Check for updates

RESEARCH ARTICLE

FI-TW: An open train–weather dataset for railway delay analysis in Finland

[version 1; peer review: 2 approved with reservations]

Vinicius Pozzobon Borin, Jean Michel de Souza Sant'Ana, Usama Raheel, Nurul Huda Mahmood

Centre for Wireless Communications (CWC), University of Oulu, Oulu, Finland

v1 **First published:** 03 Sep 2026, **6**:324
https://doi.org/10.12688/openreseurope.24750.1
Latest published: 03 Sep 2026, **6**:324
https://doi.org/10.12688/openreseurope.24750.1

Abstract

Background

Train delays result from complex interactions between operational, technical, and environmental factors. Weather strongly affects railway reliability in Nordic regions, yet publicly available datasets rarely integrate meteorological information with operational train records, which limits research on weather-driven delay.

Methods

We constructed the Finland Integrated Train-Weather (FI-TW) dataset by combining operational records from the Finland Digitraffic Railway Traffic Service with observations from 209 Finnish Meteorological Institute stations, covering January 2018 to December 2024. Train events and weather measurements were aligned in space and time using the Haversine distance, with a radial fallback strategy that recovers missing parameters from alternative nearby stations. Processing included cyclical encoding of temporal features, robust scaling of weather data to limit the effect of sensor outliers, duplicate removal, and the derivation of weather-scenario indicators together with multi-scale rolling-window aggregations.

Results

The dataset contains 138 features spanning operational variables and meteorological measurements, and approximately 38.5 million observations from Finland's 5,915-kilometer rail network. Exploratory analysis shows a clear seasonal structure, with winter delay rates

Open Peer Review

Approval Status ? ?

	1	2
version 1 03 Sep 2026	? view	? view

1. **Luca ONETO**, Università di Genova, Genova, Italy
2. **Juan Pineda-Jaramillo**, Universidad Politécnica de Valencia, Valencia, Spain

Any reports and responses or comments on the article can be found at the end of the article.

exceeding 25% compared with below 20% in summer, and geographic clustering of high-delay corridors in central and northern Finland. A baseline experiment using extreme gradient boosting (XGBoost) regression reached a mean absolute error of 2.73 minutes for station-specific delay prediction.

Conclusions

FI-TW is, to the best of our knowledge, the first publicly available dataset that couples Finnish railway operations with synchronized meteorological observations over a seven-year span. It offers multiple target variable formulations and supports applications such as delay prediction, weather impact assessment, seasonal reliability analysis, and infrastructure vulnerability mapping, providing a reusable resource for machine learning research on railway operations in severe northern climates.

Plain Language summary

Trains in Finland often run late, and the weather is one of the reasons. Snow, ice, and very low temperatures can damage equipment, slow trains down, and disrupt signalling, while in autumn fallen leaves make the rails slippery. Researchers who want to measure how much the weather really matters need data describing both the trains and the weather at the same place and the same moment. Until now that combination has been hard to find, because public railway datasets usually describe only train movements and weather datasets usually describe only the atmosphere.

This work presents FI-TW, a free and open dataset that joins the two. It brings together seven years of Finnish train records, from January 2018 to December 2024, with measurements from 209 weather stations across the country. For every point where a train passes or stops, the dataset reports how early or late it was and what the weather was doing nearby at that moment. Where a nearby weather station was missing a measurement, the value was taken from the closest station that recorded it.

The result is a single collection of about 38.5 million records, each described by 138 pieces of information. In the data, more than one train event in four is delayed during the winter months, compared with fewer than one in five during summer, and the most affected routes lie in central and northern Finland. As a demonstration, a standard machine learning model trained on the data predicted station arrival delays with an average error of under three minutes.

The dataset is free for anyone to download and reuse. It can support work on predicting delays, measuring how much specific weather conditions cost the railway, planning timetables that account for the season, and identifying the parts of the network most vulnerable to extreme weather.

Keywords
dataset, finland, freight transportation and logistics, data analytics, data science, machine learning, rail transportation

Corresponding author: Vinicius Pozzobon Borin (vinicius.borin@oulu.fi)

Author roles: Pozzobon Borin V: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Michel de Souza Sant'Ana J**: Conceptualization, Data Curation, Methodology, Supervision, Validation, Writing – Review & Editing; **Raheel U**: Data Curation, Software, Visualization; **Huda Mahmood N**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was partially supported in Finland and Sweden by the European Union through the Interreg Aurora project ENSURE-6G (Grant Number: 20361812); and in Finland by the Research Council of Finland through the project 6G Flagship (Grant Number: 369116) and by the European Commission through the Horizon Europe/JU SNS project Hexa-X-II (Grant Number: 101095759).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Pozzobon Borin V *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Pozzobon Borin V, Michel de Souza Sant'Ana J, Raheel U and Huda Mahmood N. **FI-TW: An open train–weather dataset for railway delay analysis in Finland [version 1; peer review: 2 approved with reservations]** Open Research Europe 2026, **6**:324 https://doi.org/10.12688/openreseurope.24750.1

First published: 03 Sep 2026, **6**:324 https://doi.org/10.12688/openreseurope.24750.1

Background

Railway transportation constitutes one of the most widely utilized modes of transportation globally, serving billions of passengers annually and facilitating the movement of goods across vast distances.[1] From the high-speed rail networks of Asia to the transcontinental systems of North America, railways form the backbone of sustainable mobility in both developed and developing nations. In Europe particularly, rail transportation plays a vital role in economic infrastructure and cultural connectivity, with the continent's extensive network spanning over 245,000 kilometers. This network serves millions of passengers daily, connecting major urban centers across diverse geographical and climatic conditions. As modern societies increasingly prioritize sustainable transportation alternatives to reduce carbon emissions and traffic congestion, railway systems have become critical infrastructure for both daily commuting and long-distance travel. Their reliability is therefore a key concern for passengers, government agencies, and private companies.

Within the scope of railway logistics, train delays represent one of the most significant concerns facing operators and passengers worldwide, caused by an interplay of technical, operational, environmental, and human factors.[2] Infrastructure-related issues such as track conditions, signaling system failures, and rolling stock malfunctions can cascade through interconnected networks, affecting multiple services simultaneously. Operational factors, including schedule optimization, crew management, and station dwell times, further compound these challenges. In addition, external elements such as adverse weather conditions, passenger behavior during peak hours, and unexpected incidents introduce additional variability.[2] The ability to accurately predict and analyze these delays has significant implications for both operational efficiency and passenger satisfaction. For railway operators, delay prediction enables proactive resource allocation, dynamic rescheduling, and improved maintenance planning. For passengers, reliable delay information facilitates better travel planning, reduces uncertainty, and minimizes the inconvenience associated with disrupted journeys.[3]

The convergence of advanced wireless communication technologies and artificial intelligence (AI) presents unprecedented opportunities for transforming railway operations. Next-generation networks such as 5G and emerging 6G systems enable real-time, high-bandwidth data collection from distributed sensors across railway infrastructure and station facilities.[4] This enhanced connectivity facilitates continuous monitoring of critical parameters, including track conditions, vehicle performance metrics, and passenger flows. It thereby creates rich datasets that were previously difficult or impossible to collect centrally and analyze in real time. When combined with modern AI techniques, particularly deep learning (DL) and predictive analytics, these data streams enable railway operators to anticipate potential failures, optimize maintenance schedules, and predict delays before they occur. Machine learning (ML) models trained on comprehensive operational data can identify complex patterns indicative of emerging problems, allowing for proactive interventions rather than reactive responses. This shift from traditional scheduled maintenance to predictive, condition-based approaches not only reduces operational costs but also significantly improves service reliability and safety across the entire railway network.[5]

The availability of train-related datasets has expanded in recent years across four primary thematic categories: traffic planning and management, maintenance and inspection, safety and security, and passenger mobility.[6] However, a significant limitation persists: most existing datasets either omit weather conditions entirely or incorporate only a limited set of meteorological parameters. This gap presents a critical challenge for research requiring simultaneous analysis of operational and meteorological factors. Such analysis is essential given the potential impact of weather on railway performance.

Related work in railway research

A comprehensive systematic review[6] examined 62 publicly available AI-oriented datasets for railway applications, finding that while operational data dominates the field (47 of 62 datasets), the vast majority do not incorporate meteorological data. Among the few that couple railway data with environmental conditions, RailSem19[7] provides 8,500 annotated sequences for semantic scene understanding, complementing the Cityscapes dataset,[8] while the Indian Metro Dataset[9] pairs traffic and passenger-flow data with a limited set of weather conditions (humidity, wind, visibility, and precipitation). Both, however, capture only a narrow range of meteorological parameters.

More recently, a small number of datasets have begun to integrate operational records with weather observations more explicitly. Zhang et al.[10] released a Chinese high-speed railway dataset covering 727 stations and 3,399 trains over a 16-week window (2019–2020), annotated with weather condition, temperature, and wind level. Wu et al.[11] published an Italian railway network dataset spanning 1,397 stations and 3,324 trains over six months of 2024, combining multi-type train operation records with temperature, wind, and general weather conditions retrieved from a public weather service and, as in our work, using the Haversine formula to compute inter-station distances. At nationwide scale, the RIDE benchmark[12] links 94.5 million Belgian train events with 35.7 million weather records (2023–2025) and provides a

standardized evaluation protocol for delay-prediction models. These efforts confirm the growing recognition that weather is integral to railway analytics; nevertheless, they remain constrained by comparatively short temporal spans, a narrow set of meteorological variables, and none captures the sub-Arctic conditions that dominate Nordic rail operations. The dataset proposed here is distinguished by its seven-year span (2018–2024), its breadth of meteorological coverage (13 weather parameters underlying 138 features, including engineered weather-scenario indicators and multi-scale rolling-window aggregations), its fine-grained station-level spatial matching with a radial fallback strategy, and its focus on Finland's severe northern climate.

Recent delay prediction studies have begun incorporating weather factors. Huang et al.[13] developed FCLL-Net for Chinese high-speed railways, achieving 9.4% improvement by capturing train interactions alongside temperature, wind, and rainfall. Another study[14] applied cost-sensitive deep learning (FCF-Net) to model timetables as images for delay propagation pattern recognition. Sajan et al.[15] evaluated regression models for Indian Railways, finding elastic-net regression effective when calibrated with weather variables, though departure delay remained the strongest predictor. Notably, Oneto et al.[16] achieved significant improvements using Deep Extreme Learning Machines on Italian railway data, but explicitly identified weather data integration as critical future work. In Nordic contexts, where winter weather is especially disruptive, Zakeri and Olsson[17] analysed ten years of punctuality data from Norway's Nordland Line and identified snow depth as the strongest weather-related predictor of low punctuality, with extreme cold winters driving the largest delays. More recently, Soleimani-Chamkhorami et al.[18] fused Swedish infrastructure-failure records with meteorological observations to classify climate-related failures through machine learning, reporting snow, ice, and low temperature as the dominant factors. Collectively, these studies underscore the value of a richly featured, weather-integrated dataset for Nordic railway research, yet each relies on proprietary or task-specific data rather than a reusable public resource.

Proposed dataset

To address the identified gap in integrated train-weather datasets, this work introduces the Finland Integrated Train–Weather (FI-TW) dataset, a novel resource from the Finnish railway system that systematically combines operational train data with comprehensive meteorological information. The Finnish railway network comprises over 5,900 kilometers of track and serves over 90 million passengers annually.[19] It experiences diverse climatic conditions ranging from Nordic coastal regions to severe Arctic weather, making it an interesting context for studying weather impacts on railway operations.

The dataset integrates temporal operational metrics, like train schedules, train types, routes, and delays, with synchronized weather observations, enabling comprehensive analysis of environmental impacts on railway performance. The main contributions of this work are:

- The development of the first (to the best of our knowledge) publicly available dataset integrating operational data with meteorological data for the Finnish railway network.
- Step-by-step data processing to create a cohesive dataset for ML applications.
- A comprehensive spatial-temporal coverage across diverse climatic conditions.
- A multi-dimensional integration of operational and meteorological variables with precise temporal synchronization.
- The consideration of multiple research applications beyond delay prediction: weather impact assessment, seasonal reliability analysis, weather-adaptive scheduling, and infrastructure vulnerability mapping.

A previous version of this work was made available as a preprint.[20]

Collection methods and design

This study focuses on long-distance services, which connect major cities across Finland over main lines spanning several hundred kilometers. According to VR Group, the country's primary passenger rail operator, more than 15 million of Finland's passenger journeys in 2024 were long-distance.[21] A representative route is the 875-kilometer line that traverses the country, connecting the capital Helsinki in the south to the arctic city Rovaniemi in the north.

Punctuality for these services is assessed by the Finnish Transport Infrastructure Agency based on arrival at the final destination, where a train is considered on time if it arrives within five minutes of its scheduled time. In 2024, 86.28% of

long-distance trains met this standard.[22] Maintaining such punctuality across an extensive network is challenging, as signal and track equipment malfunctions, including power failures, communication breakdowns, and point failures, frequently disrupt operations until repairs are completed.

Finland's extreme climate poses particularly severe challenges, as winter temperatures reaching −40°C can cause mechanical failures in automatic doors, couplings, and switching systems. Heavy snowfall disrupts signaling equipment and requires extensive track clearing operations.[23] During autumn, fallen leaves create slippery layers on rails, reducing adhesion and requiring trains to operate at lower speeds for safety.[24] These examples of weather-related and technical issues often cascade through the interconnected rail network, amplifying delays across multiple routes.

To analyze these operational challenges and delay patterns systematically, this study constructs a comprehensive dataset by integrating two Finnish open data sources that are described in this section. The dataset encompasses observations collected from January 2018 to December 2024. All example data presented in this section utilize observations from a long-distance train service operating between Oulu and Helsinki on December 1st, 2024.

Digitraffic railway traffic data

Operated by Fintraffic, Digitraffic Railway Traffic Service [1] provides comprehensive real-time and historical data about trains operating throughout the Finnish railway network. The dataset includes train timetables (scheduled and actual departure/arrival times), real-time location data (Global Positioning System [GPS] coordinates) and speed, train composition details, and station infrastructure metadata for Finland's state-owned railway network. For historical data, access is provided through an Application Programming Interface (API). The service requires no authentication and enforces a rate limit of 60 requests per minute per Internet Protocol (IP) address, with most responses cached for approximately one minute. All data are provided under the Creative Commons Attribution 4.0 International license (CC BY 4.0), enabling free use with appropriate attribution.

We retrieved historical data from the API endpoint [2], which returns all trains operated on a given departure date. Table 1 presents the descriptors of the data retrieved from this endpoint. As shown in this table, the dataset includes a column labeled *timeTableRows*, which contains an array of schedule information for each train number operating on that specific date. The structure of the timetable data is detailed in Table 2.

Table 3 presents a detailed example of train schedule data, showing only relevant columns. Each row in the table represents a track section traversed by the train, which does not necessarily correspond to a passenger station. The table

Table 1. Fintraffic dataset descriptors.

Column Name	Description
trainNumber	Unique train identification number
departureDate	Date of train departure
operatorUICCode	UIC code of train operator
operatorShortCode	Short code for operator (e.g., VR)
trainType	Type of train (e.g., IC)
trainCategory	Train category (e.g., Long-distance)
commuterLineID	Commuter line identifier
runningCurrently	Boolean if train is running
cancelled	Boolean if train is cancelled
version	Version number of timetable entry
timetableType	Type of timetable (e.g., REGULAR)
timetableAcceptanceDate	Date timetable was accepted
timeTableRows	Nested data object of station timing data

[1]Service webpage: https://www.digitraffic.fi/en/railway-traffic/
[2]API Endpoint: /api/v1/trains/{departure_date}

Table 2. Timetablerows data descriptors.

Column Name	Description
stationName	Name of the railway station
stationShortCode	Short code identifier for the station
type	Event type (ARRIVAL or DEPARTURE)
scheduledTime	Scheduled arrival/departure time (UTC+2)
actualTime	Actual arrival/departure time (UTC+2)
differenceInMinutes	Delay in minutes (actual - scheduled)
cancelled	Boolean indicating if stop was cancelled
stationUICCode	UIC code for the station
countryCode	ISO country code of station location
trainStopping	Boolean indicating if train stops at station
commercialStop	Boolean indicating if stop is commercial
commercialTrack	Track number for commercial operations
causes	Array of delay/disruption causes
trainReady	Train readiness information
liveEstimateTime	Real-time estimated arrival/departure time
estimateSource	Source of time estimate data

includes two temporal columns: *scheduledTime* and *actualTime.* The *differenceInMinutes* column is pre-calculated by Fintraffic and retrieved directly from the API as

$$\text{differenceInMinutes} = \text{actualTime} - \text{scheduledTime}. \tag{1}$$

This value represents the delay or early arrival at each point

$$\begin{cases} \text{On time}, & \text{if differenceInMinutes} = 0, \\ \text{Delayed}, & \text{if differenceInMinutes} > 0, \\ \text{Ahead of schedule}, & \text{if differenceInMinutes} < 0. \end{cases} \tag{2}$$

The *type* column in Table 3 indicates whether a row corresponds to an arrival (ARR) or departure (DEP) event at a given station. For intermediate stations, each stop generates two records: one for arrival and one for departure. This distinction allows for precise tracking of dwell time at stations and delay propagation throughout the journey. The origin station (Oulu asema) contains only a DEP record, while the final destination (Helsinki asema) contains only an ARR record. Consequently, a journey with n stopping points generates $2n$ records for intermediate stations, plus the origin departure and destination arrival, totaling $2n - 1$ rows when all stations involve commercial stops.

Train data overview

The geographical distribution of Finnish railway stations shows 549 stations located within Finnish territory during the data collection period (2018-2024), comprising 340 (61.9%) non-passenger stations designated for freight/cargo handling, docks, or technical service points and 209 (38.1%) passenger stations.

The functional classification of stations within the Finnish network reveals a hierarchical structure with conventional stations (passengers and non-passengers) constituting 81.07% (454 locations), stopping points representing 11.43% (64 locations), and turnouts in the open line accounting for 7.50% (42 locations). This distribution is characteristic of Finland's railway system, where approximately 5,200 km of the 5,915 km network consists of single-track lines that require strategic passing locations and auxiliary facilities for operational efficiency.

Table 3. **Train schedule data: oulu to Helsinki.**

#	stationName	stationShortCode	type	scheduledTime	actualTime	differenceInMinutes	trainStopping	commercialStop
1	Oulu asema	OL	DEP	05:49	05:50	1	Yes	Yes
2	Oulunlahti	OLL	ARR	05:53	05:57	4	Yes	Yes
3	Oulunlahti	OLL	DEP	05:53	05:57	4	Yes	Yes
4	Kempele	KML	ARR	05:57	06:01	4	Yes	Yes
5	Kempele	KML	DEP	05:57	06:01	4	Yes	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
198	Käpylä	KÄP	ARR	11:26	11:29	3	Yes	Yes
199	Käpylä	KÄP	DEP	11:26	11:29	3	Yes	Yes
200	Pasila asema	PSL	ARR	11:28	11:31	3	Yes	Yes
201	Pasila asema	PSL	DEP	11:29	11:33	4	Yes	Yes
202	Helsinki asema	HKI	ARR	11:35	11:37	2	Yes	Yes

Open Research Europe 2026, 6:324 Last updated: 01 OCT 2026

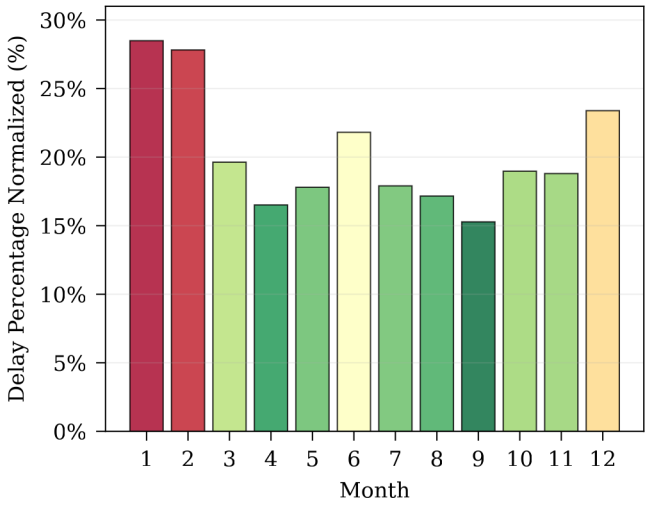


Figure 1. Aggregated normalized delays by month (2018-2024).

Train delays overview

The temporal analysis of delays during the period 2018-2024 in Finland reveals a distinct seasonal pattern. Figure 1 shows aggregated normalized delay data by month. Normalization represents the ratio of the total delays in the train schedule to the total number of train schedules. A train is considered delayed if it arrives five minutes or more after the scheduled time. Importantly, our methodology differs from the Finnish Transport Infrastructure Agency's approach: we examine each individual train schedule stop along the route rather than evaluating punctuality at the final destination only, providing a more comprehensive assessment of service reliability throughout the entire journey. It is important to note that, although the dataset flags whether each delay's officially recorded cause is weather-related (via the *causes_related_to_weather* feature introduced in Section 4), the detailed underlying causes are not further resolved in this aggregated view and could be attributed to various factors such as weather-related conditions or passenger-induced delays at stations. In particular, extreme winter months (Dec-Jan-Feb) experience substantially higher delay percentages. In addition, June exhibits a relatively higher delay percentage (21.8%) compared to other summer months, which could be related to the increased number of trains in circulation during the summer season and the higher volume of passengers traveling for holidays during this period.

Taking into account a range of seven years of data (2018-2024), the distribution of delay intensity categories between days is illustrated in Figure 2. It reveals that medium delays (10-15 minutes) are the most prevalent, with 1254 (49%) occurrences, followed by high delays (15-20 minutes), with 597 (23.35%) cases, and very high delays (more than 20 minutes) with 405 (15.84%) occurrences. Low delays (5-10 minutes) represent the smallest category with 300 (11.73%) entries.

The average delay percentage distribution presented in Figure 3 reveals weekly and seasonal patterns. Weekdays consistently exhibit higher delays compared to weekends, and Fridays demonstrate particularly high delays throughout the year, reaching up to 31.6% in February. In contrast, Saturday demonstrates the lowest delay percentages throughout the year, averaging between 12.2% and 25.5%. A clear seasonal trend is again observed, with the cold months (January, February, and December) experiencing substantially higher delays throughout all weekdays, often exceeding 25%, while the late Spring and early Autumn months (April, May, August, and September) show relatively lower delays, frequently below 20%.

Finnish meteorological institute data

The Finnish Meteorological Institute (FMI) is a government agency under Finland's Ministry of Transport and Communications. This service offers comprehensive weather, sea, and climate observation data and forecasts alongside numerical model outputs from meteorological stations throughout Finland. The dataset [3] encompasses comprehensive meteorological and environmental observations. It includes real-time weather measurements and weather radar data from 10 C-band Doppler radars, which provide reflectivity and precipitation intensity measurements. Marine observations

[3]Service webpage: https://en.ilmatieteenlaitos.fi/open-data

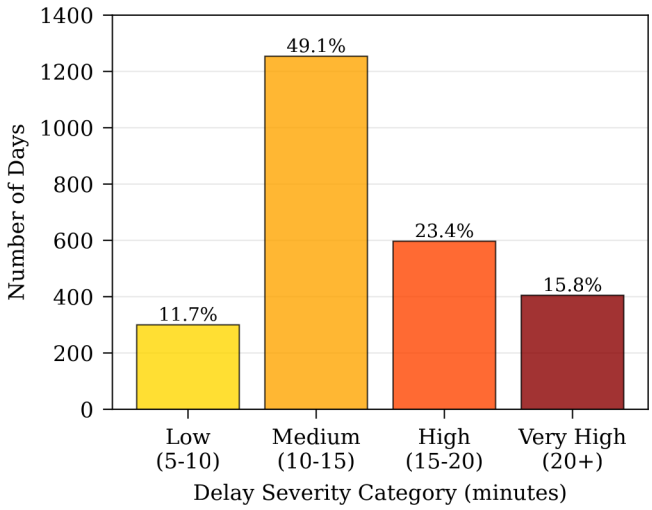


Figure 2. Distribution of delay events by delay-intensity category (2018-2024).

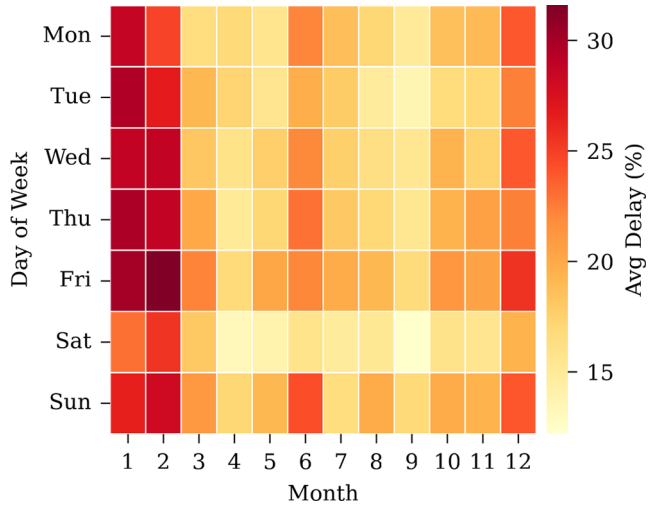


Figure 3. Average delays by week (2018-2024).

cover sea level, wave height, and ocean currents. The dataset also contains air quality measurements and numerical weather prediction model outputs from the European Centre for Medium-Range Weather Forecasts (ECMWF).

Data access is provided through a Web Feature Service (WFS) API that follows international Open Geospatial Consortium standards [4], which provides access to multipoint weather station data. The service requires no authentication, but implements rate limits of 20,000 requests per day and 600 requests per five minutes to ensure service stability. All observations are produced according to the standards of the World Meteorological Organization (WMO) with established quality control protocols. The data are provided under the GNU General Public License version 3 (GPL-3.0), allowing for free use, modification, and distribution. Table 4 shows the descriptors of the weather data fetched.

The *Present weather (auto)* field employs the WMO Code Table 4677,[25] which encodes current atmospheric conditions as integer values ranging from 0 to 99. The code structure follows a hierarchical organization: values 00–19 indicate no precipitation with various visibility conditions (haze, mist, dust); 20–29 represent recent precipitation events; 30–39 encode dust storms, sandstorms, or blowing snow; 40–49 indicate fog conditions; and 50–99 describe active precipitation at the observation time, including drizzle (50–59), rain (60–69), snow (70–79), and showery or thunderstorm-related

[4]Query: fmi::observations::weather::multipointcoverage

precipitation (80–99). This standardized encoding enables a consistent machine-readable representation of complex weather phenomena across international meteorological networks.

Table 5 details sample data from an Environmental Meteorological Station (EMS) located at Oulu Vihreäsaari satama [5] on December 1st, 2024. The Table presents all 13 weather measurements as columns. However, data for precipitation amount, precipitation intensity, and snow depth are unavailable. This absence is expected and occurs because not all EMS units measure the complete set of weather parameters. Another important detail about the data concerns the measurement time intervals, as this particular EMS records weather conditions every 10 minutes, while some stations in Finland operate at 1-minute intervals.

Weather Data Overview

This study utilizes data from 209 operational weather stations across Finland during 2018-2024. These stations are configured to collect meteorological observations at two distinct measurement intervals: high-frequency 1-minute resolution and standard 10-minute resolution measurements. The distribution reveals that 164 stations (78.47%) operate with 10-minute measurement intervals, representing the predominant temporal resolution in the FMI's observation network, while 45 stations (21.53%) provide high-resolution 1-minute measurements. This temporal configuration reflects the strategic deployment of weather monitoring infrastructure, where the majority of stations employ the World Meteorological Organization's standard 10-minute averaging period for synoptic observations,[26] whereas high-frequency 1-minute stations are typically positioned at locations requiring detailed temporal resolution for specialized applications such as aviation meteorology.[27]

The instrumental capabilities and measurement coverage of the 209 environmental meteorological stations, as detailed in Table 6, reveal a stratified distribution of meteorological parameters. Near-universal coverage (98.60%) is observed for fundamental thermodynamic variables such as air temperature, relative humidity, and dew-point temperature, representing measurements essential for weather monitoring. Wind-related parameters demonstrate moderately high coverage at 79.40%, while atmospheric pressure measurements exhibit 73.20% coverage. In contrast, hydrometeorological parameters show substantially lower coverage: precipitation measurements (56.90%), snow depth (53.60%), cloud amount (53.10%), and horizontal visibility (52.60%) are recorded at approximately half of all stations. This uneven coverage

Table 4. Weather data column descriptions.

Column Name	Unit
timestamp	ISO 8601
station_name	string
Air temperature	°C
Wind speed	m/s
Gust speed	m/s
Wind direction	° (degrees)
Relative humidity	%
Dew-point temperature	°C
Precipitation amount	mm
Precipitation intensity	mm/h
Snow depth	cm
Pressure (mean sea level, msl)	hPa
Horizontal visibility	m
Cloud amount	oktas (0-9)
Present weather (auto)	WMO code

[5]Station Coordinates. Lat: 65.006370, Long: 25.393250

Table 5. **Weather observations: Oulu Vihreäsaari satama EMS.**

Time (UTC +2)	Air Temp. (°C)	Wind Speed (m/s)	Gust Speed (m/s)	Wind Dir. (°)	Rel. Humid. (%)	Dew-point Temp. (°C)	Precip. Amount (mm)	Precip. Intens. (mm/h)	Snow Depth (cm)	Pressure (msl) (hPa)	Horiz. Visib. (m)	Cloud Amount (oktas)	Present Weather (code)
00:00	−1.4	1.1	1.4	28	97	−1.8	–	–	–	1007.1	10943	–	71
00:10	−1.3	0.8	1.6	14	96	−1.7	–	–	–	1007.3	17765	–	67
00:20	−1.2	0.8	1.2	10	96	−1.7	–	–	–	1007.4	20000	–	24
00:30	−1.2	0.3	0.7	325	96	−1.8	–	–	–	1007.4	20000	–	24
00:40	−1.2	0.3	0.6	9	96	−1.7	–	–	–	1007.4	20000	–	24
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10:09	0.6	4.4	5.9	153	99	0.4	–	–	–	1006.4	1867	–	10
10:11	0.6	4.4	5.9	153	100	0.6	–	–	–	1006.4	1706	–	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
23:57	6.3	12.0	15.7	221	83	3.6	–	–	–	995.8	20000	–	0
23:58	7.0	12.0	15.7	221	85	4.7	–	–	–	995.8	20000	–	0
23:59	6.7	12.1	15.7	221	82	3.8	–	–	–	995.8	20000	–	0

Open Research Europe 2026, 6:324 Last updated: 01 OCT 2026

Table 6. Weather feature measurement coverage.

Weather Feature	Stations Measuring
	(Coverage %)
Air temperature	206 (98.60%)
Relative humidity	206 (98.60%)
Dew-point temperature	206 (98.60%)
Wind speed	166 (79.40%)
Gust speed	166 (79.40%)
Wind direction	166 (79.40%)
Pressure (msl)	153 (73.20%)
Precipitation amount	119 (56.90%)
Precipitation intensity	119 (56.90%)
Snow depth	112 (53.60%)
Cloud amount	111 (53.10%)
Horizontal visibility	110 (52.60%)
EMS with no measurements	003 (01.43%)
TOTAL	**209 (100%)**

across meteorological parameters substantially increases the complexity of data analysis and preparation, necessitating systematic handling strategies for missing values in subsequent preprocessing steps.

Derived weather features

Beyond the instantaneous observations described above, two families of derived weather features were engineered directly on the meteorological dataset, prior to its spatial–temporal merging with the train records (Section 3): *rolling-window aggregations* and *weather scenario indicators.* Computing these features at the weather-station level, before merging, ensures that each one summarizes the genuine local weather history recorded at the measuring station, independent of the subsequent train-matching process.

The rolling-window aggregations, summarized in Table 7, capture the recent statistical behavior of the weather. Seven variables (air temperature, wind speed, relative humidity, precipitation intensity, snow depth, pressure, and horizontal visibility) are aggregated as minimum, maximum, and mean over 12 h, 24 h, and 72 h windows, while precipitation amount is aggregated as mean and cumulative total over the same windows, yielding 69 additional columns.

Table 7. Rolling-window aggregated weather features. Statistics are computed over 12 h, 24 h, and 72 h look-back windows. Gust speed, wind direction, and dew-point temperature receive no rolling aggregation.

Base variable	Statistics	Windows	Cols
Air temperature	min, max, mean	12/24/72 h	9
Wind speed	min, max, mean	12/24/72 h	9
Relative humidity	min, max, mean	12/24/72 h	9
Precipitation intensity	min, max, mean	12/24/72 h	9
Snow depth	min, max, mean	12/24/72 h	9
Pressure (msl)	min, max, mean	12/24/72 h	9
Horizontal visibility	min, max, mean	12/24/72 h	9
Precipitation amount	mean, cumulative	12/24/72 h	6
TOTAL			**69**

Table 8. Weather scenario indicator features. Each scenario is provided as a binary flag at four temporal resolutions: instantaneous and rolling 12 h/24 h/72 h windows, with the flag set if the scenario occurs at least once within the window. This yields $10 \times 4 = 40$ columns.

Scenario	Adverse-weather condition flagged
Normal_Clear	Baseline benign conditions, no adverse weather
Blizzard	Heavy snowfall with strong winds and low visibility
Heavy_Snow	Intense snowfall/rapid snow accumulation
Extreme_Cold	Severely low air temperature
Heavy_Rain	High-intensity rainfall
Freezing_Rain	Rain onto sub-zero surfaces causing icing
Black_Ice	Near-freezing conditions favouring invisible rail ice
Dense_Fog	Strongly reduced horizontal visibility
High_Winds	Strong sustained winds and gusts
Extreme_Heat	Unusually high air temperature

The weather scenario indicators, detailed in Table 8, encode ten named adverse-weather conditions (e.g., *Blizzard*, *Freezing_Rain*, *Black_Ice*) as binary flags derived through threshold rules over the measured meteorological variables. Each scenario is evaluated at four temporal resolutions: instantaneously and over rolling 12 h, 24 h, and 72 h look-back windows, where a window-level flag is set if the scenario occurs at least once within that window. This yields $10 \times 4 = 40$ columns. Together, these engineered families allow models to distinguish, for instance, a brief cold snap from a sustained sub-zero spell, a distinction that instantaneous readings alone cannot express.

Validation and quality

Ensuring data integrity is essential when combining heterogeneous data sources, as inconsistencies in temporal alignment, spatial matching, or missing values can propagate errors through subsequent analyses and compromise model performance. The proposed dataset integrates FMI meteorological data with Digitraffic railway traffic data, specifically designed and preprocessed to facilitate machine learning applications for railway operations predictions, particularly train delay forecasting.

To the best of our knowledge, this represents the first comprehensive dataset that systematically combines train schedules with corresponding real-time weather conditions along railway routes in Finland. The methodology for constructing this integrated dataset is illustrated in Figure 4, which outlines the sequential steps from data acquisition to final dataset assembly.

Merging dataset

The integration of weather and train datasets involves a two-stage process combining spatial and temporal alignment:

1. *Spatial matching:* Each train section was associated with its corresponding EMS by identifying the geographically closest station using GPS coordinates. The distance between each train station and candidate EMS was computed using the Haversine formula,[28] which accounts for the spherical geometry of Earth:

$$d = 2r \arcsin\left(\left[\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)\right]^{1/2}\right), \tag{3}$$

where d is the great-circle distance between two points, r is Earth's radius (approximately 6,371 km), ϕ_1 and ϕ_2 are the latitudes, and λ_1 and λ_2 are the longitudes of the two points in radians. The GPS coordinates metadata for train stations and EMS were obtained from their respective open databases.

2. *Temporal alignment:* A left join operation was performed with the train dataset as a base. For each train record, meteorological features were retrieved from its assigned EMS at the corresponding timestamp, preserving all train dataset records. For records lacking exact temporal correspondence, nearest-neighbor temporal matching was applied within a predefined tolerance window to ensure temporal relevance while maintaining data integrity.

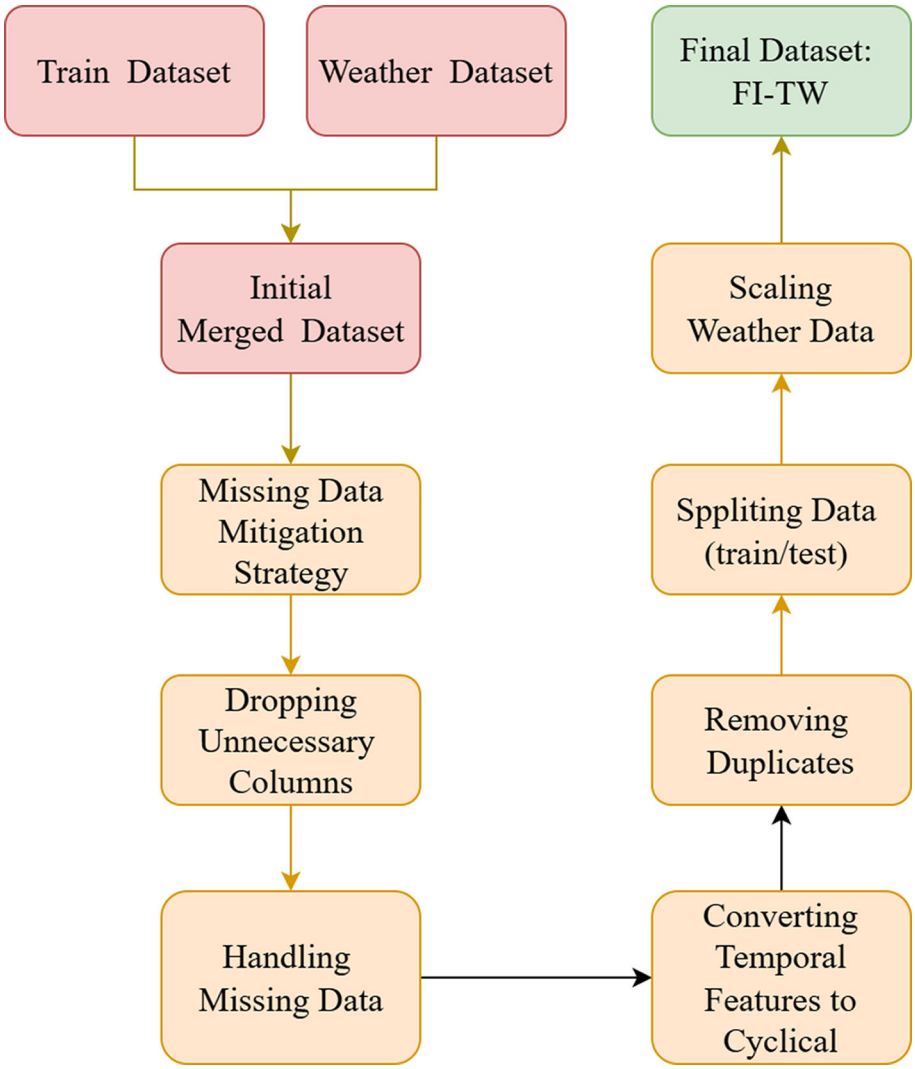


Figure 4. Flowchart of the processing steps.

This two-stage approach ensured that each train operation was matched with the most spatially and temporally relevant meteorological conditions, maximizing the reliability of the integrated dataset while preventing data loss from the primary train records. Figure 5 shows each train station and the closest EMS where weather data was fetched.

Missing data weather mitigation strategy

As discussed in Section 2.2, the environmental meteorological infrastructure exhibits heterogeneity in sensor deployment, in which not all EMS are equipped to measure the complete suite of meteorological parameters. To mitigate the impact of missing weather features in the merged dataset, a spatial fallback algorithm was implemented.

The algorithm operates as follows:

1. *Feature verification:* For each train record matched to its nearest EMS, the availability of the required meteorological features is verified.

2. *Radial search:* When a specific weather parameter is unavailable at the nearest EMS, the algorithm initiates a radial search for alternative EMS units within a 50 km radius that possess the missing measurement capability.

3. *Data retrieval:* If one or more alternative stations are identified, the weather data from the nearest available station measuring the required parameter is retrieved and incorporated into the dataset.

4. *Missing value designation:* If no stations within the search radius provide the missing measurement, the corresponding feature value is designated as missing in the dataset.

This approach represents a pragmatic balance between spatial representativeness and data completeness. The 50 km radius threshold ensures that retrieved meteorological data remain reasonably representative of conditions at the train section location while maximizing feature availability. The strategy acknowledges that localized weather measurements

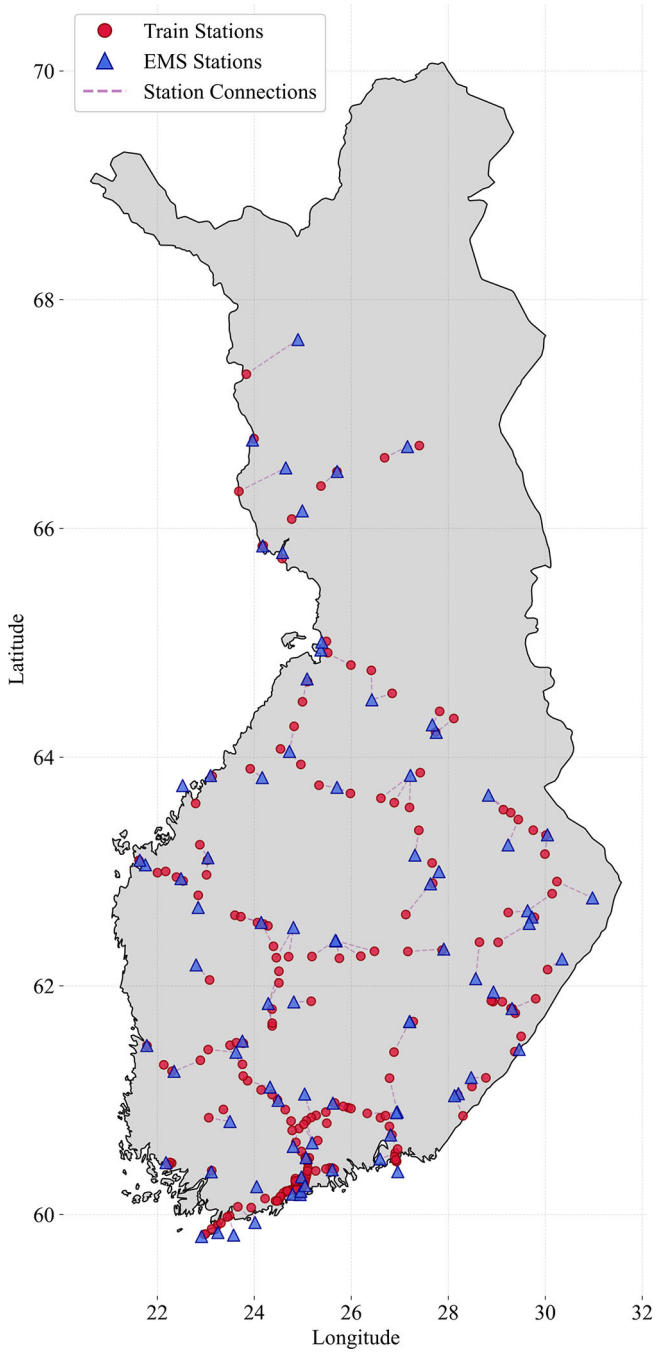


Figure 5. Spatial pairing of train stations with their nearest EMS across the Finnish railway network. Red circles denote train stations, blue triangles denote EMS, and dashed lines connect each train station to the EMS assigned to it by the Haversine nearest-neighbor criterion. Coordinates are given in decimal degrees (WGS84). Base map made with Natural Earth; free vector and raster map data available at naturalearthdata.com.

from a more distant station are preferable to the systematic absence of critical meteorological variables, particularly for regional-scale weather phenomena that exhibit spatial continuity within this distance range.

Dropping unnecessary columns

Following dataset merging, a feature selection procedure was performed to remove columns that provided negligible informational content for the predictive task. Specifically, columns that exhibit one or more of the following characteristics were removed: (1) identifier fields with no predictive value, (2) duplicate features that contain identical information, and (3) administrative metadata not related to the underlying phenomena. This reduction in dimensionality improved both computational efficiency and model interpretability by focusing analysis on domain-relevant features.

Handling missing data

Missing values were addressed through a hierarchical four-strategy approach:

1. *List-wise Deletion:* Applied when (1) timestamps were missing, as temporal integrity is essential for time-series analysis; (2) target features (delay indicators) were missing, since imputing targets introduces training bias (discussed in Section 4); and (3) all weather features were simultaneously missing, as these observations lack predictive context.

2. *Weather Feature-wise Deletion:* Weather columns exceeding 70% missingness were removed, as systematic sensor failures create sparse features that contribute more noise than signal. Table 9 shows weather features and their missing data percentages after applying the weather mitigation strategy.

3. *Constant Value Imputation:* Boolean features (trainStopping, commercialStop) with missing values were set to False (encoded as 0), based on domain knowledge that absent flags typically indicate feature negation in transportation datasets.

4. *Month-Specific Median Imputation:* Weather features passing the sparsity threshold were imputed using monthly medians to account for seasonal variability. The median was chosen for its robustness against sensor outliers.

Converting temporal features to cyclical approach

Temporal features such as hours, months, and days of the week exhibit inherent cyclical properties that are not adequately captured by their raw numerical representations. When these features are encoded as linear integers, machine learning models encounter the time wraparound problem,[29] where values at cycle boundaries are treated as maximally distant despite their temporal proximity. For instance, 23:00 (11 PM) and 00:00 (midnight) are separated by only one hour

Table 9. Missing data for weather features (2018-2024).

Feature	Missing Count	Missing (%)
Precipitation amount	33,506,086	86.91
Cloud amount	13,048,473	33.85
Wind direction	13,000,331	33.72
Pressure (msl)	6,976,767	18.10
Dew-point temperature	3,219,117	8.35
Air temperature	3,182,196	8.25
Relative humidity	3,159,488	8.20
Snow depth	2,378,448	6.17
Precipitation intensity	2,320,256	6.02
Horizontal visibility	2,251,748	5.84
Wind speed	2,153,097	5.59
Gust speed	2,153,067	5.58
Total observations	**38,551,161**	—

temporally, yet their numerical representation suggests a distance of 23 units, potentially misleading gradient-based and distance-based learning algorithms.

The temporal values of hour, month and day_of_week are transformed into $(\text{hour_sin}, \text{hour_cos})$, $(\text{month_sin}, \text{month_cos})$, and $(\text{day_week_sin}, \text{day_week_cos})$, respectively. These encodings effectively communicate to the model that these values reside at the terminal positions of their respective cycles, facilitating accurate learning of wraparound patterns.

Three distinct transformation stages were implemented to encode temporal features using sine-cosine encoding at different granularities. We employed the standard cyclical encoding formula:

$$\text{encoded_feature} = \sin\left(\frac{2\pi t}{P}\right), \quad \cos\left(\frac{2\pi t}{P}\right), \tag{4}$$

where t represents the temporal value and P is the corresponding period. Specifically, the hour feature with $t \in \{0, 1, \ldots, 23\}$ was encoded with $P = 24$; the month feature with $t \in \{1, 2, \ldots, 12\}$ with $P = 12$; and the day-of-week feature with $t \in \{1, 2, \ldots, 7\}$ (where 1 = Monday) with $P = 7$. The original hour column was subsequently removed, as the sine-cosine representation suffices for all model architectures. In contrast, the original month and day-of-week columns were retained alongside their encoded variants to support diverse model requirements.

Additional data integrity steps

Removing duplicates

Duplicate records were identified and removed to ensure that the dataset did not contain redundant observations. Duplicates were detected across all feature dimensions using an exact matching approach. Duplicates represented 23.24% of the dataset and were completely removed.

Splitting data

The dataset was partitioned into training and test sets using an 80/20 split, allocating 80% of observations for model training and 20% for evaluation. Crucially, this step was performed before any scaling operations to prevent data leakage. This division ensures that model performance is assessed on previously unseen data, providing an unbiased estimate of generalization capability.

Scaling weather data

Applied only after splitting data to prevent data leakage, scaling parameters were computed solely from the training partition only and subsequently applied to both training and test sets. This ensures that no statistical information from the test set influences the transformation.

To ensure comparability between meteorological variables with disparate measurement scales and units, a robust scaling transformation was applied to the weather features. Unlike standard normalization, which employs mean and standard deviation, this study utilized RobustScaler from the scikit-learn library. This approach is advantageous for weather data because meteorological measurements are susceptible to outliers and extreme values that can distort conventional scaling methods. By leveraging quartile-based measures, RobustScaler minimizes the influence of such anomalies, ensuring that scaling remains representative of the typical data distribution.

The RobustScaler normalization can be defined as

$$x_{\text{scaled}} = \frac{x - Q_1(X_{\text{train}})}{Q_3(X_{\text{train}}) - Q_1(X_{\text{train}})}, \tag{5}$$

where Q_1 denotes the first quartile (25th percentile) and Q_3 denotes the third quartile (75th percentile), both calculated from the training partition. Quartiles are obtained by sorting the training observations and identifying the values that divide the distribution into four equal parts, such that 25% of observations fall below Q_1 and 75% fall below Q_3. The denominator— of the equation, the interquartile range, represents the range spanning the central 50% of the training data, a statistic fundamentally resistant to extreme values.

Table 10. **Final Dataset Feature Descriptors.**

No.	Feature Name	Description	Derived?
Target Features			
1	differenceInMinutes	Delay in minutes (actual - scheduled)	No
2	differenceInMinutes_offset	Delay in minutes (delay offset removed from 1st station only)	Yes
3	differenceInMinutes_eachStation_offset	Delay in minutes (delay offset removed from all stations)	Yes
4	trainDelayed	Boolean indicating if train is delayed	Yes
5	cancelled	Boolean indicating if train was cancelled	No
Operational Features			
6	trainStopping	Boolean indicating if train stopped at station	No
7	commercialStop	Boolean indicating if stop is commercial	No
8	hour	Hour of the day (0-23)	No
9	hour_sin	Sine component of cyclical hour encoding	Yes
10	hour_cos	Cosine component of cyclical hour encoding	Yes
11	month	Month of the year (1-12)	No
12	month_sin	Sine component of cyclical month encoding	Yes
13	month_cos	Cosine component of cyclical month encoding	Yes
14	day_of_week	Day of the week (1-7)	No
15	day_week_sin	Sine component of cyclical day of week encoding	Yes
16	day_week_cos	Cosine component of cyclical day of week encoding	Yes
17	day_of_month	Day of the month (1-31)	No
18	causes_related_to_weather	Boolean indicating if the recorded delay cause is weather-related	Yes
Weather Features (base)			
19	Air temperature	Air temperature (°C)	No
20	Wind speed	Wind speed (m/s)	No
21	Gust speed	Wind gust speed (m/s)	No
22	Wind direction	Wind direction (degrees)	No
23	Relative humidity	Relative humidity (%)	No
24	Dew-point temperature	Dew-point temperature (°C)	No
25	Precipitation intensity	Precipitation intensity (mm/h)	No
26	Snow depth	Snow depth (cm)	No
27	Pressure (msl)	Mean sea level pressure (hPa)	No
28	Horizontal visibility	Horizontal visibility (m)	No
29	Cloud amount	Cloud amount (oktas, 0-9)	No
Weather Features (engineered)			
30–69	weather_scenario_∗	adverse-weather scenario indicators at 4 temporal windows (Table 8)	Yes
70–138	*rolling aggregations*	Rolling-window statistics of weather variables over 12/24/72 h (Table 7)	Yes

Proposed final dataset

Table 10 presents a comprehensive overview of the feature set comprising our final dataset for the train delay prediction in Finland. The dataset encompasses 138 distinct features organized into three primary categories: target features, operational features, and weather features. The weather features are further divided into eleven base measurements and two families of engineered variables: 69 rolling-window aggregations (Table 7) and 40 weather scenario indicators (Table 8). Approximately 86.2% of these features are derived columns, generated through feature engineering techniques to enhance the predictive capability of our models.

The **operational features** capture the temporal patterns and characteristics of the train service. Two boolean variables, *trainStopping* and *commercialStop*, distinguish stopping behaviors. Recognizing the cyclical nature of temporal data, we applied trigonometric encoding to month, hour, and day of week, generating sine and cosine components that preserve circular relationships. Importantly, the primary temporal columns (*hour*, *month*, *day_of_week*, *day_of_month*) are retained in the final dataset for generic use purposes, as tree-based algorithms may better utilize these raw features, while cyclical encodings are more suitable for numeric/regression-based techniques. An additional boolean feature, *causes_related_to_weather*, is derived from the Digitraffic *causes* array and flags whether the officially recorded cause of a delay is weather-related, providing a label for weather-impact studies.

The **weather features** comprise eleven *base* meteorological measurements from the FMI (Features 19–29, Table 10), reporting the instantaneous conditions at each train section, together with the two families of engineered weather features introduced in Section 2.2.2: 69 rolling-window aggregations (Table 7) and 40 weather scenario indicators (Table 8). Collectively, these features enable our models to account for weather-related disruptions, which may constitute a factor in railway delays.

The **target features** constitute the dependent variables and delay-related indicators that our models aim to predict or utilize for classification tasks:

- *differenceInMinutes*: The primary target variable representing the raw delay calculated as the difference between actual and scheduled arrival times.
- *differenceInMinutes_offset*: An offset feature representing the delay in minutes excluding the initial delay accumulated at the first station, thus isolating delays that develop during the journey.
- *differenceInMinutes_eachStation_offset*: An offset feature that isolates the delay contribution specific to each individual station by removing propagated delays from previous stops.

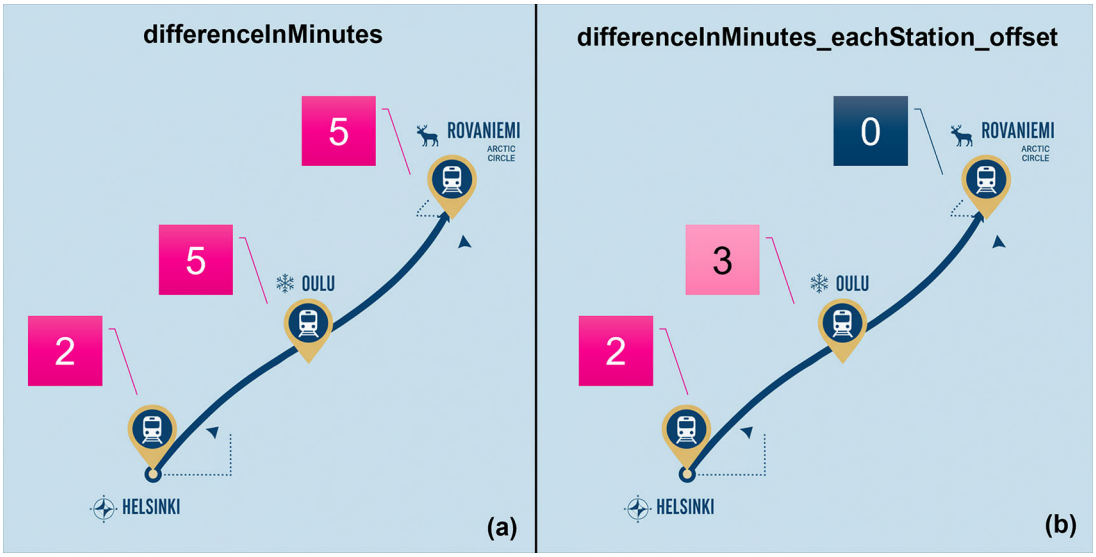


Figure 6. Delay target variables comparison: (a) *differenceInMinutes*, (b) *differenceInMinutes_eachStation_offset*.

To clarify the construction of *differenceInMinutes_eachStation_offset*, consider the example illustrated in Figure 6. This figure depicts a hypothetical train route from Helsinki to Rovaniemi via Oulu. The number above each station represents the delay in minutes at that station. Figure 6a shows that the train departed Helsinki 2 minutes late, arrived at Oulu 5 minutes late, and reached Rovaniemi 5 minutes late. However, these values represent the total accumulated delay at each point, not the delay introduced on each segment. Figure 6b illustrates the delay offset for each station segment after removing the inherited delay from previous stations. If the train departed Helsinki 2 minutes late and arrived at Oulu 5 minutes late, the delay introduced specifically in the Helsinki–Oulu segment is 3 minutes (5-2 = 3), not 5 minutes. Similarly, since the train arrived at both Oulu and Rovaniemi with the same 5-minute delay, the Oulu–Rovaniemi segment did not introduce additional delay (5-5 = 0). The *differenceInMinutes_eachStation_offset* target feature thus represents the incremental delay added to each route segment, isolating the delay contribution between consecutive stations rather than the cumulative delay from the origin.

Additionally, two boolean indicators facilitate binary classification tasks and enable analysis of severe disruptions in rail operations:

- *trainDelayed*: A boolean feature containing True/False information using a 5-minute threshold to classify whether a train is delayed.

- *cancelled*: A boolean feature that contains raw information that indicates whether a train was canceled by the operator.

Dataset usage and configuration

The dataset has been designed with maximum flexibility in mind, allowing researchers and practitioners to adapt it to various predictive modeling scenarios. Users have complete autonomy to select relevant weather features according to their specific research questions and to choose appropriate target variables based on their analytical objectives. This modular design enables the dataset to serve multiple use cases, from weather-impact studies to operational delay prediction systems.

A correlation analysis was conducted to identify potential feature redundancies and inform feature selection strategies. The analysis revealed several strong correlations between weather variables: gust speed and wind speed exhibit very high correlation (0.946), dew-point temperature shows strong correlations with air temperature (0.898) and relative humidity (0.595), while snow depth demonstrates moderate correlations with dew-point temperature (0.575) and horizontal visibility (0.571). The heatmap enables identification of strongly correlated features, detection of multicollinearity issues, and insights into which weather conditions most significantly impact train operations. These highly correlated features could be candidates for removal during model training to reduce multicollinearity and improve model efficiency, though they remain in the dataset to allow flexibility in feature selection strategies for different modeling approaches.

The dataset includes both numeric and binary target variables to accommodate different modeling approaches and research objectives. Numeric targets include delay measurements (Features 1–3, Table 10), while binary targets comprise operational status indicators (Features 4–5, Table 10).

Figure 7 compares delay distributions under two different formulations for Oulu asema (Oulu station), considering all long-distance trains that pass through Oulu during the seven-year period. Figure 7a shows differenceInMinutes, which measures cumulative delay from the origin. Here, 72.6% of observations show some delay (zero threshold), and 27.1% exceed five minutes. Figure 7b shows differenceInMinutes_eachStation_offset, which isolates the delay generated at each station by removing propagation effects. This formulation yields 51% delayed observations (zero threshold) and 14.2% exceeding five minutes.

The difference reflects how delays accumulate: most observed delays result from propagation rather than station-specific causes. The choice of target variable depends on the modeling objective, as differenceInMinutes suits end-to-end delay prediction for passenger information systems, and differenceInMinutes_eachStation_offset identifies local operational issues and weather impacts at specific stations. The class imbalance also differs significantly between formulations, requiring different handling strategies for classification tasks.

Preliminary baseline experiment

To illustrate the dataset's utility for ML applications, we conducted a preliminary baseline prediction experiment using XGBoost regression, a gradient boosting algorithm widely employed for predictive modeling tasks across diverse

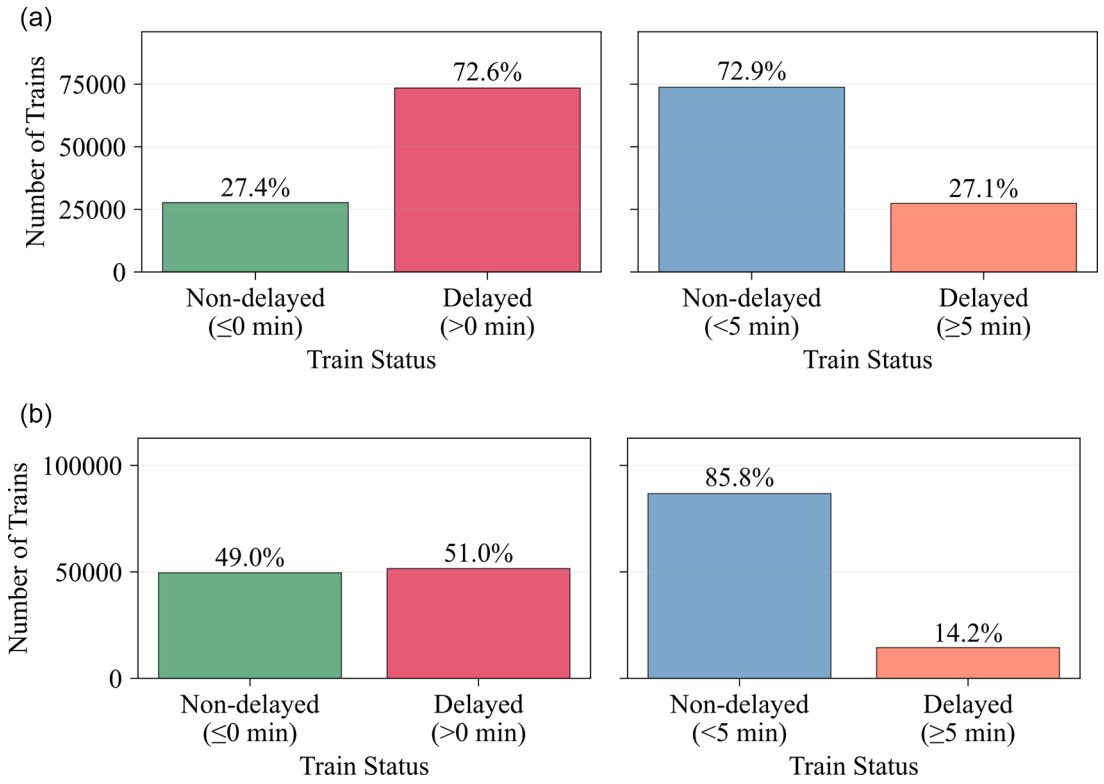


Figure 7. Target feature (delay in minutes) distributions for Oulu asema. (a) Delay distribution with propagation delay (*differenceInMinutes* target variable). (b) Delay distribution per station, without delay propagation (*differenceInMinutes_eachStation_offset* target variable).

domains.[30] This experiment is intended as an illustrative demonstration of how the dataset can be used, rather than an attempt to obtain an optimized or state-of-the-art delay-prediction model.

Experimental setup

The experiment focused on long-distance trains passing through Oulu asema during the 2018–2024 period, comprising 101,146 observations. The target variable was differenceInMinutes_eachStation_offset, which isolates station-specific delay contributions by removing propagated delays from previous stops. This formulation enables the model to learn local delay patterns influenced by weather conditions rather than cumulative delay propagation.

The feature set comprised 18 variables spanning operational and meteorological domains: trainStopping, cyclical temporal encodings (month_sin, month_cos, hour_sin, hour_cos, day_of_week, day_week_sin, day_week_cos), and ten weather features (air temperature, wind speed, gust speed, wind direction, relative humidity, dew-point temperature, precipitation intensity, snow depth, pressure, and horizontal visibility).

Hyperparameter optimization was performed using randomized search with 50 iterations and 5-fold cross-validation on the training partition. The dataset was split 80/20 for training and testing, with weather features scaled using RobustScaler parameters derived exclusively from the training set, as detailed in Section 3.6.

Result and discussion

Figure 8 presents the Mean Absolute Error (MAE) performance of XGBoost models across three delay prediction targets over 50 random search iterations for hyperparameter optimization.

The station-specific delay target (differenceInMinutes_eachStation_offset) demonstrates superior predictive performance, achieving an MAE of 2.73 minutes on the test set for Oulu asema. This represents a substantial improvement over cumulative delay prediction targets: differenceInMinutes (4.21 minutes MAE) and differenceInMinutes_offset (4.81 minutes MAE). The performance gap illustrates that station-specific delays, which isolate local operational

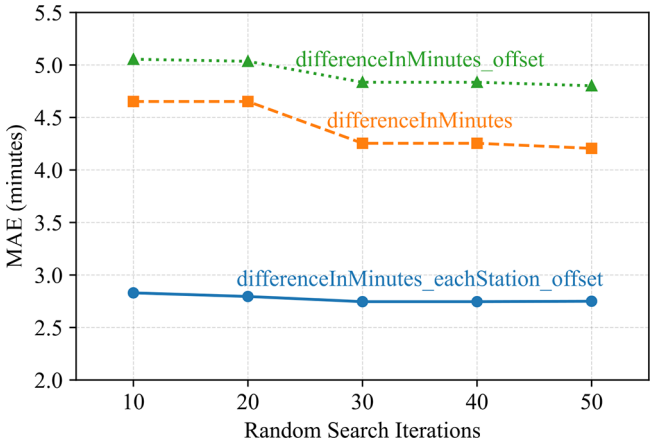


Figure 8. MAE comparison of XGBoost models across random search iterations for three delay prediction targets.

conditions, are inherently more predictable than accumulated delays that propagate through the network. All models exhibit convergence by 30 iterations, with minimal performance variation thereafter.

Direct comparison with existing literature is challenging due to differences in railway systems, prediction targets, and evaluation contexts. Huang et al.[13] employed different algorithms on Chinese high-speed railways; for long-distance trains, Random Forest and Support Vector Machine achieved MAE values of 2.16 minutes, while their proposed FCLL-Net deep learning architecture achieved 1.87 minutes (one-step predictor). Li et al.[31] applied Random Forest and XGBoost to Dutch conventional railways, achieving MAE values of 1.71 and 1.72 minutes for delays exceeding 3 minutes. However, both the Chinese high-speed and Dutch conventional systems operate predominantly on double-track infrastructure with milder climatic conditions. The Finnish context introduces unique challenges, including extreme seasonal weather variations (temperatures ranging from -40°C to $+30$°C) and a predominantly single-track network with different operational characteristics.

As a preliminary demonstration, this experiment is deliberately limited in scope. It considers a single station (Oulu asema) with a single learning algorithm, reports a single error metric, and uses only the base weather and operational features. The engineered weather-scenario indicators and rolling-window aggregations, a systematic assessment of the incremental predictive value of weather (e.g., through with- and without-weather ablations), and a network-wide evaluation across multiple stations are all left for future work. The intent here is to establish that FI-TW supports end-to-end machine-learning workflows and to provide a reproducible reference point, rather than to deliver an optimized delay-prediction model.

Conclusions

This paper presented FI-TW, the Finland Integrated Train–Weather dataset, which is, to the best of our knowledge, the first publicly available resource integrating Finnish railway operational data with synchronized meteorological observations from 2018-2024. The dataset combines train schedules and delay records from Digitraffic with weather measurements from 209 FMI stations, covering approximately 38.5 million observations across Finland's 5,915-kilometer rail network. Exploratory analysis revealed distinct seasonal patterns, with winter months exhibiting delay rates exceeding 25% compared to below 20% in summer, and geographic clustering of high-delay corridors in central and northern Finland. The dataset offers multiple target variable formulations, cumulative delays for passenger information systems and station-specific offset delays for isolating local operational impacts, enabling diverse modeling approaches from binary classification to regression-based prediction. A baseline XGBoost experiment achieved MAE of 2.73 minutes for station-specific delay prediction, demonstrating the dataset's utility for machine learning applications.

Several research directions emerge from this work. The temporal structure of train journeys, where delays propagate through successive stations, suggests that sequence modeling architectures such as Long Short-Term Memory networks and Transformers may capture delay dynamics more effectively than traditional approaches. Graph Neural Networks could exploit network-wide dependencies between stations and routes. Extending the dataset with streaming data pipelines and weather forecasts would enable real-time prediction systems capable of advance warning rather than post-hoc analysis. Finally, causal inference methods could quantify the direct impact of specific meteorological

conditions on railway performance, while incorporating additional data sources such as passenger flows, maintenance records, or rolling stock information could further enhance predictive accuracy.

Ethics and consent

Ethical approval and consent were not required.

Data availability

Underlying data

Zenodo: FI-TW: Finland Integrated Train-Weather Dataset. https://doi.org/10.5281/zenodo.21675622.[32]

This project contains the following underlying data:

- matched_data_flat_YYYY_MM.parquet (96 Apache Parquet files, one per month from January 2018 to December 2025, holding the merged train and weather records that constitute the dataset. Each row is a train event at a track section, with the train identifier, train type and category, the station short code, name, and geographic coordinates, the scheduled and actual times, the delay in minutes in both its cumulative and station-specific offset formulations, the 13 raw meteorological parameters observed at the matched Finnish Meteorological Institute station, and all derived features, namely the cyclical temporal encodings, the weather-scenario indicators, and the multi-scale rolling-window aggregations, giving 138 features in total. The analysis reported in this article uses the files from January 2018 to December 2024).

A mirror of this dataset is also hosted on Kaggle at https://www.kaggle.com/datasets/viniborin/finland-integrated-train-weather-dataset-fi-tw. The Zenodo deposit cited above is the version of record.

Extended data

There is no extended data associated with this article.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC BY 4.0).

References

1. Gkoumas K, dos Santos FLM, Stepniak M, *et al.*: **Rail transport research and innovation in europe: An abessment based on recent european union projects.** *Transportation research procedia.* Elsevier B.V; 2023; pp. 3633–3640.
Publisher Full Text
2. Mukunzi G, Palmqvist CW: **The impact of railway incidents on train delays: A case of the swedish railway network.** Journal of Rail Transport Planning and Management. Jun. 2024; **30**: 100445.
Publisher Full Text
3. Monsuur F, Enoch M, Quddus M, *et al.*: **Modelling the impact of rail delays on passenger satisfaction.** Transp. Res. A Policy Pract.. Oct. 2021; **152**: 19–35.
Publisher Full Text
4. Taleb T, *et al.*: *White Paper on 6G Networking.* University of Oulu; 2020. 6G Research Visions 6.
5. Davari N, Veloso B, de Assis Costa G, *et al.*: **A survey on data-driven predictive maintenance for the railway industry.** *Sensors.* Aug. 2021; **21**: 5739.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
6. Pappaterra MJ, Flammini F, Vittorini V, *et al.*: **A systematic review of artificial intelligence public datasets for railway applications.** *Infrastructures.* Oct. 2021; **6**.
Publisher Full Text
7. Zendel O, Murschitz M, Zeilinger M, *et al.*: **RailSem19: A dataset for semantic rail scene understanding.** *IEEE computer society conference on computer vision and pattern recognition workshops.* IEEE Computer Society; Jun. 2019; pp. 1221–1229.
Publisher Full Text
8. Cordts M, *et al.*: **The Cityscapes Dataset.**Accessed: Nov. 05, 2025.
Reference Source
9. Ansari U: **Indian Metro Data: Prediction of the Future Traffic.** Accessed: Nov. 05, 2025.
Reference Source
10. Zhang D, *et al.*: **A high-speed railway network dataset from train operation records and weather data.** *Sci. Data.* 2022; **9**(1): 244.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
11. Wu J, *et al.*: **A railway network dataset incorporating multi-type train operation records and train scheduling.** *Sci. Data.* 2026; **13**(1): 77.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
12. Elliker C, Bail ML, Mantoux C, *et al.*: **RIDE: An open dataset and benchmark for train delay prediction.** *arXiv preprint arXiv: 2606.05070.* 2026.
Publisher Full Text
13. Huang P, *et al.*: **Modeling train operation as sequences: A study of delay prediction with operation and weather data.** *Transp. Res. Part E Logist. Transp. Rev.* Sep. 2020; **141**: 102022.
Publisher Full Text
14. Huang P, Li Z, Wen C, *et al.*: **Modeling train timetables as images: A cost-sensitive deep learning framework for delay propagation pattern recognition.** *Expert Syst. Appl.* Sep. 2021; **177**: 114996.
Publisher Full Text
15. Sajan GV, Kumar P: **Forecasting and analysis of train delays and impact of weather data using machine learning.** *2021 12th international conference on computing communication and networking technologies (ICCCNT).* IEEE; 2021.
Publisher Full Text
16. Oneto L, *et al.*: **Train delay prediction systems: A big data analytics perspective.** *Big Data Research.* 2018; **11**: 54–64.
Publisher Full Text
17. Zakeri G, Olsson NOE: **Investigating the effect of weather on punctuality of Norwegian railways: A case study of the Nordland line.** *Journal of Modern Transportation.* 2018; **26**(4): 255–267.
Publisher Full Text

18. Soleimani-Chamkhorami K, Karbalaie A, Kasraei A, *et al.*: **Identifying climate-related failures in railway infrastructure using machine learning.** *Transp. Res. Part D: Transp. Environ.* 2024; **135**: 104371.
Publisher Full Text

19. Finnish Transport Infrastructure Agency: **Railway Network.** Accessed: Nov. 05, 2025.
Reference Source

20. Borin VP, de Souza Sant'Ana JM, Raheel U, *et al.*: **FI-TW: An open train-weather dataset for railway delay analysis in Finland.** *arXiv.* 2026.
Publisher Full Text

21. VR Group: **VR long-distance traffic in finland had an all-time year – 15.3 million train journeys and excellent customer satisfaction.**Accessed: Sep. 15, 2025.
Reference Source

22. Finnish Transport Infrastructure Agency: **Railway Statistics: Punctuality in Long Distance and Commuter Traffic.**Accessed: Sep. 15, 2025.
Reference Source

23. Lotfi A, Virk MS: **Railway operations in icing conditions: A review of issues and mitigation methods.** *Public Transport.* Oct. 2023; **15**: 747–765.
Publisher Full Text

24. Lewis R, *et al.*: **Leaves on the line: Characterising leaf based low adhesion on railway rails.** *Tribol. Int.* Jul. 2023; **185**: 108529.
Publisher Full Text

25. World Meteorological Organization: **WMO code table 4677: Present weather reported from a manned weather station.**
Reference Source

26. World Meteorological Organization: *Guide to meteorological instruments and methods of observation.* World Meteorological Organization;Geneva, Switzerland: 2008. WMO-No. 8.
Reference Source

27. SKYbrary: **Weather observations at aerodromes.**
Reference Source

28. Sinnott RW: **Virtues of the haversine.** *Sky and Telescope.* 1984; **68**(2): 158–159.

29. Cai L, Janowicz K, Mai G, *et al.*: **Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting.** *Trans. GIS.* Jun. 2020; **24**: 736–755.
Publisher Full Text

30. Chen T, Guestrin C: **XGBoost: A scalable tree boosting system.** *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* ACM; 2016; pp. 785–794.
Publisher Full Text

31. Li ZC, Wen C, Hu R, *et al.*: **Near-term train delay prediction in the dutch railways network.** *Int. J. Rail Transp.* 2021; **9**: 520–539.
Publisher Full Text

32. Borin VP, de Souza JM, Sant'Ana UR, *et al.*: **FI-TW: Finland integrated train-weather dataset.** *Zenodo.* 2026.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 23 September 2026

https://doi.org/10.21956/openreseurope.26824.r80612

© **2026 Pineda-Jaramillo J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Juan Pineda-Jaramillo
Universidad Politécnica de Valencia, Valencia, Spain

The article presents FI-TW, an open dataset that couples Digitraffic train records for Finnish long-distance services with observations from Finnish Meteorological Institute (FMI) stations over 2018-2024, describes the spatial-temporal matching and preprocessing pipeline, and illustrates its use with an XGBoost baseline for station-specific delay prediction at Oulu asema. The resource is genuinely useful: the seven-year span, the Nordic setting, the open licences of both sources and the public Zenodo deposit are real strengths, the related work is current, and the writing is clear. Since Open Research Europe publishes data articles irrespective of perceived novelty, my assessment focuses on accuracy, methodological soundness and reproducibility, which is precisely where the article falls short. I downloaded one of the deposited monthly files (December 2024) to check the article against the data, and the released files do not correspond to the dataset described in the text: they have 122 columns rather than 138, they retain raw identifiers and physical units, and they contain neither the weather-scenario indicators nor the cyclical encodings and derived boolean targets listed in Table 10. In addition, several claims in the abstract and conclusions (geographic clustering of high-delay corridors, a correlation heatmap) are not supported by any analysis in the article, key processing parameters (temporal tolerance, scenario thresholds, deduplication key) are missing, the processing code is neither cited nor archived, and the baseline experiment is reported without a naive benchmark, so its 2.73-minute MAE cannot be interpreted. Some of the points below overlap with the first reviewer's report; I keep them because they remain necessary, and I add issues that arise from inspecting the deposit directly. My status is Approved with reservations, which in practice corresponds to a major revision: the article must be brought into agreement with the version of record of the data before it can be approved. Comments tagged [Essential] below are the points that must be addressed to make the article scientifically sound; comments tagged [Recommended] would improve the article but do not condition approval.

1. Background and Related Work
The background is well motivated, and the related work is current and correctly identifies the closest datasets (Zhang et al., Wu et al., RIDE) and the relevant Nordic studies.
• [Essential] The scope of the dataset is not stated where it matters. The deposited files contain

only records with train Category equal to Long-distance (verified on the December 2024 file), and the text itself says that the study focuses on long-distance services, yet the title, abstract, plain language summary and conclusions speak of "railway operations in Finland" and "Finland's 5,915-kilometer rail network", and the keywords include "freight transportation and logistics". Commuter and freight traffic are absent. Please state explicitly, in the abstract and in the Data availability section, that FI-TW covers long-distance passenger services only, remove the freight keyword, and report which train types (IC, S, PYO, etc.) and which share of the network are actually represented.
• [Recommended] The claim of being "the first publicly available dataset integrating operational data with meteorological data" is repeated three times and should be confined to Finland, since the same section acknowledges RIDE and the Chinese and Italian datasets. The distinguishing features (seven-year span, 13 meteorological parameters, sub-Arctic conditions, engineered features) are the real contribution and should be the basis of the positioning. The paragraph on 5G/6G networks and AI is tangential to a dataset article and reads as project-driven; it should be shortened or removed. Figures such as the 245,000 km European network need a citation.

2. Data Sources and Descriptive Analysis
The description of the Digitraffic and FMI services, their licences, rate limits and record structures (Tables 1 to 5) is detailed and will be helpful to users.
• [Essential] Time reference. Tables 2, 3 and 5 label the timestamps as UTC+2, but the deposited files store ISO 8601 timestamps in UTC (suffix Z): the example train of Table 3 (IC 22, Oulu 05:49) appears in the December 2024 file as 05:49Z, which is 07:49 local time. Finland uses UTC+2 in winter and UTC+3 in summer, so a fixed UTC+2 label is wrong for seven months of the year. Please state the reference clock of both sources, how daylight saving time was handled in the temporal matching, and whether the hour, hour_sin and hour_cos features are computed in UTC or in local time, which is what matters for operational patterns (peak hours, night operations).
• [Recommended] Station accounting. The text reports 549 stations split into 340 non-passenger and 209 passenger stations, but the functional classification that follows (454 + 64 + 42) sums to 560. The coincidence between 209 passenger stations and 209 FMI stations should also be double-checked. More useful for users would be the counts that describe the released data: the December 2024 file contains 372 distinct station Short Code values and 84 distinct closest_ems values, so please report how many track sections and how many EMS actually appear in FI-TW over 2018-2024, rather than national totals.
• [Essential] Delay definitions and the unit of analysis. Equation 2 defines a train as delayed when difference In Minutes is greater than zero, Figure 1 uses five minutes or more, train Delayed uses a five-minute threshold, and the official punctuality standard is arrival within five minutes; please adopt one definition and use it consistently. The unit of analysis in Figures 1 to 3 is the train event (an ARRIVAL or DEPARTURE record), not the train, and since arrival and departure records at intermediate stops carry the same delay in about 86% of cases (December 2024 file), every stop is counted twice; the text should say so and should not describe these percentages as the share of trains delayed. Figure 2 is undefined: the y-axis counts days (the four bars sum to 2,556, i.e. seven years), but the text does not explain which daily statistic is assigned to the 5-10, 10-15, 15-20 and 20+ minute categories. Finally, June at 21.8% contradicts the statement that summer delay rates are "below 20%".

3. Dataset Construction and Quality
The two-stage matching, the radial fallback and the hierarchical treatment of missing values are sensible design choices, and the step-by-step presentation is easy to follow.

• [Essential] Temporal and spatial matching parameters are incomplete. The tolerance of the nearest-neighbour temporal matching is not given, nor whether weather is matched to the scheduled or the actual time, how ties and the 1-minute versus 10-minute stations are treated, and whether the rolling-window aggregations (computed at station level before merging) are taken from the fallback EMS when a base variable is retrieved from a different station within the 50 km radius. The nearest-EMS assignment has no maximum distance, so remote sections may be matched to stations far beyond 50 km, and localised phenomena such as precipitation, fog and snow depth are poorly represented by observations 50 km away. Please report the distribution of station-to-EMS distances, the share of records and variables that used the fallback, and add to the released files the matching distance and a per-variable source or fallback flag (the files currently contain the EMS name in closest_ems but no distance or coordinates).
• [Essential] Missing data. Table 9 reports 33.85% missing cloud amount, 33.72% missing wind direction, 18.10% missing pressure and 8.25% missing air temperature over 2018-2024, whereas in the December 2024 file these variables have 0.00% missing values and only precipitation amount (86.3%), snow depth (4.2%) and precipitation intensity (3.8%) contain gaps. Please clarify at which stage Table 9 is computed (before or after the fallback, and whether missingness is concentrated in particular years), and state whether the deposited values are observed or already imputed; if imputed, an indicator column is needed, because monthly median imputation erases precisely the extreme episodes the dataset is meant to capture. Median imputation is also inappropriate for wind direction (a circular variable) and questionable for cloud amount (ordinal oktas). Precipitation amount exceeds the 70% threshold and is accordingly absent from the base features of Table 10, yet Table 7 keeps six rolling aggregations of it; this should be reconciled.
• [Essential] Duplicates. The 23.24% of "exact duplicates" is not a property of the merged data: the December 2024 file has zero exact duplicates and zero duplicates on (trainNumber, departureDate, stationShortCode, type, scheduledTime). The figure can only arise after identifiers and the type column are dropped, at which point the ARRIVAL and DEPARTURE records of intermediate stops (identical scheduledTime and delay in about 80% of pairs) collapse into identical rows. These are distinct events, not duplicates, and removing them discards dwell-time and departure information. Please specify the columns used for duplicate detection, explain the origin of the 23.24%, and either keep the type column or justify the removal.
• [Essential] Train/test split and scaling. A random 80/20 split at row level of events that are strongly dependent (the same train on the same day at consecutive sections, matched to the same EMS at nearly the same time) leaks information and inflates test performance; a chronological or grouped split (by train and departure date) is needed, or the limitation must be stated. The order of the monthly median imputation relative to the split is not stated, and if the medians were computed on the full data the leakage extends to preprocessing. Equation 5 does not describe scikit-learn's RobustScaler with default settings, which centres on the median and scales by the interquartile range; please correct the equation or the description. More fundamentally, splitting, scaling and duplicate removal are experiment-specific choices, not properties of a dataset, and the released files are in fact unscaled and unsplit (the December 2024 file contains physical units), which is the right decision; Figure 4 and the text should therefore separate the construction of FI-TW from the preprocessing used in the baseline experiment.

4. Released Dataset, Data and Software Availability
The choice of Zenodo as version of record, the CC BY 4.0 licence and the monthly Parquet layout are appropriate, and keeping raw identifiers and physical units in the files is the right decision for reuse.
• [Essential] The article does not describe the deposited data, which is the central problem for a

data article. Table 10 lists 138 features; the Zenodo description states "Features: 30"; the December 2024 file has 122 columns. The files retain identifier and metadata columns that the text says were dropped (trainNumber, operator codes, timetable version and acceptance date, causes, trainReady, live Estimate Time, estimate Source, closest ems), contain all 13 raw weather parameters including Present weather and Precipitation amount, and hold 78 rolling aggregations (including nine for cloud amount that Table 7 does not list), but they do not contain the 40 weather-scenario indicators, the cyclical encodings, hour, month, day of week, day of month, train Delayed or causes related to weather. Please rewrite Table 10 so that it documents exactly the columns of the released files (name, type, unit, derivation), update the Zenodo description accordingly and, if the engineered features are part of the intended resource, deposit a new versioned release that includes them or the script that generates them. The Data availability statement should also be corrected: the deposit has 96 files covering 2018-2025 while the article covers 2018-2024, and it should be stated whether the 2025 files went through the same processing; the Zenodo creator list omits U. Raheel; and reference 32 garbles the author names.
• [Essential] The engineered features are not reproducible from the article. The ten weather-scenario indicators (Blizzard, Black_Ice, Freezing Rain, Extreme Cold, etc.) are defined only qualitatively in Table 8; the numerical thresholds, the variables involved and the logical combinations must be given, ideally as a table, together with the rule used for the 12/24/72 h windows at stations with different sampling rates. The derivation of causes related to weather from the Digitraffic causes array (which category codes count as weather-related) and the exact definition of difference In Minutes offset also need to be specified.
• [Essential] Software availability. The Zenodo record links three GitHub repositories (Railway-FMI-Data fetcher, Railway-FMI-Data viewer, Railway-FMI-Data training) that implement the pipeline, but the article does not cite them, none of them carries an open-source licence, and none has a tagged release or an archived snapshot with a DOI. Open Research Europe requires a Software availability section with the source code repository, an archived version and the licence; please add it and make sure that the archived code reproduces the released files and the baseline experiment.

5. Baseline Experiment
The experiment is honestly framed as an illustration rather than a state-of-the-art result, and the limitations paragraph at the end of the section is appreciated.
• [Essential] Without a naive benchmark the reported MAE has no interpretation. The station-specific offset target is exactly zero for about 70% of events network-wide, and at Oulu asema a constant predictor equal to the median gives an MAE of roughly 3.2 minutes on the December 2024 file, a high-delay month; over the full period the naive MAE is likely lower still, so the 2.73 minutes obtained by XGBoost may be only marginally better than predicting a constant. Please report naive baselines (zero, median, previous-station delay), an ablation with and without weather features, and at least one scale-free metric (for example MASE or MAE relative to the naive predictor). The conclusion that station-specific delays "are inherently more predictable than accumulated delays" should be withdrawn or reformulated: the three targets have very different scales (in December 2024 the offset target has a standard deviation of about 1.7 minutes against 12.6 for the cumulative delay), so a lower MAE on the offset target is mechanical rather than evidence of predictability.
• [Recommended] The reporting of the experiment is incomplete. It is not stated whether the curves in Figure 8 are cross-validation or test-set errors (if the latter, the test set was used for model selection), the randomized search space and the selected hyperparameters are not given, the feature list includes day of week both raw and encoded, and it is unclear whether the 101,146

observations include both ARRIVAL and DEPARTURE records at Oulu or the deduplicated set. Since weather is matched at the actual time of the event, the task is contemporaneous estimation rather than forecasting, and this should be said explicitly. The comparison with the MAE values of Huang et al. and Li et al. is not informative given the different targets and networks and could be reduced to one sentence.

6. Conclusions and Overall Presentation
The conclusions summarise the resource faithfully in most respects, and the future research directions are reasonable.
• [Essential] Several statements are not supported by the content of the article. The "geographic clustering of high-delay corridors in central and northern Finland" appears in the abstract, the plain language summary and the conclusions, but no figure, table or analysis of delays by station or corridor is presented (Figure 5 only shows the station-to-EMS pairing); either add the analysis or remove the claim. The text refers to a correlation heatmap that is not in the article. The statement that "most observed delays result from propagation rather than station-specific causes" goes beyond what the comparison of two target distributions can establish. Finally, the "approximately 38.5 million observations" (Table 9 total) refers to the data before removing the 23.24% duplicates; please state clearly which count describes the released files.
• [Recommended] Presentation. The text contains cross-references to numbered sections (Section 2.2, 2.2.2, 3, 3.6 and 4) that do not exist in the published layout and are presumably inherited from the preprint; please replace them with the section titles. Figure 4 misspells "Splitting", omits the fallback step and the feature engineering, and presents scaling and splitting as part of dataset construction. Minor issues include "Timetablerows" and "oulu" in the captions of Tables 2 and 3, "003 (01.43%)" in Table 6, a stray dash after "The denominator" in the paragraph that follows Equation 5, "abessment" in reference 1, and references 8, 9, 19, 21 and 22 that give "Reference Source" without a URL in the reference list.

Question 1: Is the work clearly and accurately presented and does it cite the current literature?
Ans: Partly
The literature is current and well chosen. The presentation is not accurate in several respects: the scope (long-distance passenger services only) is not stated, the timestamps are mislabelled as UTC+2, the delay definitions are inconsistent, and Table 10 does not describe the released files (Sections 1, 2 and 4).

Question 2: Is the study design appropriate and does the work have academic merit?
Ans: Partly
The design of the resource (spatial-temporal matching with a radial fallback, several target formulations, a seven-year span in a sub-Arctic climate) is appropriate and the dataset has clear merit. However, the pipeline conflates dataset construction with experiment-specific preprocessing (duplicate removal, splitting, scaling), and the baseline uses a random row-level split of strongly dependent events (Section 3).

Question 3: Are sufficient details of methods and analysis provided to allow replication by others?
Ans: No
The temporal tolerance and tie-breaking rule, the duplicate-detection key, the numerical

thresholds of the ten weather-scenario indicators, the derivation of causes related to weather and the hyperparameter search space are not given, and the processing code is neither cited nor archived. The engineered features cannot be reproduced from the article (Sections 3, 4 and 5).

Question 4: If applicable, is the statistical analysis and its interpretation appropriate?
Ans: Partly
The descriptive analysis is reasonable, but the unit of analysis (train events, with stops counted twice) and the delay threshold must be stated. The baseline experiment lacks naive benchmarks and a scale-free metric, the comparison across targets with very different variances is not valid, and the random split is likely to leak (Sections 2, 3 and 5).

Question 5: Are all the source data underlying the results available to ensure full reproducibility?
Ans: Partly
The Zenodo deposit is open, licensed CC BY 4.0 and usable, which is a strength. However, it does not contain the features described in the article, its description states a different feature count, and the code needed to regenerate the derived features has no licence and no archived version (Section 4).

Question 6: Are the conclusions drawn adequately supported by the results?
Ans: Partly
The description of the resource is largely supported, but the claims of geographic clustering of high-delay corridors, of a correlation heatmap, of delays being mostly due to propagation, and of station-specific delays being inherently more predictable are not supported by the analyses presented (Sections 5 and 6).

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Railway Engineering, LLM systems, Machine Learning, Data Science, MLOps

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 16 September 2026

https://doi.org/10.21956/openreseurope.26824.r80611

© 2026 ONETO L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Luca ONETO
Università di Genova, Genova, Italy

The manuscript presents FI-TW, a large open dataset integrating Finnish railway operational information with meteorological observations. The dataset covers the period 2018-2024 and includes approximately 38.5 million observations and 138 operational, temporal, meteorological, and engineered features. I consider the dataset potentially valuable for railway reliability analysis, weather-impact assessment, and machine-learning research, particularly because of its long temporal coverage and its focus on Nordic climatic conditions. The manuscript is generally well motivated and cites relevant recent literature. I therefore support indexing, subject to the authors addressing the points below.

1.The temporal matching procedure should be described more precisely. The manuscript states that nearest-neighbor temporal matching was performed within a predefined tolerance window, but the actual tolerance is not reported. Please state the exact tolerance used, explain the matching rule when observations are equidistant, and clarify how weather observations recorded at 1-minute and 10-minute intervals were handled. It would also be important to clarify whether weather observations were matched to scheduled train time, actual train time, or another timestamp

2. The train/test splitting procedure requires clarification because of possible information leakage. The authors use an 80/20 split and 5-fold cross-validation for the XGBoost baseline. However, it is not stated whether the split is random, chronological, grouped by train journey, or otherwise structured. Because observations from the same train, route, or day may be highly dependent, a random row-level split could result in optimistic performance estimates. For this type of time-dependent railway dataset, a chronological or grouped evaluation would be preferable, or the limitation of the current evaluation should be discussed explicitly.

3. The order of the missing-value imputation procedure should be clarified. The manuscript describes month-specific median imputation before discussing the train/test split, whereas only the scaling procedure is explicitly stated to have been fitted using the training partition. If monthly median values were calculated using the complete dataset, this would introduce information from the test set into preprocessing. The authors should confirm that all imputation statistics used in the baseline experiment were derived from the training data only, or revise the procedure

accordingly

4. The duplicate-removal procedure requires further explanation. The manuscript reports that exact duplicates represented 23.24% of the dataset, which is a relatively large proportion. The authors should specify exactly which columns were used to identify duplicates, explain why such a high proportion of duplicates occurred, and demonstrate that genuinely distinct train events were not inadvertently removed. This is especially important if identifiers were removed before duplicate detection

5. The weather-scenario variables are not described with sufficient detail for full reproducibility. Conditions such as Blizzard, Heavy_Snow, Extreme_Cold, Freezing_Rain, and Black_Ice are described qualitatively, but the numerical thresholds and logical rules used to construct these variables are not provided in the manuscript. The authors should provide the precise thresholds and combinations used, either in the paper, supplementary material, or openly available processing code

6. Additional information about the spatial fallback procedure would improve the manuscript. When a weather variable is unavailable at the nearest meteorological station, the authors search for another station within a 50-km radius. The rationale is understandable, but the manuscript should report the distribution of distances between railway locations and meteorological stations and the percentage of observations or variables requiring the fallback procedure. The potential limitations of using observations from relatively distant stations for localized weather phenomena such as precipitation, fog, or snow should also be discussed

7. Several internal inconsistencies in the dataset description should be corrected. For example, the manuscript reports 549 Finnish railway stations, comprising 340 non-passenger and 209 passenger stations, but the subsequent functional categories appear to total a different number. In addition, one section states that the original hour feature was removed after cyclical encoding, while the final feature table and later dataset description indicate that hour is retained. These inconsistencies should be reconciled so that the final dataset structure is unambiguous

8. The description of the meteorological variables in the released dataset should also be clarified. The final feature table identifies 11 base meteorological variables, while the Data Availability section refers to 13 raw meteorological parameters and still reports a total of 138 features. The authors should clearly distinguish between variables in the raw or intermediate datasets and those retained in the final processed dataset.

9. The RobustScaler equation should be checked. The equation shown in the manuscript appears to subtract the first quartile from each observation. If the standard scikit-learn RobustScaler with default settings was used, the transformation is normally centered using the median and scaled using the interquartile range. The equation and implementation description should therefore be verified and corrected if necessary

10. Some conclusions should be expressed more cautiously. The manuscript states that summer delay rates are below 20%, while the text reports June at 21.8%. This should be corrected or rephrased. In addition, the statement that most observed delays result from propagation rather than station-specific causes appears stronger than what can be established from the comparison of cumulative and station-specific delay distributions. The analysis shows a difference between the

two target formulations, but it does not by itself establish a causal explanation

11. The predictive interpretation of the baseline experiment should be clarified. The authors correctly describe the XGBoost experiment as a preliminary demonstration using one station, one algorithm, and a limited feature set. This cautious framing should be maintained throughout the manuscript. In particular, if weather measurements are matched to the actual time of the train event, the experiment may be better interpreted as contemporaneous or retrospective estimation rather than true advance forecasting. Clarification of the temporal matching procedure would resolve this issue.

12. A few additional presentation details would improve the usability of the dataset. The authors should define more explicitly how the delay-severity categories shown in Figure 2 were calculated, clarify how cancelled trains are handled when delay targets are missing, and indicate clearly whether the released meteorological variables are provided in their original physical units or in scaled form.

Overall, I consider FI-TW to be a useful and potentially important open dataset. The main contribution is the dataset itself rather than the preliminary prediction experiment. The work appears suitable for indexing once the methodological ambiguities, reproducibility issues, and internal inconsistencies described above are addressed.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.