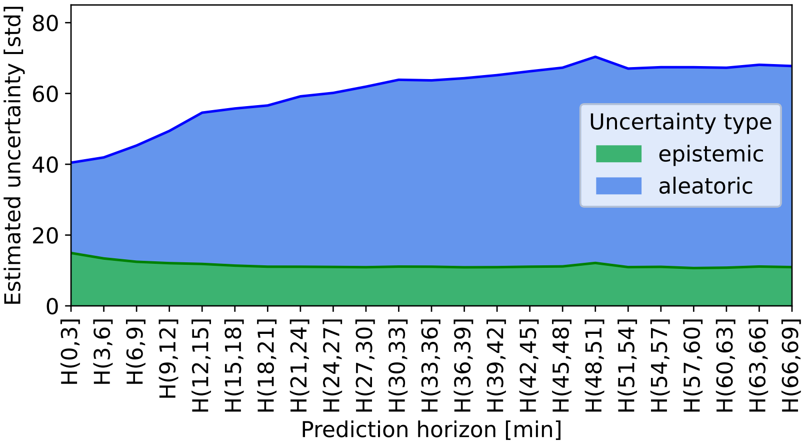


Fig. 6. Comparing aleatoric and epistemic uncertainty over the prediction horizon for the NN model.

4.3. Uncertainty and confidence of predictions in DPHF

The proposed uncertainty-aware neural network approach, combined with results obtained from information theory (see Eq. (1)), enables us to quantify the uncertainty within the generated predictions. This uncertainty is comprised of two components: the estimated standard deviation of a normal distribution (aleatoric component) and the standard deviation of the dropout-modified models (epistemic component). Therefore, the provided uncertainty quantification does not have a directly interpretable unit.

To evaluate the dynamic prediction quality within the DPHF, we categorize the prediction horizons into classes of three minutes. The results, depicted in the stacked plot in Fig. 6, reveal that the aleatoric uncertainty component significantly overshadows the epistemic component. This holds already for the minimal prediction horizon of zero to three minutes, with an evaluated share of 57% aleatoric uncertainty and 43% epistemic uncertainty.

Moreover, the aleatoric component displays a pronounced increase as the prediction horizon extends, whereas the epistemic uncertainty remains almost constant and even decreases. For prediction horizons of up from 60 minutes, we find significantly higher shares of aleatoric uncertainty of 84%. This observation underscores that the primary source of uncertainty stems from the inherently stochastic nature of the data, particularly evident in long-term train delay predictions.

These insights drawn from Fig. 6 indicate that the potential reduction of uncertainty through the adoption of more powerful models is limited. Instead, we posit that the reduction of uncertainty could be achieved through the incorporation of additional information that explains data variance, such as weather conditions, crowding information, or network circumstances. The higher epistemic uncertainty at short-term predictions (< 10 min before arrival) hints at an opportunity to find better methods for this specific use case.

Nonetheless, our efforts to include such supplementary information; specifically, weather data, delay of preceding trains on the same track, and long-term historic delay, did not yield significant improvements in the results (see Appendix G). Evidently, the variability in the data is largely attributable to unexpected events, such as technical issues on the train or human factors which are hard to parameterize.

4.3.1. Dynamic evolution of prediction intervals

As outlined in Section 3.5.3, we analyse precision using the PI width $\mathcal{W}(\phi)$. The PI width serves as an estimated (symmetric) level of confidence for the predictions. When setting the desired coverage to 90% ($\phi = 0.9$), we derive the scaling factor $\hat{q} = 1.381$ for the predicted uncertainty on the validation dataset (see Section 3.5.3). In Appendix A, we show that the distribution of errors of the neural network is distributed symmetrically around the ground truth, justifying the use of symmetric intervals.

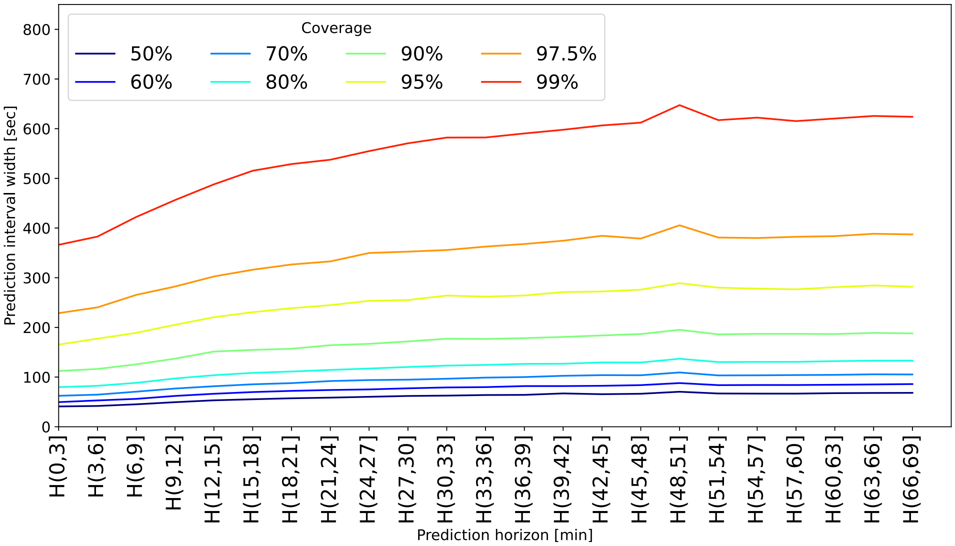


Fig. 7. Prediction interval width by coverage.

The level of confidence varies for different prediction horizons, and rates of desired coverage, as shown in Fig. 7. The results for prediction horizon classes of three minutes show that the mean PI width is increasing with the prediction horizon and the desired coverage. For example, we observe that for a desired coverage of 90%, the PI width is $\mathcal{W}(0.9) = 87$ seconds for predictions up to 0–3 minutes before arrival, but converges to around 180 seconds for prediction horizons beyond 60 minutes (i.e., 90% of the realized final arrival delays y lie within $\hat{y} \pm 90s$). Thus, already at the beginning of the train journey from Zurich to Chur and vice versa, the final arrival delay can be predicted within $\pm 90s$ for 90% of the samples.

Interestingly, the resulting PI widths for desired coverage rates of 50%–80% are significantly closer than for desired coverage rates of 80%–95%. This is a result of the distribution of the final arrival delay, exhibiting a significant tail of higher than average delays, requiring disproportionately larger PIs to achieve a high coverage rates of 90% or 95%.

4.3.2. Dynamic prediction accuracy evaluation

To better understand the value of online information, we analyse the prediction accuracy for different prediction horizons within the DPHF. The MAE evaluations of the proposed uncertainty-aware neural network and the benchmark models are illustrated in Fig. 8. In addition to the commonly-used benchmark models, we consider the current delay and the median historic delay as naive benchmarks. While the MAE of the median historic delay remains constant over time as it does not take into account online information, the current delay solely relies on this online data obtained in real-time. For short prediction horizons (less than 15 minutes), the current delay achieves reasonable performance, since the train approaches its final destination and randomness decreases. Conversely, the median historic delay is a solid predictor for horizons beyond 45 minutes, where online information seems to have only limited influence.

The comparison of the MAE evolution of complex prediction models and the naive benchmarks highlights the strength of complex prediction models. Specifically for a prediction horizon ranging from 10 minutes up to 40 minutes, prediction models provide a more accurate future outlook. Within this time frame, all discussed models outperform the two naive benchmark predictors in our study. Interestingly, the machine learning models NN and NGB continue to exhibit a slight advantage over the median historic delay benchmark event at longer prediction horizons, gradually converging towards the accuracy of that naive benchmark, whereas the markov chain models and the bayesian network show a similar MAE to the historical average.

Only for a prediction horizon of less than 10 minutes, the MC-2stepP model outperforms the proposed NN. This is a result of the MC-2stepP model design (see Section 3) exploiting the low variability of process time deviations, which is especially useful for short-term predictions.

4.4. Quantification of predictability as exponential decay

The likeliness of realization (LoR), given a prediction with confidence estimation, allows us to evaluate the generated stochastic predictions in a meaningful way. Essentially, the LoR reflects how probable a given prediction is to come true. In the context of the DPHF, we can study the evolution of the LoR over time. This analysis enables us to conduct a detailed exploration of the predictability of train delays.

Fig. 9 visualizes the dynamic evaluation of the LoR for the proposed uncertainty-aware NN, commonly used stochastic models and naive benchmark models. The LoR is shown for a symmetric interval of $\Delta = 30s$ around the prediction, but results provided

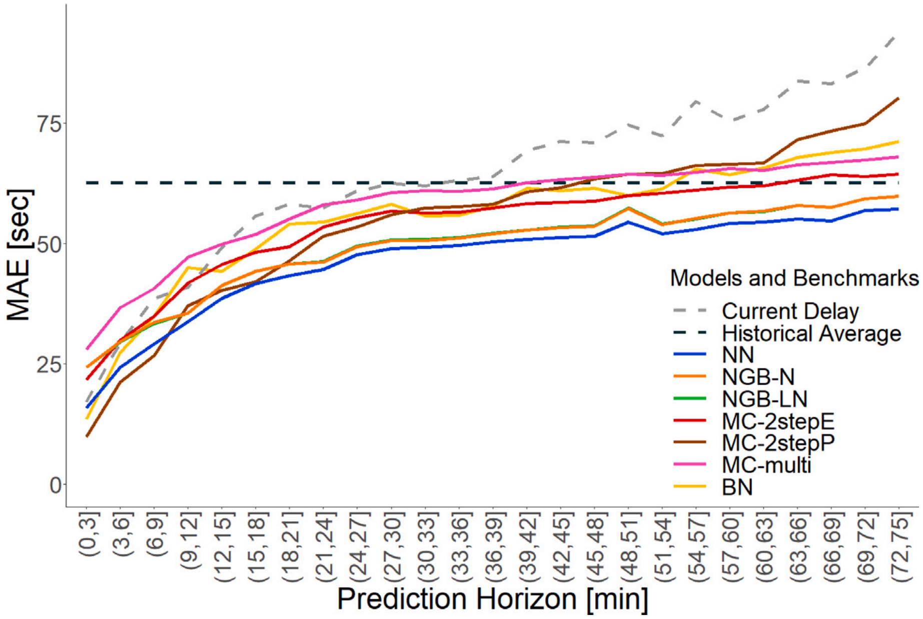


Fig. 8. Mean absolute error per model for increasing prediction horizon intervals.

Table 2
Parameters of the exponential fit per model. The β-parameters are similar across model, indicating that the differences between models are rather due to a horizon-independent ability of the model to fit the data, expressed by the offset γ.

	BN	MC-multi	MC-2stepP	MC-2stepE	NN	NGB-LN	NGB-N
λ	0.15	0.13	0.11	0.13	0.11	0.11	0.10
γ	0.19	0.33	0.30	0.36	0.41	0.29	0.37
α	0.64	0.61	0.85	0.67	0.58	0.43	0.47

in Appendix E suggest a similar shape for other intervals. Similar to the MAE evaluation, we find that the proposed model also outperforms the implemented NGB, MC and BN models in terms of LoR for any prediction horizon above ten minutes.

As depicted in Fig. 9, we find that the decay of the LoR for an increasing prediction horizon can be approximated by an exponential decay function of form $\mathcal{P}(\tau) = \alpha \cdot e^{-\lambda\tau} + \gamma$ for all models. For a detailed analysis of the quality of fit and a comparison to fits of other parametric functions (linear, logarithmic and polynomial), see Appendix I. The observed exponential decay of the LoR is in alignment with the approximately logarithmic increase of the MAE (see Fig. 8), as the LoR can be viewed as an inverse measure of the MAE under consideration of uncertainty.

In Table 2, we discover that the decay rate λ remains remarkably consistent across different models, hovering around $\lambda = 0.11$. The LoR curves mainly differ in their shifts, characterized by the offset parameter γ (e.g., $\gamma = 0.41$ for the NN compared to $\gamma = 0.19$ for the BN), in combination with the factor α. While the offset parameter γ reflects the long-term convergence level of the prediction performance of a model, the decay parameter λ offers insights into how the confidence of train delay predictions evolves with increasing prediction horizons. The results indicate that the level of confidence increases rapidly within thirty minutes before the planned time of the event. However, for longer prediction horizons, the decay rate significantly diminishes.

In Appendix J, we extend the analysis of the NN to the prediction of the delay at intermediate stations. The results show that a similar exponential decay is observed, at least in the cases where the horizon is sufficiently long.

4.4.1. Predictability with respect to offline information

After having demonstrated that the reduction of the LoR for increasing prediction horizon can be well-described by an exponential decay, we seek a deeper understanding of the effects of offline information, such as the train category, direction, and time of the day, on the exponential decay parameters. In other words, we aim to understand how the context influences the quantified predictability of train delays. For categorizing the time of day, we differentiate between trains that depart before noon and those that depart afterwards, keeping in mind that train journeys have a maximum duration of 90 minutes.

Table 3 provides a numeric comparison of the resulting parameters of the exponential decay fits of the LoR, specifically obtained for these offline variables. We find a significant difference in the estimated decay rates λ for fast and slow trains, verifying that the dynamic predictability decay characteristics clearly depend on the train category. The direction and time of the day only minorly affect λ. Also, for the offset parameters γ and α, we find significant differences for different train categories and directions. The

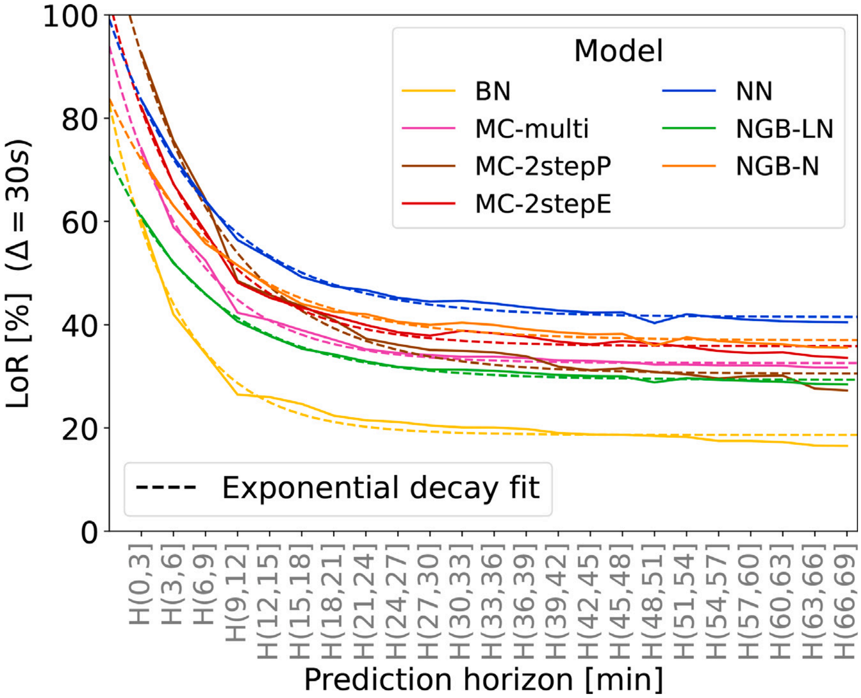


Fig. 9. Fitted exponential decay on the model performance. The models' LoR degrade with a similar decay rate, but shifted.

Table 3
Parameters of the LoR's exponential decay fits, distinguished by direction, train category, and time of the day.

	Direction		Train category		Time of the Day	
	Chur - Zurich	Zurich - Chur	fast	slow	Morning	Afternoon
λ	0.06	0.08	0.06	0.09	0.08	0.07
γ	0.53	0.60	0.47	0.68	0.58	0.55
α	0.49	0.41	0.54	0.34	0.43	0.46

higher level of γ and the lower level of α for trains running from Zurich to Chur indicate that the final delay is slightly easier to predict than for trains running in the opposite direction. The same holds for slow trains in comparison to fast trains.

We extend our analysis by employing a linear regression model to elucidate the underlying relationship between the fitted exponential decay parameters of the LoR and the direction, departure time, train category and average final delay of the train. In this refined analysis, we fit an exponential decay model for the LoR evolution to each individual train. The objective is to systematically examine the influence of these factors on the shape of the decay function.

The findings are detailed in Table 4, listing the coefficients of the regression model. An inspection of the R-squared scores reveals that offline information and the final delay predominantly influence the α and γ parameters, attesting to the substantive effect on the level of the LoR. Interestingly, the fitted exponential decay parameter λ is not to be found significant, implying that the decay itself is consistent across different trains.

Notably, the train category emerges as a significant factor, with fast trains correspondingly exhibiting an increase in α and a decrease in γ in comparison to slow trains. Further, λ exhibits directional dependencies, providing additional insights into the complex interplay between train characteristics and decay parameters. In the Appendix, we complement these numeric results with a visual comparison of LoR (Fig. 17), the epistemic and aleatoric uncertainty (Fig. 13) and the PI width (Fig. 14) divided by train category and direction.

4.4.2. Predictability decay by distance or time?

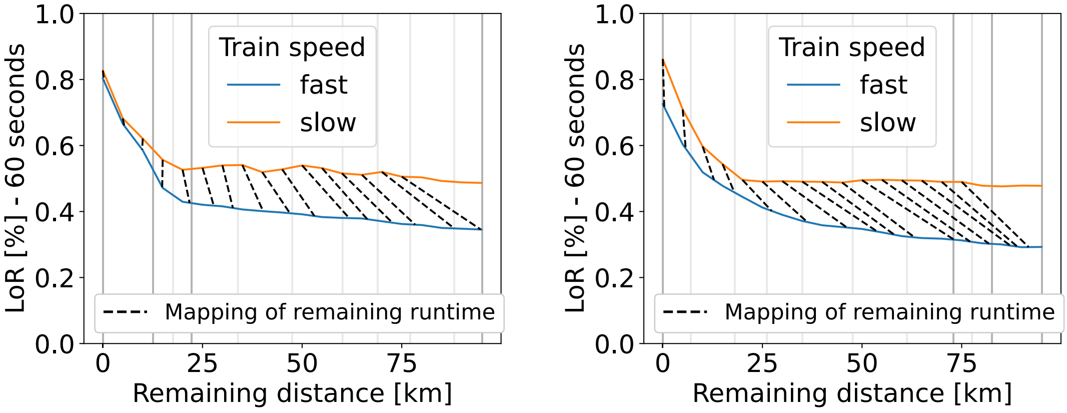
Furthermore, our dataset enables a nuanced analysis of the interplay between time and distance in the decay of predictability for trains. The key question we want to address is whether this decay is more closely linked to an increase in distance to the destination (attributed to stochastic events occurring along the route) or to the planned remaining duration of travel (owing to the accumulation of stochastic effects over time). The relationship between time and distance is inherent to train movement; however teasing apart these effects can be insightful. Analysing the LoR for different train categories, fast and slow — which cover the same distance but vary in runtime, provides a valuable framework to explore this question.

Table 4
Coefficients of a regression model fit to explain the parameters of the exponential decay.

Target variable	α	λ	γ
R-squared	0.170	0.063	0.234
Intercept	0.34 ± 0.04**	0.1 ± 0.06	0.5 ± 0.04**
Daytime (0: before 12, 1: after 12)	0.04 ± 0.03	0.01 ± 0.04	-0.05 ± 0.03
Direction (0: down, 1: up)	-0.01 ± 0.03	0.12 ± 0.04**	0.03 ± 0.03
Final delay [min]	-0.03 ± 0.01**	0.02 ± 0.01	-0.01 ± 0.01*
Train type (0: slow, 2: fast)	0.19 ± 0.04**	0.04 ± 0.05	-0.24 ± 0.03**

* Significant coefficients with $p < 0.05$

** Significant coefficients with $p < 0.01$



(a) Zurich to Chur

(b) Chur to Zurich

Fig. 10. Comparing the LoR evolution between fast and slow trains with respect to distance and time.

Fig. 10 illustrates the evolution of the LoR (in terms of remaining distance) for both train categories and both directions. The dotted lines reveal the matching remaining running time of the slow and fast trains. First, Fig. 10 once again shows that there is a clear difference in the predictability of fast and slow trains in both directions: Slow trains consistently prove to be more predictable for any prediction horizon. Longer runtime thus does not negatively impact predictability; instead, predictability is influenced by a myriad of factors, such as the distribution of supplement time and network effects that deserve investigation in future work. In this study, the distance to the destination, particularly influenced by the number of stops, emerges as the dominant determinant of predictability. More stops appear to enhance predictability, a phenomenon that varies between directions. This variation can be attributed to the skewed distribution of stops for fast trains (see Fig. 5), which are predominantly closer to Chur and consequently enhance the predictability of fast trains at the start of the route.

4.4.3. *Predictability of unpunctual trains*

We also want to analyse the prediction performance by punctuality level. Therefore, Fig. 11 visualizes the dynamic evolution of the LoR, specifically for different punctuality levels. The delay cutoffs were determined by the quantiles of the final delay. Only 1% trains arrive with more than 207 seconds of final delay in our case study.

We find that the prediction performance decreases for higher prediction horizons for punctual and unpunctual trains. Furthermore, higher levels of final delay have a negative impact on the temporal dynamic predictability. Especially for longer prediction horizons, we can find a significant prediction performance gap. This is due to the fact that significant delays can be acquired anywhere within the prediction horizon, and, therefore the final delay of currently punctual trains becomes more difficult to predict for longer prediction horizons.

5. Discussion

We proposed a neural network enriched with epistemic and aleatoric uncertainty estimation, compared it to commonly-used stochastic models and found that the proposed method outperforms commonly-used prediction models as point predictor *and* also provides the best uncertainty estimates. We essentially quantified the temporal dynamic predictability based on the uncertainty-aware prediction quality and showed that the dynamic predictability of train delay for increasing remaining running time and distance can be well-described by an exponential decay. This finding is consistent across models and data subsets and in line with findings from related fields, such as bus arrival time estimation.

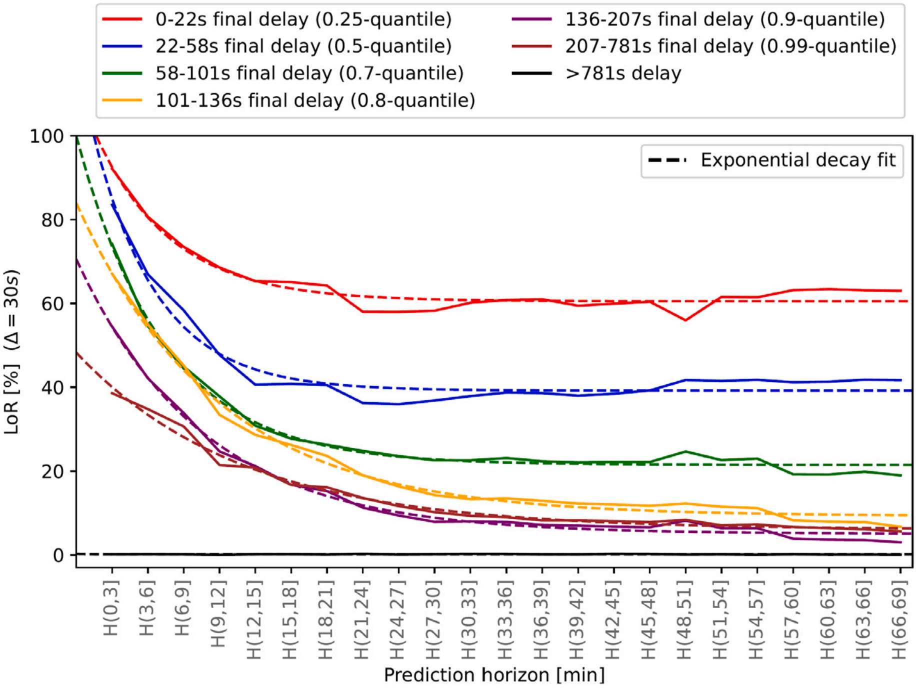


Fig. 11. LoR for different levels of punctuality.

The exponential decay of the predictability of train delay is in strong contrast to other phenomena with a similar analysis of predictability, such as trajectory completion. For the latter, Tsiligkaridis et al. (2022) and Tenzer et al. (2022) have found that the prediction accuracy increases very fast at an early stage of the process (high prediction horizon) and remains at a high level for shorter prediction horizons. We, however, showed that the predictability of train delay is significantly affected by chaotic effects along the remaining distance, where the number of potential problems and disruptions accumulate for increasing prediction horizons. These effects are intertwined with the linearly spaced supplement time. Future studies can, therefore, focus on factors that can interrelate the availability of running time supplement to predictability, which include driver behaviour, availability of a microscopic timetable, traffic density, vehicle circulation, and network bottlenecks.

In our analysis, we explored both aleatoric and epistemic components of uncertainty, uncovering valuable insights for railway operators in optimizing resource investments. Specifically, the limited proportion of epistemic uncertainty, particularly for prediction horizons exceeding twenty minutes in our study, implies that further investments in prediction accuracy might lead to only marginal improvements given the existing input data. Conversely, the notable presence of aleatoric uncertainty points to an opportunity for meaningful enhancements in predictions. By improving data collection methods for currently unobserved variables that impact train delays, it may be possible to tap into this unexploited potential.

Our exploration into predictability offers a more model-independent examination of predictability of train delays than prior studies, which focused solely on the variability for a individual models, such as in Büchel and Corman (2022). Despite still quantifying predictability in terms of model performance, which could theoretically increase with superior modelling, our model comparison strives to minimize epistemic uncertainty. In this way, we isolated the aleatoric uncertainty component, which describes the inherent variability of the data and, and thereby provided a meaningful quantification of the predictability of the phenomenon train delay. While the inclusion of more features might explain some of this variability, the development of novel graph-based models could enable the capture of complex dependencies within train networks, further reducing the errors. This represents a distinct research challenge, yet our work lays a foundation framework for comparing emerging models against established benchmarks.

It is also worth noting that accuracy is likely to vary significantly across different data sets and train lines, particularly considering the low rate of delays in the Swiss network. As expected, punctual trains are more predictable. We found that train delay prediction models clearly outperform the two naive benchmark models, "current delay" and "historical average", for short- to mid-term prediction horizons (i.e., 10–60 min in our case). For shorter prediction horizons, the current delay is a reasonable predictor, and for long-term predictions, the historical average cannot be improved significantly by prediction models taking into account real-time delay information. However, the exponential decay behaviour of the LoR for increasing prediction error in our study was robust and is expected to translate to train delay predictions in other regions, countries or with different methodologies. Indeed, our analysis of the delay at intermediate stations demonstrates the applicability of the framework for online updates of the expected delay along the route. Future research should aim to identify further factors that affect predictability to improve our understanding when, where, and how train delay prediction is beneficial for operators and passengers. Also, the analysis of the delay predictability of longer train journeys and an even more detailed analysis of time-varying characteristics, i.e., a more fine-grained description of peak hours, possible changes in traffic and or traffic management due to peak hours (i.e., prioritization of commuter trains) could benefit the understanding of train delay predictability.

Finally, it is essential to realize the practical usefulness of train delay predictions that extends beyond marginal improvements in accuracy. Quantifying the increasing total uncertainty in train delay predictions empowers railway operators to efficiently allocate resources such as stand-by drivers and rolling stock. By understanding the level of uncertainty associated with different prediction horizons, operators can make well-informed decisions, ensuring effective contingency plans and minimizing disruptions to services. Further research is needed to understand how to optimally provide stochastic information in the design of decision support systems for traffic controllers and for travellers, with the goal to maximize transparency while avoiding information overload.

From a passenger's perspective, the quantified predictability of train delays provides fundamental information for risk-based decision-making. For instance, passengers inside a train might use the upper limit of the prediction interval to evaluate their onward travel connections, while those waiting at stations could use the lower limit of the prediction interval to prepare for boarding. In our study, we observe a significant increase in prediction quality within a prediction horizon of less than thirty minutes. This knowledge allows passengers to make informed choices, such as arriving at the platform just in time or preparing for potential delays when planning their onward journey upon arrival.

6. Conclusion

In this research study, we have proposed, for the first time in the context of train delay prediction, to apply an uncertainty-aware neural network within a dynamic prediction horizon framework (DPHF) to precisely quantify the predictability of train delays. This innovative approach has allowed us to evaluate prediction performance based on a-priori provided uncertainty estimations, evaluated by the proposed likeliness of realization (LoR), and not only based on traditional accuracy measures. The insights drawn from the influencing factors of the temporal dynamic predictability foster a more comprehensive understanding of the prediction quality.

Our findings demonstrate high confidence levels in short-term predictions, which gradually decrease for longer prediction horizons. By employing an exponential decay model, we have managed to encapsulate the predictability with only three parameters, providing an ideal base for comparison. The uncertainty-aware neural network also allowed us to split the total uncertainty into its aleatoric and epistemic components, shedding light on a majority of inherent data variability and model uncertainty, especially for increasing prediction horizons.

Our main contribution lies in the enriched understanding of train delay prediction quality across varying prediction horizons. The introduced methodology holds significant potential to aid railway operators in optimizing resource allocation. Moreover, it enables passengers to make risk-based decisions by providing quantified predictability of train delays.

For future research, a promising direction lies in incorporating more data sources. This could potentially reduce the aleatoric uncertainty within train delay predictions, particularly for longer prediction horizons, thus further enhancing the efficacy of this innovative approach to train delay prediction.

CRediT authorship contribution statement

Thomas Spanninger: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Nina Wiedemann:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Francesco Corman:** Conceptualization, Funding acquisition, Supervision, Validation, Writing – original draft, Writing – review & editing.

Acknowledgements

This work is partially supported by the Swiss National Science Foundation, Switzerland under Project 1481210/DADA, and by the ETH Zürich Foundation, Switzerland project number: 2023-FS-331 on Research in the area of Public Transport Design and Operations. We sincerely thank Swiss Federal Railways for sharing expertise and data, without which those analyses would not have been possible, at this level of detail. Furthermore, we thank Martin Raubal for the valuable feedback.

Appendix A. Distribution of errors of the neural network

We argue that the use of symmetric prediction intervals, as well as the assumption of normally distributed errors for estimating aleatoric uncertainty, is justified by the distribution of errors presented in Fig. 12.

Appendix B. Epistemic vs aleatoric uncertainty by direction and train speed

Fig. 13 shows the epistemic and aleatoric uncertainty by train category and direction.

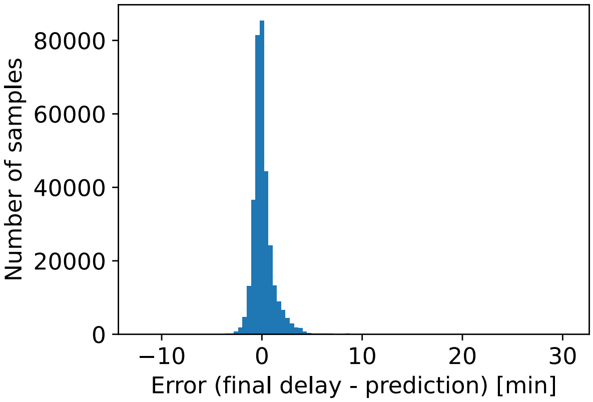
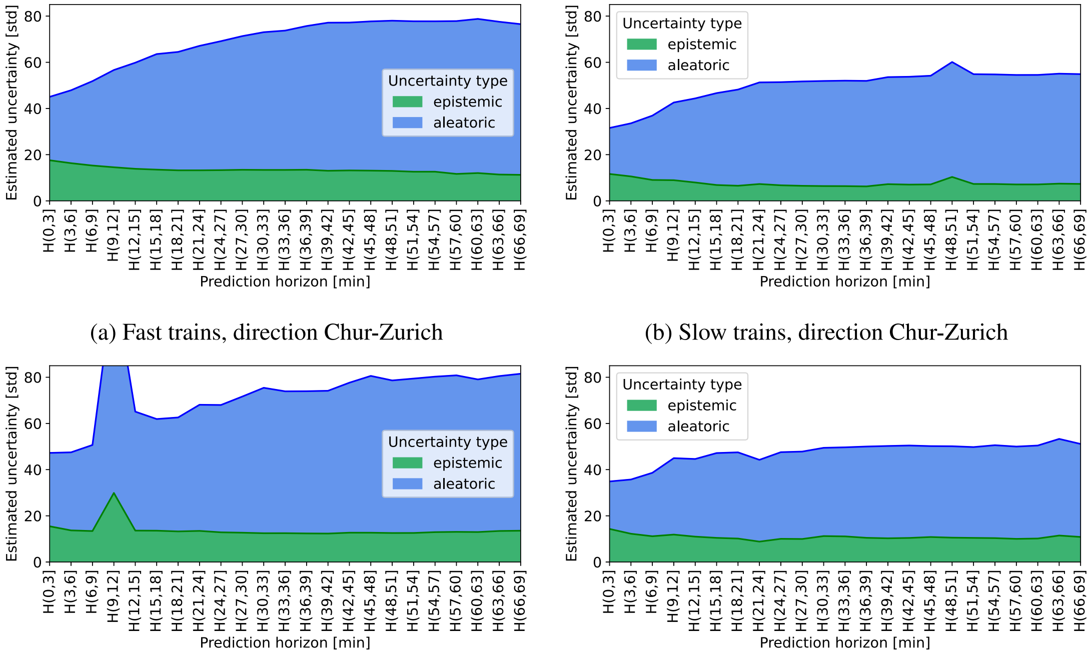


Fig. 12. Distribution of errors of the point predictions by the neural network. The errors are distributed approximately symmetrically, with few outliers where the final delay is underestimated.



(a) Fast trains, direction Chur-Zurich

(b) Slow trains, direction Chur-Zurich

(c) Fast trains, direction Zurich-Chur

(d) Slow trains, direction Zurich-Chur

Fig. 13. Epistemic and aleatoric uncertainty by train category and direction.

Appendix C. Prediction interval width by direction and train speed

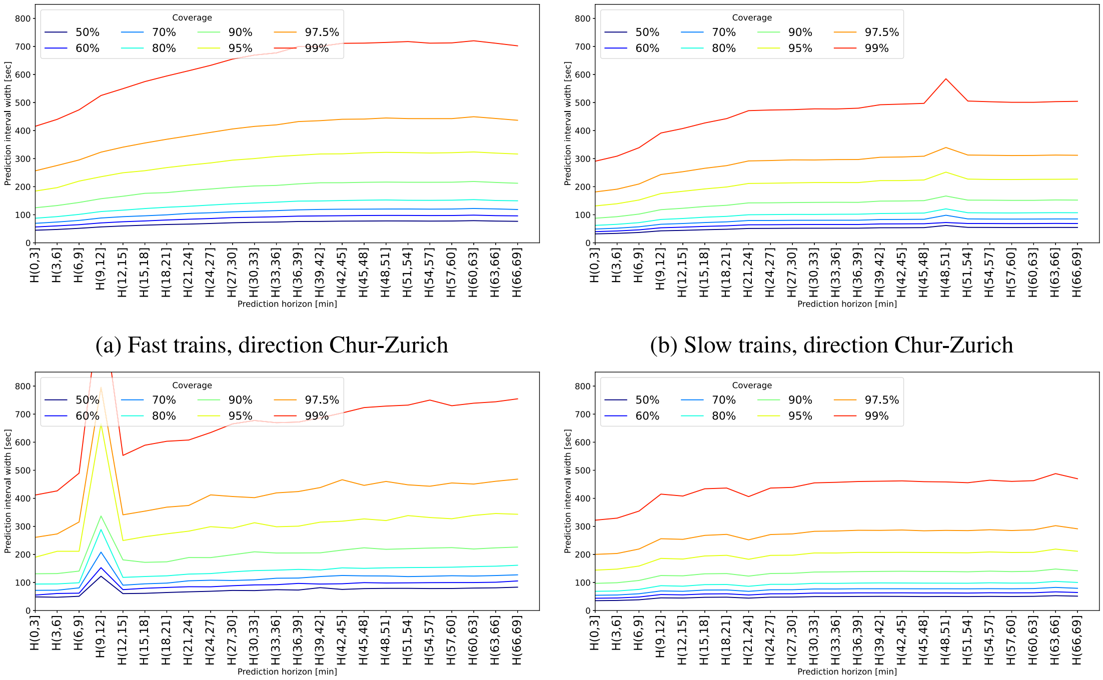
Fig. 14 shows the prediction interval width by train type (direction and speed).

Appendix D. Detailed MAE evaluation for increasing prediction horizon per direction and train speed

Fig. 15 visualizes the evolution of the MAE for distinct directions and train categories.

Appendix E. Levels of LoR evolution for increasing prediction horizon

Fig. 16 visualizes the levels of LoR for a 30, 60 and 90 seconds interval ($\pm 15, 30, 45$ seconds around the ground truth). The uncertainty-aware NN achieves the highest (=best) LoR for all prediction horizons. Therefore, we only visualize its results.



(a) Fast trains, direction Chur-Zurich

(b) Slow trains, direction Chur-Zurich

(c) Fast trains, direction Zurich-Chur

(d) Slow trains, direction Zurich-Chur

Fig. 14. Prediction interval width by train category and direction.

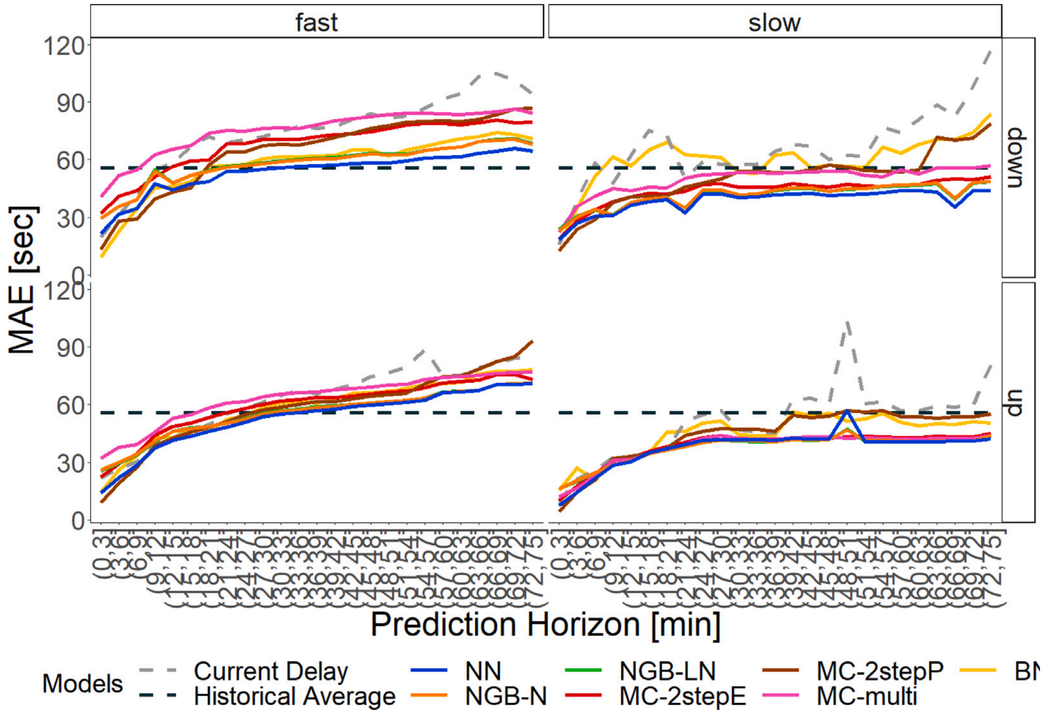


Fig. 15. MAE for an increasing prediction horizon per train category and direction.

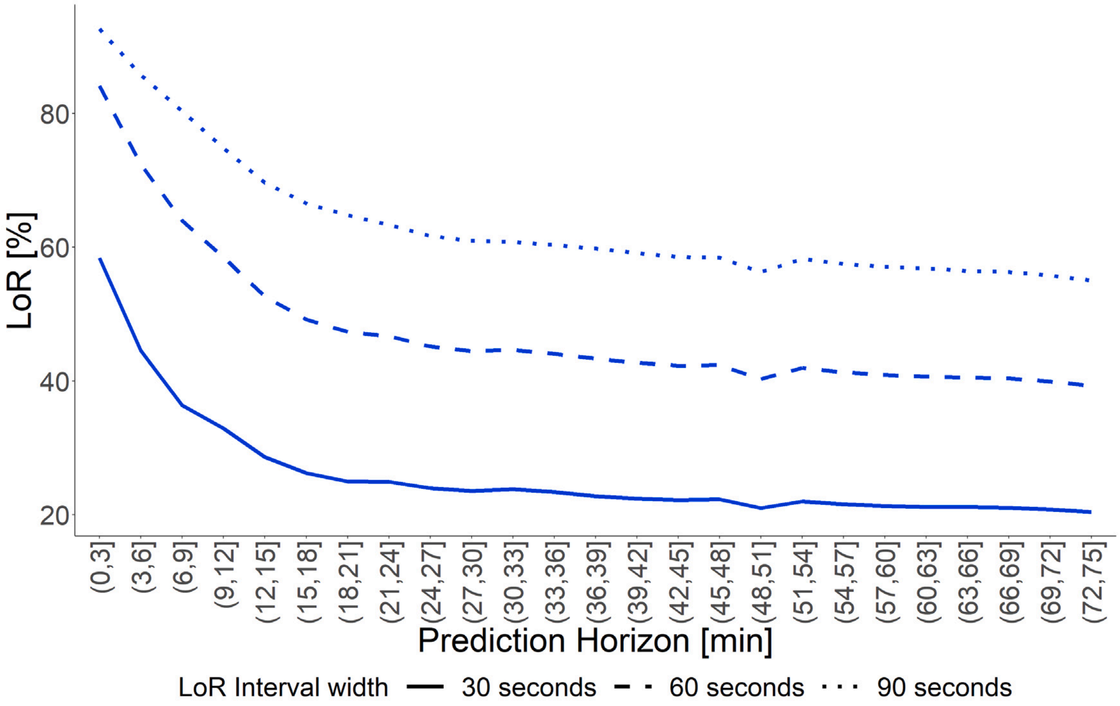


Fig. 16. Likeliness of realizations for given predictions for an increasing prediction horizon and varying levels of LoR interval widths.

We can observe that the LoR (level 30 seconds) decreases from 56% to 21% for an increasing prediction horizon. The 90 seconds LoR level, on the other hand, decreases from 92% to 57%.

The probabilistic methods do not seem to benefit, in terms of LoR performance, from their ability to approximate a full probability distribution; a property that we expected to offer more precise confidence estimates when measured by the LoR.

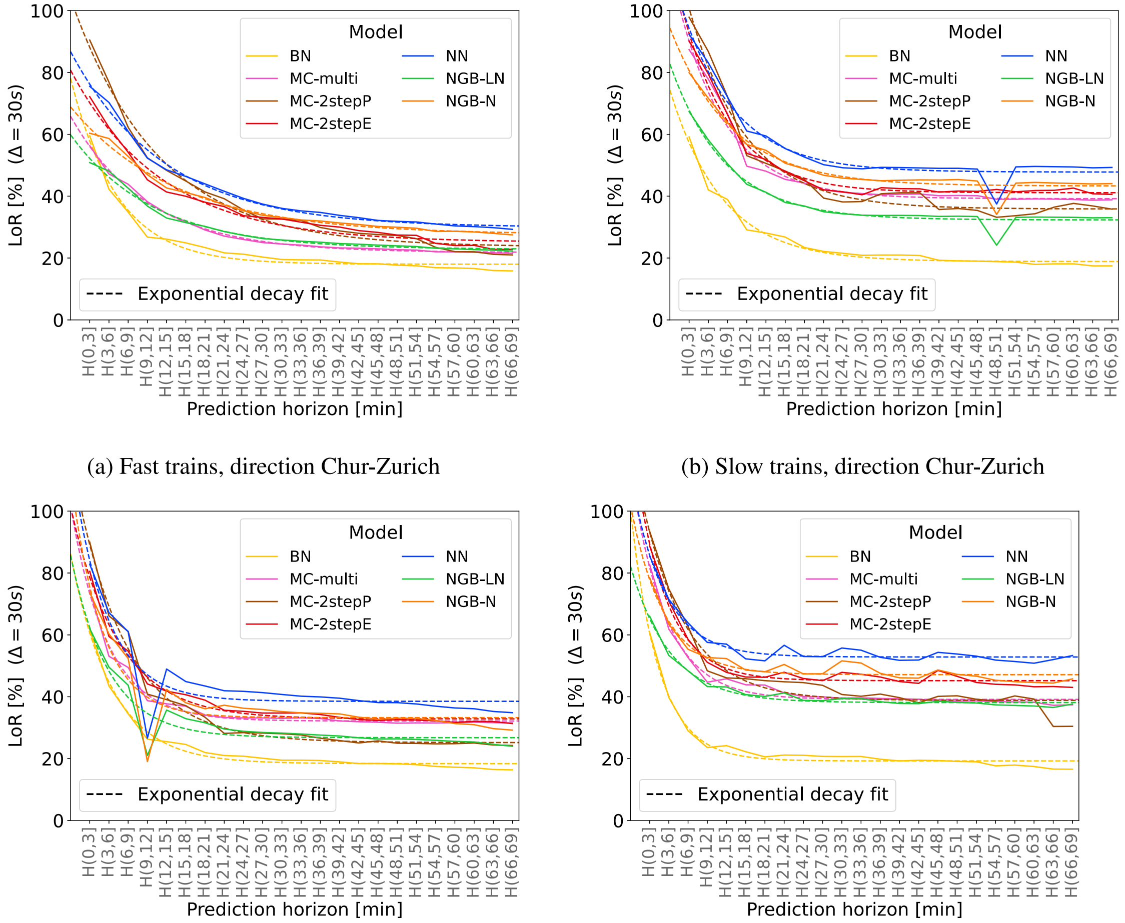
Appendix F. Exponential decay of LoR by train type

In addition to the overall exponential decay shown in Fig. 9, we distinguish by train type (slow/fast) and journey direction in Fig. 17. The outliers in the curves are due to a different subset of observations being present in each time interval. It can be observed that fast and slow trains are significantly different, corresponding to the results presented in Section 4.4.1.

Appendix G. Sensitivity analysis with respect to NN parameters and input features

To test for the sensitivity of the neural network to its hyperparameters, we vary the network's architecture and compare the performance on the test data. We further compare the performance of the NN as well as the NGBoost baseline with additional input features. Specifically, three input feature sets are compared: The set we are currently using ("Ours") as discussed in Section 3, a comprehensive set of all features that we computed ("All features"),[3] a small set of features selected with a Lasso-regression enforcing sparse coefficients ("Lasso-selected"), and two intermediate version where supplement time ("Ours + supplement time") as well as buffer time and weather features ("Ours + weather + supplement & buffer time") are respectively added to our current set of features. Table 5 shows that the architecture or feature set only causes minor differences in the RMSE or LoR, at least in relation to the difference between methods. The MAE of the NN increases if only using Lasso-selected features and decreases when including all features, while it is the other way round for the RMSE. Supplement time seems to be noisy since it generally decreases performance. Overall, the differences are minor, testifying sufficient robustness of the NN.

[3] Features comprise: Planned arrival – planned observation time, planned arrival – actual observation time, weather features (Average temperature in Celsius, dew point in Celsius, relative humidity, wind speed, peak wind gust, air pressure; extracted with the `meteostat` Python library from the weather stations closest to the observation), supplement and buffer time, current delay, whether the train is a fast or slow train, its direction, its number of stops, the delay at the last five observations, the delay at this observation in the last five days, the final delay in the past five days, the average delay over all historic data of the train, the delay of other trains on that day, the current time and the planned arrival time (converted to sine and cosine features since circular).



(a) Fast trains, direction Chur-Zurich

(b) Slow trains, direction Chur-Zurich

(c) Fast trains, direction Zurich-Chur

(d) Slow trains, direction Zurich-Chur

Fig. 17. Likeliness of realization by train category and direction.

Appendix H. Ablation study: Training ML models on observation subsets

In the results presented so far, the ML models were trained on a single dataset comprising all trains and observations. However, it may be beneficial to train one model *per observation* to avoid "contaminating" the model with the data from other observations. We test this hypothesis by training one model per observation, and comparing its performance to the general model (trained on all observations) on test data for this observation point.

For the sake of efficiency, we only do this experiment for every 10th observation. Fig. 18 shows that, surprisingly, the errors are larger for the specialized model, indicating that it is more important to feed the model with as many samples as possible, and that it is beneficial to learn across observations. Thus, in all other experiments we only report the performance for the model that was trained on all samples.

Appendix I. Modelling the decay of the LoR

In our analysis of the decay in the likeliness of realization (LoR) for increasing prediction horizons of train delay, we evaluate different model fits using both the Akaike Information Criterion (AIC) and the coefficient of determination (R^2). The AIC results are suggestive of the exponential model being the most appropriate, with a value of −532.115, representing the optimal balance between fit and complexity. This is further corroborated by the R^2 evaluation, where the exponential model yields an impressive 0.9950473, meaning it accounts for 99.5% of the variance in the decay of LoR. The polynomial model follows closely but the exponential model, being simpler, is the preferred choice. Both the linear and logarithmic models fall short in comparison to the exponential model

Table 5
Sensitivity to hyperparameters and input features. The employed architecture is marked bold.

Feature set	Model	Parameters[a]	MAE	RMSE	$LoR_{\Delta=30s}$ [%]
All features (45)	NGBoost (Lognormal)		49.74	103.07	0.33
	NGBoost (Normal)		49.72	101.44	0.42
	Neural network	2, 128, 128, 1e−5	45.41	96.22	0.49
	Simple median		62.6	137.89	0.38
Lasso-selected (7)	NGBoost (Lognormal)		49.91	101.77	0.33
	NGBoost (Normal)		49.89	100.75	0.42
	Neural network	2, 128, 128, 1e−5	47.93	95.49	0.45
	Simple median		62.6	137.89	0.38
Ours (11)	NGBoost (Lognormal)		49.67	102.05	0.33
	NGBoost (Normal)		49.69	100.98	0.42
	Neural network	**2, 128, 128, 1e-5**	**46.74**	**95.91**	**0.47**
	Neural network	2, 64, 128, 1e−5	46.54	95.85	0.48
	Neural network	2, 128, 64, 1e−4	46.77	96.49	0.47
	Neural network	2, 128, 64, 1e−5	46.68	96.21	0.48
	Neural network	2, 256, 64, 1e−5	46.48	95.86	0.48
	Neural network	2, 128, 128, 1e−6	47.19	96.15	0.47
	Neural network	3, 128, 128, 1e−5	46.66	96.37	0.48
	Simple median		62.6	137.89	0.38
Ours + supplement time (12)	NGBoost (Lognormal)		51.53	104.38	0.33
	NGBoost (Normal)		52.22	104.73	0.42
	Neural network	2, 128, 128, 1e−5	48.09	98.1	0.45
	Simple median		62.6	137.89	0.38
Ours + weather + supplement & buffer time (19)	NGBoost (Lognormal)		49.5	102.58	0.33
	NGBoost (Normal)		49.37	101.12	0.42
	Neural network	2, 128, 128, 1e−5	46.28	98.56	0.51
	Simple median		62.6	137.89	0.38

[a] Parameters of the NN are given as (number of layers, number of neurons in 1st layer, number of neurons in 2nd layer, learning rate)

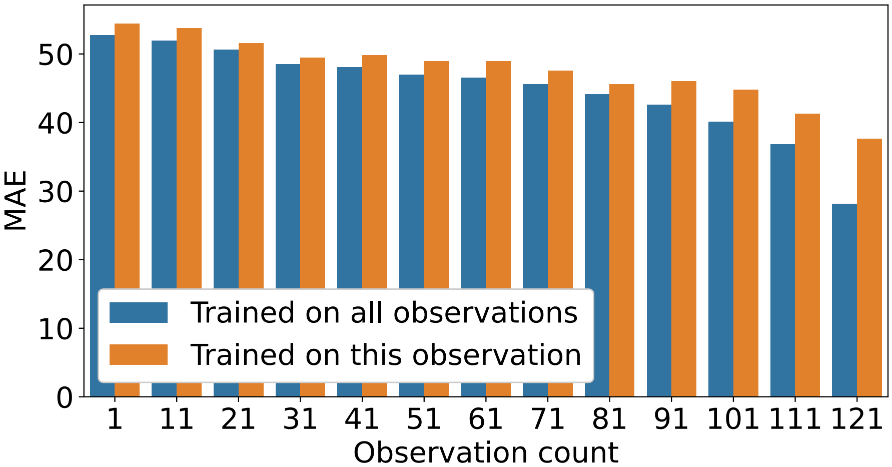


Fig. 18. Comparing a general model that was trained on all points of operations simultaneously to a specialized model that was trained on one point of operation at the time. Separate training does not improve model performance on this observation.

in terms of both AIC and R^2, reinforcing the exponential fit as the most suitable representation for the decay of the likeliness of realization for train delays over increasing prediction horizons. Fig. 19 visualizes the fits of the decay of the LoR given prediction provided by the suggested uncertainty-aware neural network approach. Table 6 provides the numerical results in details.

We also fit the LoR decay specifically per direction and train category. The results are visualized in Fig. 20. Table 7 provides the resulting R^2 values in detail.

Appendix J. Predictability of delay at intermediate stations

To validate our results obtained on predicting the final delay of the train ride from Zurich to Chur or the other direction, we re-train and evaluate the NN for predicting the delay at intermediate stations. Therefore, we selected Landquart as an intermediate

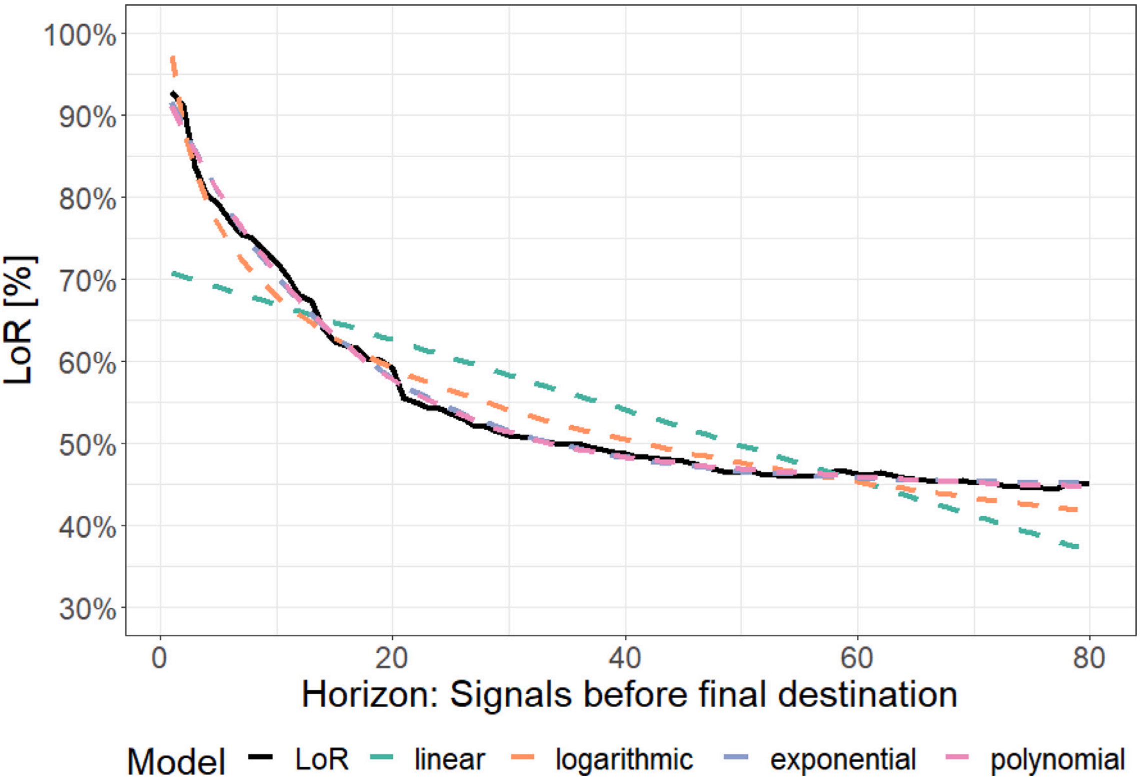


Fig. 19. Comparison of the fitted linear, logarithmic, exponential and polynomial models for the evolution of the LoR for an increasing prediction horizon.

Table 6
Comparison of AIC and R^2 values for different model fits.

Model	AIC	R^2
Linear	−208.91	0.71
Exponential	−532.12	0.99
Logarithmic	−380.03	0.96
Polynomial	−529.86	0.99

Table 7
R^2 values for the fitted decay models per direction and train category, and averaged.

Train category	Direction	Linear	Exponential	Logarithmic	Polynomial
Fast	Down	0.62	0.98	0.93	0.97
Fast	Up	0.87	1.00	0.97	1.00
Slow	Down	0.42	0.97	0.82	0.92
Slow	Up	0.72	0.98	0.92	0.99
Average		0.66	0.98	0.91	0.97

stop for fast trains and Thalwil as an intermediate station for slow trains. For fast trains, Landquart is the first stop on the ride from Chur to Zurich. Slow trains stop for the first time in Thalwil from Zurich to Chur. Four NNs are trained, one for each intermediate station and direction.

Fig. 21 presents the decay of the LoR for predicting the arrival delay at intermediate stops. The comparison of the decay of the respective prediction performances confirms previously obtained results of the general exponential decay finding for the final delay prediction quality for all trains (grey line). When predicting the delay at the first station (Landquart on the route from Chur to Zurich), the performance is consistently high, since the train is hardly unpunctual on this short segment. Another reason for the lack of decay in predictability may be a large supplement time in Landquart. Fig. 5(a) supports this explanation as a large drop

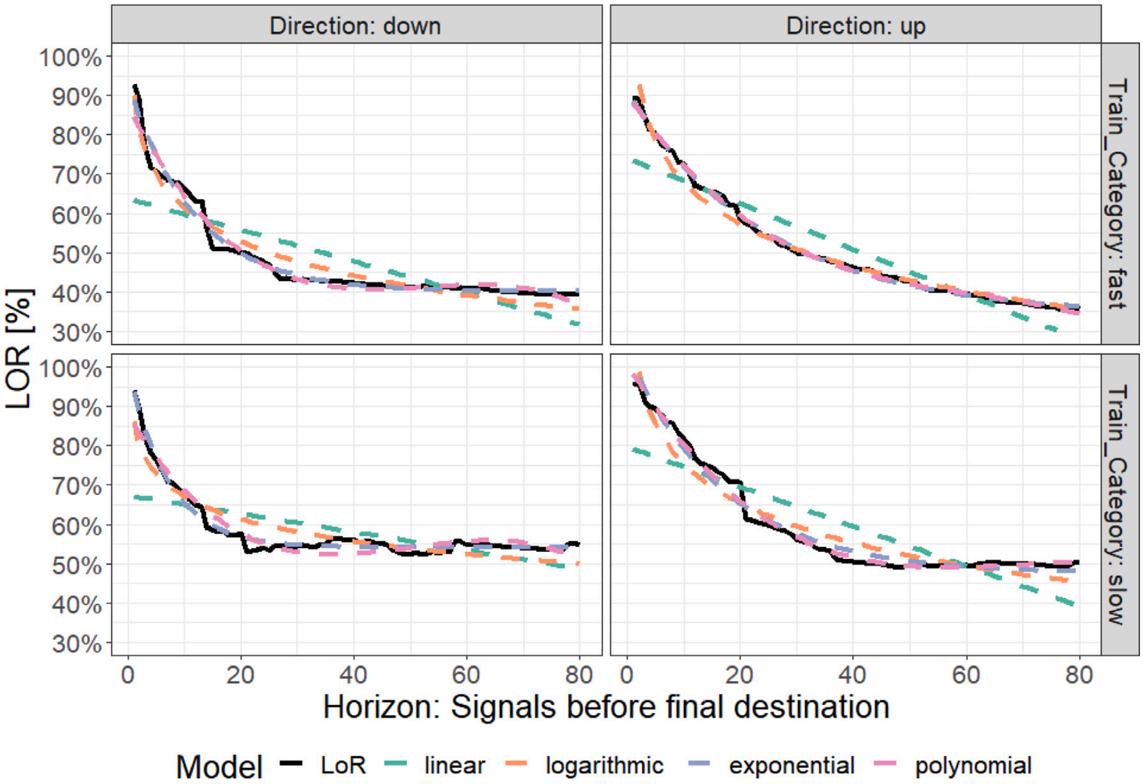


Fig. 20. Comparison of the fitted linear, logarithmic, exponential and polynomial models for distinct direction and speed for the evolution of the LoR for an increasing prediction horizon.

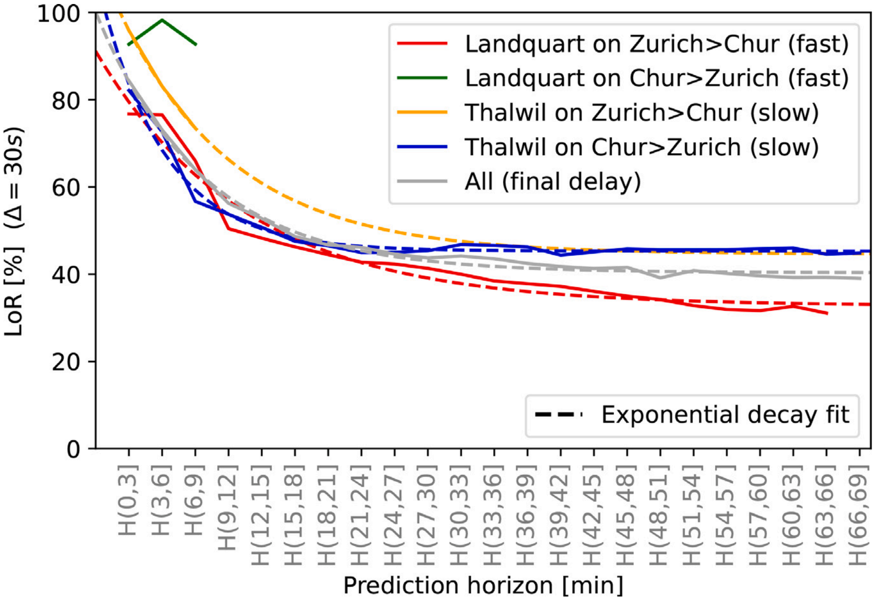


Fig. 21. LoR for predicting the delay at intermediate stations.

Table 8
Comparison of the average prediction accuracy in terms of MAE and RMSE and the average quality of predicted confidence (LoR) for the proposed models.

	MAE [s]	% to NN	RMSE [s]	% to NN	$LoR_{\Delta=30s}$ [%]	% to NN
Results reported in Section 4.2						
NN	46.1	0.0%	88.28	0.0%	46.92	0.0%
NGB-LN	48.84	+5.9%	92.06	+4.3%	33.17	−29.3%
NGB-N	48.92	+6.1%	91.7	+3.9%	41.55	−11.4%
MC-2stepE	52.59	+14.1%	107.29	+21.5%	40.53	−13.6%
MC-2stepP	57.73	+25.2%	94.74	+7.3%	36.83	−21.5%
MC-multi	57.05	+23.8%	109.8	+24.4%	36.49	−22.2%
Simple median	60.2	+30.6%	121.94	+38.1%	38.22	−18.5%
Current delay	68.46	+48.5%	108.32	+22.7%	28.9	−38.4%
Results including all available data per model:						
NN	46.74	0.0%	95.91	0.0%	47.36	0.0%
NGB-LN	49.67	+6.3%	102.05	+6.4%	33.46	−29.3%
NGB-N	49.69	+6.3%	100.98	+5.3%	41.85	−11.6%
MC-2stepE	54.39	+16.4%	121.22	+26.4%	40.81	−13.8%
MC-2stepP	57.73	+23.5%	94.74	−1.2%	36.83	−22.2%
MC-multi	59.0	+26.2%	124.04	+29.3%	36.75	−22.4%
BN	55.04	+17.8%	95.96	+0.1%	26.36	−44.3%
Simple median	62.6	+33.9%	137.89	+43.8%	38.08	−19.6%
Current delay	68.66	+46.9%	112.49	+17.3%	29.06	−38.6%
Results excluding outliers:						
NN	42.98	0.0%	69.07	0.0%	47.54	0.0%
NGB-LN	45.43	+5.7%	69.47	+0.6%	33.59	−29.3%
NGB-N	45.55	+6.0%	69.62	+0.8%	42.01	−11.6%
MC-2stepE	49.27	+14.6%	78.71	+14.0%	40.95	−13.9%
MC-2stepP	56.27	+30.9%	86.4	+25.1%	36.91	−22.4%
MC-multi	53.75	+25.1%	82.45	+19.4%	36.89	−22.4%
BN	52.08	+21.2%	76.19	+10.3%	26.44	−44.4%
Simple median	56.91	+32.4%	94.74	+37.2%	38.22	−19.6%
Current delay	65.43	+52.2%	94.85	+37.3%	29.17	−38.6%

in the variation between current and final delay is observed at the station in Landquart. For Thalwil, on the hand, we can already observe a decay within the first 10 minutes of the ride.

Appendix K. Comparison of results with and without outliers included

The RMSE is strongly affected by outliers. Since the NN output is normalized, the model fails to predict the delay of strongly delayed trains. In addition, the MC requires a minimum number of observations to fit the transition probability matrices, resulting in the exclusion of samples at ops with only very limited observations in Section 4.2. For the sake of completeness, we compare the reported results to the ones with all samples included in Table 8. When the models are evaluated on all possible samples, the performance of the NN decreases in comparison to the MC-2stepP model. However, this is due to the inclusion of samples with large final delay in the NN results, which are not included for the MC-2stepP model.

Furthermore, Table 8 shows the results when removing all outliers from the analysis, defined as trains with a final delay of more than 15 minutes. In this case, the RMSE is generally lower and the NN also achieves best performance in all three metrics.

References

Agarap, Abien Fred, 2018. Deep learning using rectified linear units (ReLU). arXiv preprint arXiv:1803.08375.

Angelopoulos, Anastasios, Bates, Stephen, 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

Artan, Mehmet Şirin, Şahin, İsmail, 2022. Exploring patterns of train delay evolution and timetable robustness. IEEE Trans. Intell. Transp. Syst. 23 (8), 11205–11214.

Artan, Mehmet Şirin, Şahin, İsmail, 2023. A stochastic model for reliability analysis of periodic train timetables. Transportmetrica B 11 (1), 572–589.

Ash, Robert B., 1965. Information Theory. Dover Publications.

Bao, Xu, Li, Yanqiu, Li, Jianmin, Shi, Rui, Ding, Xin, 2021. Prediction of train arrival delay using hybrid ELM-PSO approach. J. Adv. Transp. 2021, 1–15.

Barta, János, Rizzoli, Andrea Emilio, Salani, Matteo, Gambardella, Luca Maria, 2012. Statistical modelling of delays in a rail freight transportation network. In: Proceedings of the 2012 Winter Simulation Conference. WSC, IEEE.

Berger, Annabell, Gebhardt, Andreas, Müller-Hannemann, Matthias, Ostrowski, Martin, 2011. Stochastic delay prediction in large train networks. In: 11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Boateng, Veronica, Yang, Bo, 2023. Ensemble stacking with the multi-layer perceptron neural network meta-learner for passenger train delay prediction. In: 2023 IEEE Conference on Artificial Intelligence. CAI, IEEE, pp. 21–22.

Büchel, Beda, Corman, Francesco, 2022. What do we know when? Modeling predictability of transit operations. IEEE Trans. Intell. Transp. Syst. 23 (9), 15684–15695.

Büchel, Beda, Spanninger, Thomas, Corman, Francesco, 2021. Modeling evolutionary dynamics of railway delays with Markov chains. In: 7th International Conference on Models and Technologies for Intelligent Transportation Systems. MT-ITS.

Büker, Thorsten, Seybold, Bernhard, 2012. Stochastic modelling of delay propagation in large networks. J. Rail Transp. Plan. Manag. 2, 34–50.

Carey, Malachy, Kwieciński, Andrzej, 1994. Stochastic approximation to the effects of headways on knock-on delays of trains. Transp. Res. B 28 (4), 251–267.

Cats, Oded, Loutos, Gerasimos, 2016. Real-time bus arrival information system: An empirical evaluation. J. Intell. Transp. Syst. 20 (2), 138–151.

Chien, Steven I-Jy, Ding, Yuqing, Wei, Chienhung, 2002. Dynamic bus arrival time prediction with artificial neural networks. J. Transp. Eng. 128 (5), 429–438.

Corman, Francesco, Kecman, Pavle, 2018. Stochastic prediction of train delays in real-time using Bayesian networks. Transp. Res. C 95, 599–615.

Duan, Tony, Anand, Avati, Ding, Daisy Yi, Thai, Khanh K, Basu, Sanjay, Ng, Andrew, Schuler, Alejandro, 2020. NGBoost: Natural gradient boosting for probabilistic prediction. In: International Conference on Machine Learning. PMLR, pp. 2690–2700.

Gal, Yarin, Ghahramani, Zoubin, 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gaurav, Ramashish, Srivastava, Biplav, 2018. Estimating train delays in a large rail network using a zero shot Markov model. In: 2018 21st International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 1221–1226.

Gawlikowski, Jakob, Tassi, Cedrique Rovile Njieutcheu, Ali, Mohsin, Lee, Jongseok, Humt, Matthias, Feng, Jianxiang, Kruspe, Anna, Triebel, Rudolph, Jung, Peter, Roscher, Ribana, et al., 2021. A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342.

Grinsztajn, Léo, Oyallon, Edouard, Varoquaux, Gaël, 2022. Why do tree-based models still outperform deep learning on tabular data? arXiv preprint arXiv:2207.08815.

Guo, Chuan, Pleiss, Geoff, Sun, Yu, Weinberger, Kilian Q., 2017. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR, pp. 1321–1330.

Hallowell, Susan, Harker, Patrick, 1996. Predicting on-time line-haul performance in scheduled railroad operations. Transp. Sci. 30, 364–378.

van Hooff, Madelon L.M., 2015. The daily commute from work to home: Examining employees' experiences in relation to their recovery status. Stress Health 31 (2), 124–137.

Huang, Ping, Spanninger, Thomas, Corman, Francesco, 2022. Enhancing the understanding of train delays with delay evolution pattern discovery: A clustering and Bayesian network approach. IEEE Trans. Intell. Transp. Syst. 23 (9), 15367–15381.

Huang, Ping, Wen, Chao, Fu, Liping, Lessan, Javad, Jiang, Chaozhe, Peng, Qiyuan, Xu, Xinyue, 2020. Modeling train operation as sequences: A study of delay prediction with operation and weather data. Transp. Res. E 141, 102022.

Kecman, Pavle, Corman, Francesco, Meng, Lingyun, 2015. Train delay evolution as a stochastic process. In: 6th International Conference on Railway Operations Modelling and Analysis - RailTokyo2015. pp. 1–27.

Kecman, Pavle, Goverde, Rob M.P., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transp. 7, 295–319.

Kendall, Alex, Gal, Yarin, 2017. What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Keyhani, Mohammad, Schnee, Mathias, Weihe, Karsten, Zorn, Hans-Peter, 2012. Reliability and delay distributions of train connections. In: 12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Kiureghian, Armen Der, Ditlevsen, Ove, 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112.

Lakshminarayanan, Balaji, Pritzel, Alexander, Blundell, Charles, 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 30.

LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey, 2015. Deep learning. Nature 521 (7553), 436–444.

Lemnian, Martin, Rückert, Ralf, Rechner, Steffen, Blendinger, Christoph, Müller-Hannemann, Matthias, 2014. Timing of train disposition: Towards early passenger rerouting in case of delays. In: 14th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Leng, Nuannuan, Corman, Francesco, 2020. The role of information availability to passengers in public transport disruptions: An agent-based simulation approach. Transp. Res. A 133, 214–236.

Lessan, Javad, Fu, Liping, Wen, Chao, 2019. A hybrid Bayesian network model for predicting delays in train operations. Comput. Ind. Eng. 127, 1214–1222.

Lijesen, Mark G., 2014. Optimal traveler responses to stochastic delays in public transport. Transp. Sci. 48 (2), 256–264.

Liu, Yafei, Tang, Tao, Xun, Jing, 2017. Prediction algorithms for train arrival time in urban rail transit. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 1–6.

Lyons, Glenn, Chatterjee, Kiron, 2008. A human perspective on the daily commute: Costs, benefits and trade-offs. Transp. Rev. 28 (2), 181–198.

Malavasi, Gabriele, Ricci, Stefano, 2001. Simulation of stochastic elements in railway systems using self-learning processes. European J. Oper. Res. 131 (2), 262–272.

Martin, Layla, Wittmann, Michael, Li, Xinyu, 2021. The influence of public transport delays on mobility on demand services. Electronics 10 (4), 379.

Mazloumi, Ehsan, Rose, Geoff, Currie, Graham, Moridpour, Sara, 2011. Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. Eng. Appl. Artif. Intell. 24 (3), 534–542.

Meester, Ludolf E., Muns, Sander, 2007. Stochastic delay propagation in railway networks and phase-type distributions. Transp. Res. B 41 (2), 218–230.

Mentch, Lucas, Hooker, Giles, 2016. Quantifying uncertainty in Random Forests via confidence intervals and hypothesis tests. J. Mach. Learn. Res. 17 (1), 841–881.

Nabian, Mohammad Amin, Alemazkoor, Negin, Meidani, Hadi, 2019. Predicting near-term train schedule performance and delay using bi-level random forests. Transp. Res. Rec. 2673 (5), 564–573.

Nix, David, Weigend, Andreas, 1994. Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 IEEE International Conference on Neural Networks. ICNN'94, IEEE, pp. 55–60.

Oneto, Luca, Fumeo, Emanuele, Clerico, Giorgio, Canepa, Renzo, Papa, Federico, Dambra, Carlo, Mazzino, Nadia, Anguita, Davide, 2018. Train delay prediction systems: A big data analytics perspective. Big Data Res. 11, 54–64.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037.

Peters, Jan, Emig, Bastian, Jung, Marten, Schmidt, Stefan, 2005. Prediction of delays in public transportation using neural networks. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, CIMCA-IAWTIC'06, vol. 2, IEEE, pp. 92–97.

Şahin, İsmail, 2017. Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances. J. Rail Transp. Plan. Manag. 7 (3), 101–113.

Şahin, İsmail, 2022. Data-driven stochastic model for train delay analysis and prediction. Int. J. Rail Transp. 1–20.

Scutari, Marco, Denis, Jean-Baptiste, 2021. Bayesian Networks: With Examples in R. CRC Press.

Shafer, Glenn, Vovk, Vladimir, 2008. A tutorial on conformal prediction. J. Mach. Learn. Res. 9 (3).

Shi, Rui, Wang, Jing, Xu, Xinyue, Wang, Mingming, Li, Jianmin, 2020. Arrival train delays prediction based on gradient boosting regression tress. In: Proceedings of the 4th International Conference on Electrical and Information Technologies for Rail Transportation, EITRT 2019: Rail Transportation Information Processing and Operational Management Technologies. Springer, pp. 307–315.

Spanninger, Thomas, Büchel, Beda, Corman, Francesco, 2021. Probabilistic predictions of train delay evolution. In: 7th International Conference on Models and Technologies for Intelligent Transportation Systems. MT-ITS.

Spanninger, Thomas, Büchel, Beda, Corman, Francesco, 2023. Train delay predictions using Markov chains based on process time deviations and elastic state boundaries. Mathematics 11 (4), 839.

Spanninger, Thomas, Trivella, Alessio, Büchel, Beda, Corman, Francesco, 2022. A review of train delay prediction approaches. J. Rail Transp. Plan. Manag. 22, 100312.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, Salakhutdinov, Ruslan, 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Sun, Dihua, Luo, Hong, Fu, Liping, Liu, Weining, Liao, Xiaoyong, Zhao, Min, 2007. Predicting bus arrival time on the basis of global positioning system data. Transp. Res. Rec. J. Transp. Res. Board 2034 (1), 62–72.

Taylor, John Robert, Thompson, William, 1982. An introduction to error analysis: The study of uncertainties in physical measurements, volume 2. Springer.

Tenzer, Mark, Rasheed, Zeeshan, Shafique, Khurram, Vasconcelos, Nuno, 2022. Meta-learning over time for destination prediction tasks. In: Proceedings of the 30th International Conference on Advances in Geographic Information Systems. pp. 1–10.

Tiong, Kah Yong, Ma, Zhenliang, Palmqvist, Carl-William, 2023. A review of data-driven approaches to predict train delays. Transp. Res. C 148, 104027.

Tsiligkaridis, Athanasios, Zhang, Jing, Paschalidis, Ioannis Ch, Taguchi, Hiroshi, Sakajo, Satoko, Nikovski, Daniel, 2022. Context-aware destination and time-to-destination prediction using machine learning. In: 2022 IEEE International Smart Cities Conference. ISC2, IEEE, pp. 1–7.

Vovk, Vladimir, Gammerman, Alexander, Shafer, Glenn, 2005. Algorithmic Learning in a Random World. Springer Science & Business Media.

Wang, Yuexin, Wen, Chao, Huang, Ping, 2022. Predicting the effectiveness of supplement time on delay recoveries: A support vector regression approach. Int. J. Rail Transp. 10 (3), 375–392.

Wen, Chao, Mou, Weiwei, Huang, Ping, Li, Zhongcan, 2020. A predictive model of train delays on a railway line. J. Forecast. 39 (3), 470–488.

Wilson, Andrew, Izmailov, Pavel, 2020. Bayesian deep learning and a probabilistic perspective of generalization. Adv. Neural Inf. Process. Syst. 33, 4697–4708.

Yaghini, Masoud, Khoshraftar, Mohammad M., Seyedabadi, Masoud, 2013. Railway passenger train delay prediction via neural network model. J. Adv. Transp. 47 (3), 355–368.

Yuan, Jianxin, Hansen, Ingo, 2007. Optimizing capacity utilization of stations by estimating knock-on train delays. Transp. Res. B 41 (2), 202–217.