Predictive Classification and Understanding of Weather Impact on Airport Performance through Machine Learning

Michael Schultz[1], Stefan Reitmann[2], Sameer Alam[3]

[1] *Dresden University of Technology, Institute of Logistics and Aviation*
[2] *Freiberg University of Mining and Technology, Institute of Informatics*
[3] *Nanyang Technological University, Air Traffic Management Research Institute*

Abstract

Efficient airport operations depend on appropriate actions and reactions to current constraints. Local weather events and their impact on airport performance may have network-wide effects. The classification of expected weather impacts enables efficient consideration in airport operations on a tactical level. We classify airport performance with recurrent and convolutional neural networks considering weather data. We are using London–Gatwick Airport to apply our developed approach. The weather data is derived from local meteorological reports and airport performance is derived from both flight plan data and reported delays. We show that the application of machine learning approaches is an appropriate method to quantify the correlation between decreased airport performance and the severity of local weather events. The developed models could achieve prediction accuracy higher than 90% for departure movements. We see our approach as one key element for a deeper understanding of interdependencies between local and network operations in the air transportation system.

Keywords: machine learning, airport performance, weather impact, feature importance, performance prediction

1. Introduction

The performance of the airport system depends on internal and external constraints. In this context, weather conditions have a significant impact on aircraft and airport operations, which can also affect the entire aviation network. In 2019, reactionary delays from previous flights continued to be the main delay cause with 44.4% and are followed by delays originated by aircraft turnaround at airports with 32.6% (Eurocontrol, 2020). To mitigate the effect of delayed operations, the prediction of aircraft processes along the whole trajectory in the air and at the ground is required. As depicted in Fig. 1, the average time variability during the flight phase is at least three times smaller than the variability of arrival and departure times. Flight delays are important for Air Traffic Management (ATM) and could be induced by weather and traffic situations as well as controller actions (e.g., allowing direct routes (Bongiorno et al., 2017)). Typical standard deviations for airborne flights are 30 s at 20 min before arrival (Bronsvoort et al., 2009), but could increase to 15 min when the aircraft is still on the ground (Mueller and Chatterji, 2002). Fig. 1 exhibits that the average time variability (measured as standard deviation) during the flight phase (5.3 min) is higher than in the taxi-out (3.8 min) and in the taxi-in (2.2 min) phases. However, it is still significantly lower than the variability of both the departure (16.1 min) and arrival (18.1 min) phases (Eurocontrol, 2017).

Since the variability at the gate-to-gate (flight) phase is small, the departure variability at the upstream airport has a high impact on the actual arrival time (Tielrooij et al., 2015). In this

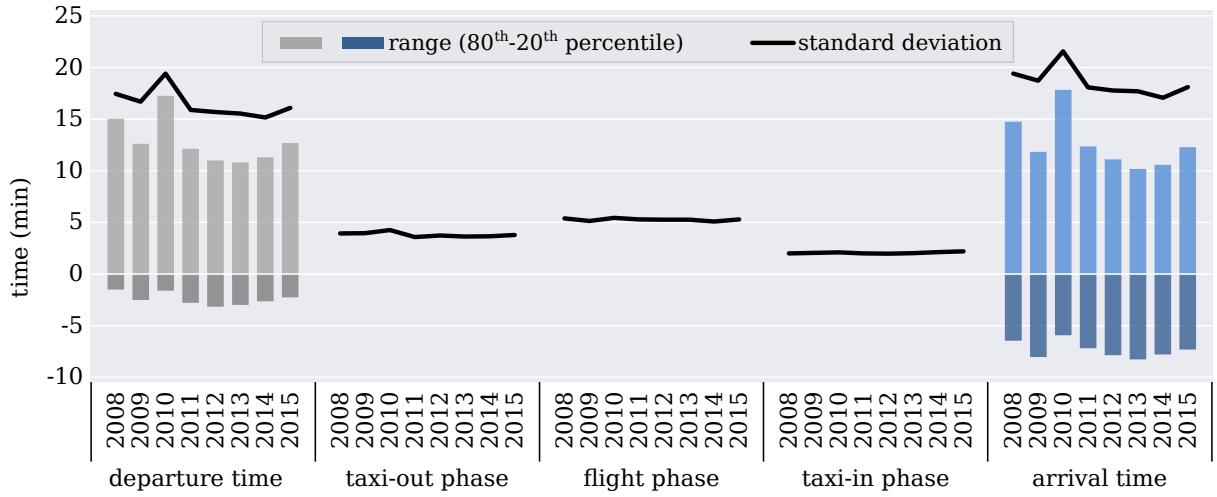

Figure 1: Analysis of European flights from 2008–2015 regarding time variability of flight phases, not considering flights departing to or arrival from outside Europe. The data are provided by Eurocontrol until 2015 but are no longer part of the latest performance review reports (Eurocontrol, 2017, 2016).

context, the air transport network could be seen as a network of interconnected airports rather than a network of airlines. Mutual interdependencies between airports result in system-wide, far-reaching effects, as departing delays propagate through the network. Thus, today's ground operations gain more relevance in providing reliable timestamps for aircraft departure. All local airport stakeholders (airlines, airport, ground handling agencies, network manager, air navigation service provider) play a significant role to improve the network punctuality (Helm et al., 2015; Rosenow and Schultz, 2018; Ali et al., 2019). Providing reliable and predictable departure times is one of the main tasks of these ground activities. For example, airlines strategically implement buffers to absorb a part of the delay generated by tactically reducing its propagation and achieving the desired target of punctuality (Cook et al., 2010; Sohoni and Erat, 2015).

The flow management positions report weather-related delays as the second most common cause of en-route Air Traffic Flow Management (ATFM) delays (21.2%) according to the Performance Review Report 2019 (Eurocontrol, 2020). Suppose the traffic demand exceeds the available airport capacity. In that case, the traffic flow will be regulated, and the corresponding delays will be assigned to the airport ATFM delay (6.5 Mio. delay minutes in the Eurocontrol area in 2019). Besides these delay measurements, additional flight times in the Arrival Sequencing and Metering Area (ASMA) around airports (40 NM radius) are used as an indicator for efficient operations (Cappelleras, 2015). The top 30 airports cause three-quarters of the airport ATFM delay. These major airports are already operating at their maximum capacity (see Fig. 2), which results in severe operational inefficiencies in the case of operational deviations and disturbances. In this context, nearly half of the delay is caused by adverse weather conditions (47.6%) (Eurocontrol, 2020). In our analysis, we consider London–Gatwick Airport (LGW) as an exemplary use case, since this airports operates at its declared capacity since years.

This airport was already used for data-based analysis in the context of airport performance (Schultz et al., 2019). LGW has a dependent parallel runway system with a distance of 200 m between and operates one of the busiest single runway globally. Since only 08R/26L possesses an instrument landing system, 08L/26R is mainly used as a taxiway or a backup runway during maintenance. Furthermore, LGW shows the second-highest additional ASMA time with 4.6 minutes per arrival (average of 2.17 minutes at the top 30 airports) and a share of 6.4% of the airport ATFM

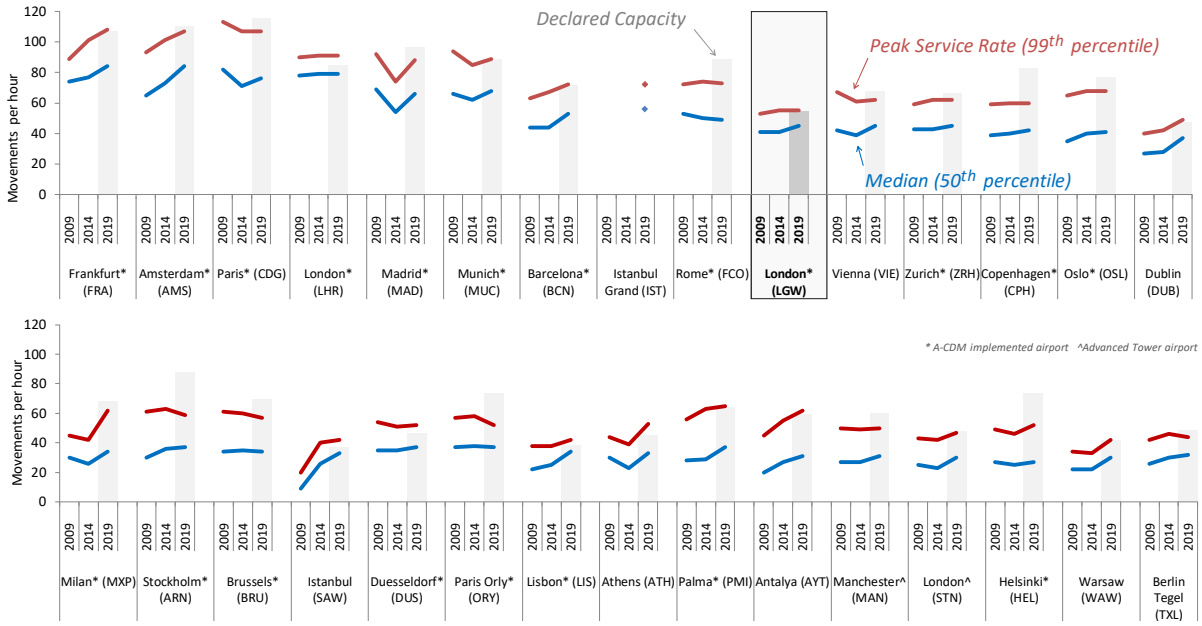

Figure 2: Aircraft movements (arrival and departures) at 30 major airports in Europe in comparison to their declared capacities in 2019 (Eurocontrol, 2020). We will focus on LGW in our current research.

delay. This means that the local airport's performance will significantly impact aircraft movements that are already regulated by the network manager (e.g., shifting departure times). The additional amplifying effect can lead to further delay cascades in the network. Given the high traffic volume, the single runway layout, and the highly utilized runway system at LGW, deviations from optimal conditions (e.g. weather events) immediately impact the airport's performance and make it difficult to recover from these situations. At this point, forecasting the expected performance of the airport based on local boundary conditions can contribute significantly to improving flow and capacity management at the airport.

During periods of operation near the maximum capacity, the anticipation of predicted operating conditions could mitigate cascading effects and delay propagation in the network. Nearly all of the 30 major airports are equipped with modern data processing systems, such as an Airport-Collaborative Decision Making (A-CDM) implementation (Eurocontrol Airport CDM Team, 2017) or advanced towers (Eurocontrol Network Manager, 2020). The available data will allow identifying precursors for future operational system states (e.g. delays) and indications of correlations between the severity of weather events and the corresponding impact to the airport performance. In (Schultz et al., 2018) the expert-based Air Traffic Management Airport Performance (ATMAP) algorithm (Performance Review Commission, 2009) was applied for statistical analysis (ex-post) to provide arrival and departure delay distributions in correlation to prior classified weather conditions. However, correlation analysis does not necessarily have to be based on given inputs from domain experts, so machine learning approaches can also be used to compile these correlations (supervised or unsupervised). The domain knowledge is still needed to enrich the input data and check the plausibility of the obtained results. Data-driven methods enable new approaches to provide improved situational awareness for aviation stakeholders (airline, airport, air navigation service providers). Thus, efficient tactical and operational measurements can be derived to mitigate the negative impact of operational deviations and disturbances.

1.1. Status quo

Current research in the field of aviation and airport operations addresses a broad range of contributions to improve the economic, operational and ecological efficiency of the air transportation system (Gerdes et al., 2018; Standfuss et al., 2018; Rosenow et al., 2017; Niklaß et al., 2017; Santos et al., 2017; Ingrid Gerdes et al., 2016; Kaiser et al., 2012; Carlier et al., 2007; Gerdes et al., 2020; Rosenow et al., 2019). In particular, challenges arise from the climate change and the expected increase of severe weather events are analysed with a focus on resilience and vulnerability of the transportation system conditions (Zhou and Chen, 2020; Markolf et al., 2019; Taszarek et al., 2020; Burbidge, 2016; Stamos et al., 2015).

Dynamic traffic situations emerge from traffic flow patterns across Europe and to/from intercontinental flows, military operations (Islami et al., 2017), volcanic ash eruptions (Luchkova et al., 2015), zones of convective weather (Kreuz et al., 2016), prevention of contrails (Rosenow et al., 2017, 2018), consideration of commercial space operations (Kaltenhaeuser et al., 2017) and integration of new entrants (Sunil et al., 2015; Schultz et al., 2019). Current research also address passengers metrics to evaluate flight performance (Montlaur and Delgado, 2017), which can be particularly relevant when optimizing arrival flows at airports under uncertainty (Delgado and Prats, 2014; Buxi and Hansen, 2013). Thus, delay generation due to weather impacts including location and time of the primary delay generation and its evolution are relevant to capture the complexity of the system dynamics.

The propagation of delay in the network is paramount when assessing the impact of (local) congestion (Campanelli et al., 2016; Ivanov et al., 2017; Baspinar et al., 2016). Particularly, research is conducted to evaluate the impact of disruptions on airport turnaround operations (Postorino et al., 2020), the impact of sudden and slow onset weather events on departure delays (Borsky and Unterberger, 2019), the handling of uncertainties in the arrival management (Schultz et al., 2012; Förster et al., 2021), and the management of airports in extreme winter conditions (Merkert and Mangia, 2012). The delay propagation is particularly critical when estimating the resilience of the air traffic management system and the impact of different mechanisms on the expected performances' variations (Cook et al., 2016; Proag and Proag, 2014; Cook et al., 2009).

The analysis of local airport situations allows anticipating congested times and mitigating the negative impact of delays to operations to airline and airport operations (Arnaldo Scarpel and Pelicioni, 2018; Henriques and Feiteira, 2018). Here, historical data provide a profound basis to improve weather forecast using data analytics and machine learning approaches (Rozas Larraondo et al., 2018; Schultz et al., 2019; Reitmann and Schultz, 2018; Herrema et al., 2019) for fog forecast (Ming et al., 2019), forecast of poor-visibility episodes near complex terrain (Fernández-González et al., 2019) or develop a robust model for learning and recognizing weather pattern (Salman et al., 2018). Besides analysis and forecast, the optimization of specific operational environments considering volatile constraints will also affect the performance of the entire airport system, such as efficient stand/gate allocation at the airport (Bagamanova and Mota, 2020) or dynamic reconfiguration of the capacity-limited terminal airspace system during convective weather (Serhan et al., 2019).

1.2. Objective and scope of the research and structure of the document

The research objective presented in this paper is to quantify the impact of local (severe) weather conditions on airport performance using a machine learning approach. We consider local weather data and airport performance data (scheduled and actual flight times) to correlate the complex dependencies in the airport systems. Herein, we are not focusing on the causes (input) but the consequences (output) to the airport performance. For this purpose, we categorize the airport performance and backtrack/ evaluate possible causes from the observed weather phenomenon. Finally, we will provide a methodology to map specific weather conditions to airport performance

impacts. Our new approach will overcome the limitations of current expert-based decisions and enable predicting future system states at the airport, considering permanent data updates. We use London–Gatwick Airport as a reference case to demonstrate how our data-driven approach performs.

The document is structured as follows. Sec. 1 provides an introduction of the topic and a status quo of related research activities. In Sec. 2, we introduce the datasets for weather and airport performance. Here we set a particular focus on the local weather data, which are taken from the local Aviation Routine Weather Reports (METARs). The METAR messages are parsed and translated into aviation relevant severity scores for weather events by using the expert-based ATMAP algorithm (Eurocontrol, 2011). We briefly describe the algorithm using METAR messages from London–Gatwick Airport and show a correlation between the algorithm-derived weather score and the delay situation at the airport. This approach provides an initial indication of how weather conditions will affect airport performance. In the following section, we provide the fundamental background for our machine learning approach, information about the process of classification, and the model setup (Sec. 3). In Sec. 4, several neural networks are implemented accordingly and applied to the prior processed datasets (weather conditions, airport performance). Furthermore, features are extracted to provide information about the importance of particular input values to the current airport and weather conditions. From this point on, the expert-based assessment of weather events by the ATMAP algorithm is no longer required. Finally, the document closes with a conclusion and outlook (Sec. 5).

2. Weather and airport performance

The dataset we used for the analysis consists of flight plans and weather data from major European airports for the years 2013-2015. The flight plans include scheduled and actual times of aircraft movements. Weather events, which are relevant for air traffic operations at airports, are derived from the local METAR data. From this dataset, we used a subset with a focus on LGW.

2.1. Airport performance

The performance of an airport is mainly related to the number of aircraft movements handled in comparison to the maximum of (observed) movements in a comparable situation (airport capacity). In this case, the term capacity generally refers to a given transportation facility's ability to accommodate a traffic volume (e.g., movements) in a given period (e.g., on an hourly, daily, or yearly basis). When the air traffic demand approaches or exceeds airport capacity, the congested infrastructure leads to delays and cancellations of air traffic movements. This imbalance between demand and capacity is one of the main causes of unpunctual operations and affects several components of the whole airport system on the airside (e.g., runways, taxiways, aprons) and landside (e.g., passenger handling (O'Flynn, 2016; Schultz and Fricke, 2011)). Flight delays are defined as the difference between the scheduled and actual times of arrivals and departures, where the reference measure points are usually on-/off-block times at the gate/apron position. Punctuality is determined as the proportion of flight delays with less than 15 min, widely accepted as a performance indicator in aviation. To avoid delays during periods of high traffic demand (peak times), airlines apply buffering strategies, which improve punctuality and minimizes tactical delay costs (Eurocontrol, 2020; Cook et al., 2016; Evler et al., 2020). The definition of delay varies between stakeholders and several terms have been established, such as acceptable delay, network delay, on-time performance, reactionary delays, delays per flight (gate-to-gate view), arrival delays, additional time in the arrival sequencing and metering area, departure delays, surface taxiing delays, and passenger delay minutes (cf. O'Flynn (2016)).

2.2. Weather data

Weather conditions are usually recorded at each airport using the Meteorological Aviation Routine Weather Report (METAR) (Administration, 2016). METARs are reported in combination with Terminal Aerodrome Forecasts (TAFs). While TAF provides forecast values for the next 6 hours, METAR data are usually measured every 30 min. The update interval of a METAR weather report is not harmonized and implemented differently worldwide. For example, a METAR message is released every half an hour at larger German airports (20 min past and 10 min to the full hour). The Unscheduled Special Weather Report (SPECI) is another format representing significant changes in airport weather conditions. Current and historical weather data are accessible at different publicly available websites. In addition to information about the location, the day of the month, and the UTC-time, METAR contains relevant information for airport operations, such as wind speed and direction, visibility, precipitation, clouding, air temperature, and pressure.

An exemplary METAR message is provided in Tab. 1 and contains the following information: (a) LGW, 21st day of the month, time 08:20 UTC, (b) wind direction 310 degrees, wind speed of 13 knots, (c) a visibility of 3000 m, broken sky, (d) light snow, (e) temperature -1°C, dew point 0°C, and (f) sea-level pressure of 1003 hPa.

Table 1: Main components of Meteorological Aviation Routine Weather Report (METAR) message.

Example: *EGKK 210820Z 31013KT 3000 -SN BKN006 M01/00 Q1003*

Parameter	Measurement	METAR Code (Example)
wind	direction azimuth in degrees/speed [kn]	31013KT
visibility	horizontal visibility [m]	3000
precipitation	significant weather phenomenon	−SN
cloud	cover/height*100 [ft] above aerodrome level	BKN006
temperature	air temperature/ dew point [°C]	M01/M00
pressure	sea-level pressure (QNH) [hPa]	Q1003

Fig. 3 depicts general weather information derived from METAR massages of LGW for the first 60 days of operations in 2014.

Besides this general weather information, additional measurements were available related to adverse weather situations, such as information about wind gusts, runway conditions (e.g., ice layer), thunderstorm-related cloud formations, or measurements of runway visual range. The use of METAR weather records for data analysis demands a comprehensive analysis since the data provider does not assure the data integrity. Typically, the data (partially) lacks significant information, such as wind data, dew-point data, runway condition information (e.g., depth of deposit), or incomplete information about airport runway conditions.

For the following analysis of the weather-performance correlation, the METAR messages are parsed and filtered to enable a quantification of the weather measurements regarding their impacts on the aviation domain. Eurocontrol provides a framework for measuring airport airside and nearby airspace performance for this quantification (Performance Review Commission, 2009). Here, weather conditions are generally separated into nominal, degraded, and disruptive conditions with an increasing impact on airport performance. Furthermore, the ATMAP algorithm was developed to describe weather conditions at European airports, considering expert judgments and allowing quantification of current weather conditions (Eurocontrol, 2011). We implemented the ATMAP approach to show how weather conditions can affect airport operations.

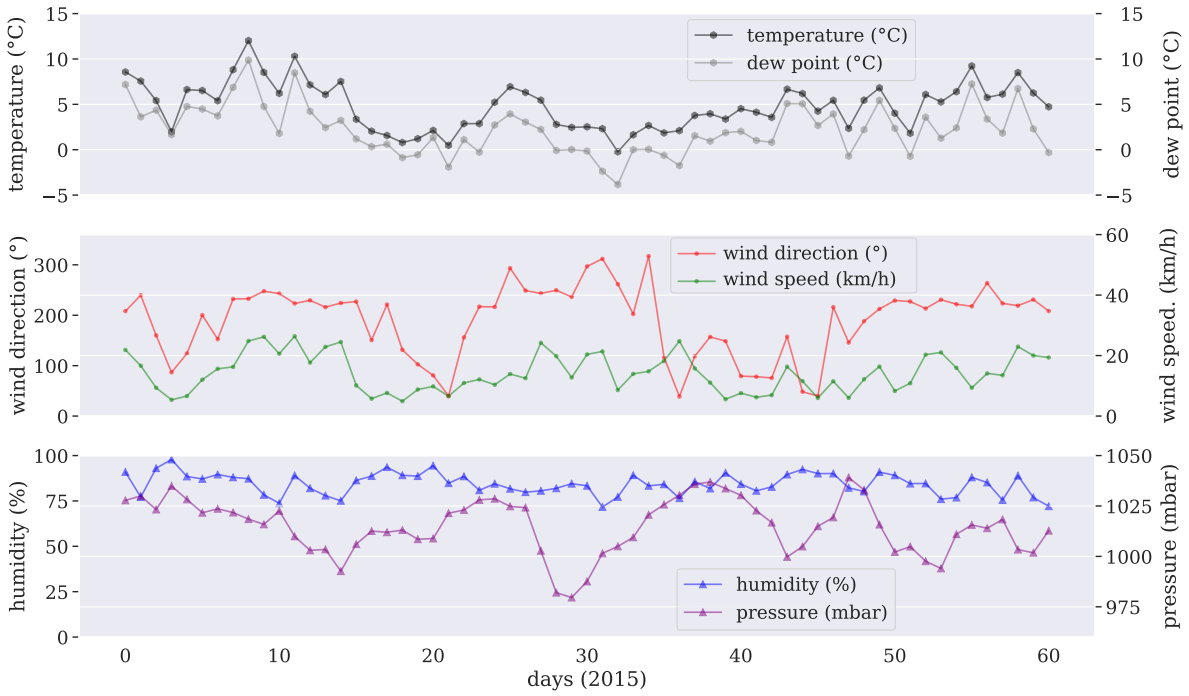

Figure 3: Weather data from the first 60 days of 2014 at LGW.

2.3. Concept of the ATMAP algorithm

The following definitions are used in applied the ATMAP algorithm (Eurocontrol, 2011): *weather phenomenon* is a single meteorological element that impacts the safety of aircraft during air and ground operations; *weather class* is a group of one or more weather phenomena affecting the airport performance; *severity code* is a ranking number of the weather class status (from best to worst); *coefficient* represents the assignment of a score to a given severity code to describe the nonlinear behavior of various weather phenomena (zero is the default value for nominal conditions). A multi-step procedure is proposed to determine the ATMAP weather score: in the first step, a given METAR observation at the airport will be assessed by specifying the severity code and its associated coefficient for each weather class. This METAR message is parsed, filtered, and transformed to the quantified measures (coefficients). In a second step, these weather class coefficients are summed up to the corresponding ATMAP score (per METAR message). Finally, for a given time interval (hours of operations), the sum of all ATMAP scores are divided by the number of METAR reports to calculate an average ATMAP score per time interval (e.g., per hour, per day).

The ATMAP algorithm quantifies and aggregates major weather conditions at airports, significantly impact airport operations. Five different weather classes with a significant influence on aircraft and airport operations include: (1) visibility and cloud ceiling; (2) wind; (3) precipitation; (4) freezing conditions; and (5) dangerous phenomena. In Tab. 2, these five different weather classes are shown, described with meteorological conditions, and linked to the associated maximum coefficient defined by the ATMAP algorithm. Compared to the other weather classes, dangerous phenomena have a particularly high impact on airport operations, resulting in the highest coefficients. For both towering cumulus clouds (TCU) and cumulonimbus (CB), the ATMAP coefficients are depending on the cloud coverage (FEW, SCT, BKN, OVC) and range from 3 to 10 (TCU) or 4 to 12 points (CB). Showery precipitation and intensive precipitation can lead to a further increase

in coefficient values up to 18 (TCU) or 24 points (CB). Other dangerous phenomena with an impact on safe aircraft operations can be divided into three groups with 30 points (heavy thunderstorm), 24 points (e.g., sandstorm, volcanic ash), and 18 points (small hail and/or snow pellets).

Table 2: Weather classes defined in the ATM Airport Performance (ATMAP) algorithm.

Weather Class	Description	Meteorological Conditions	Coefficient
(1) ceiling and visibility	deterioration of visibility	precision approach runways (CAT I-III)	max. 5
(2) wind	strong head-/cross-wind	Wind speed $>$ 16 knots (+gusts)	max. 4 (+1)
(3) precipitations	runway friction influencing runway occupancy time	e.g., rain, (+/−) snow, frozen rain	max. 3
(4) freezing conditions	reduced runway friction, de-icing	T $\leq 3°$C, visible moisture, any precipitation	max. 4
(5) dangerous phenomena	unsafe ops, unpredictable impact	TCU/CB, cloud cover, (+/−) shower, storm	max. 30

In Tab. 3, an example METAR report from London–Gatwick Airport is given to show the transformation from the raw METAR message to a quantified ATMAP score (cf. Tab. 1). The given METAR report is evaluated by specifying the severity of the specific meteorological condition and the associated coefficient for each of the five weather classes. These particular weather coefficients are summed up to the corresponding ATMAP score.

Table 3: ATMAP weather score is based on local airport METAR messages (London–Gatwick Airport).

Example message: *EGKK 210820Z 31013KT 3000 -SN BKN006 M01/00 Q1003*

	Weather Classes					ATMAP
	Visibility	Wind	Precipitations	Freezing	Dangerous	Score
METAR	**3000**	**31013KT**	**−SN**	**M01/00**, **−SN**	-	
ATMAP	0	0	2	3	0	**5** (sum)

2.4. Flight plan and weather data

The introduction of the ATMAP algorithm (Eurocontrol, 2011), and a first application of the algorithm to evaluate the effect of weather events on airport performance on a European scale (Schultz et al., 2018) emphasize a correlation of the on-time performance of flights with the present weather at airports by using 20.5 million flights and local weather data. In this contribution, we set a focus on the LGW because two major benefits could be observed: (1) the absence of parallel or cross runways allows an isolated impact analysis, and (2) the high utilization of the single runway close to its maximum capacity leads to an intermediate system response when weather conditions change. When the ATMAP algorithm is applied to the first 60 days of operation in 2014 at LGW, wind and dangerous phenomena will be identified as the most severe weather classes (Fig. 4).

A more comprehensive picture of airport performance dependencies and weather conditions arises if scheduled/actual flight plan data and weather data are combined. Fig. 5 (left) provides

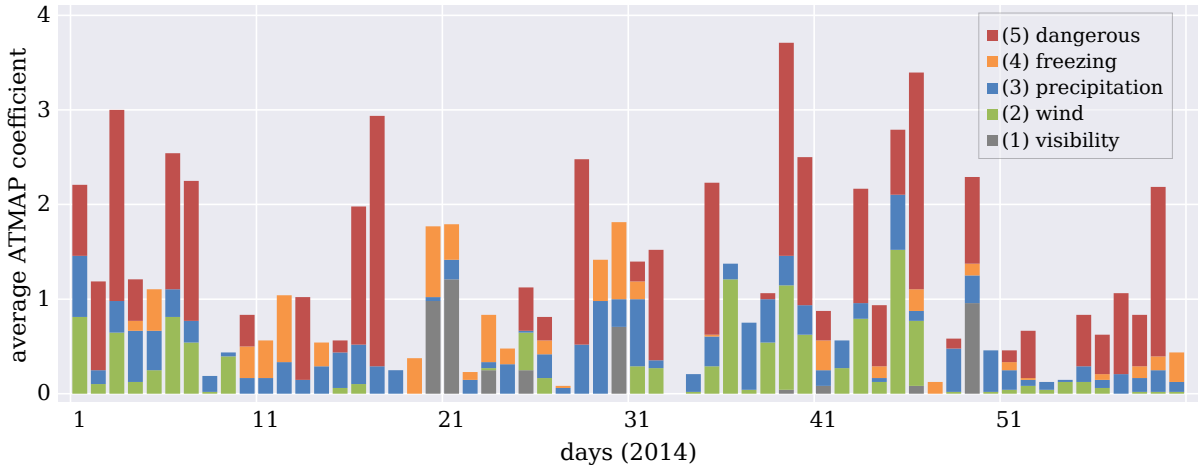

Figure 4: Corresponding ATMAP weather score for the first 60 days in 2014 at LGW.

an example of how the delay at the airport increases rapidly to 795 minutes (accumulated delay minutes from all flights in 1 hour) at the beginning of the day of operations due to a 2 hour upstream period of fog (hourly ATMAP weather score of 5). The observed delay is different from a situation with nominal weather conditions, where the delay generally does not exceed 400 min. The data from LGW also confirm that the impact of weather events on airport performance is even higher with an increasing severity level. Fig. 5 (right) exhibits the operational consequences of 4 hours of thunderstorms and rain (06:50 - 10:20 hours) in the vicinity of the airport. Since the traffic demand is increasing due to this time, the accumulated delay could only be reduced slowly over the whole day and affects almost all aircraft operations.

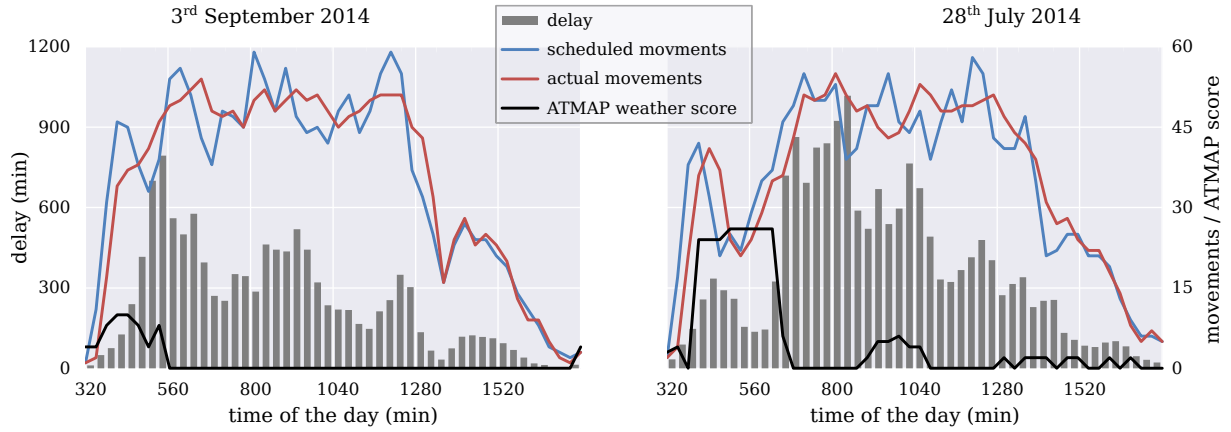

Figure 5: Airport performance data (delay minutes) and ATMAP weather score at LGW airport for 3^{rd} September 2014 (left) and 28^{th} July 2014 (right). The graphs show a disruptive weather event (peak of black line) followed by a severe delay.

3. Modular Machine Learning Approach

Our approach differs significantly from the basic considerations made by the prior introduced ATMAP algorithm, where expected impacts on airport operations are associated with five weather classes. These classes are assumed to be independent and universal for European airports, but are not related to specific airports or regions. We have a critical view of this, as an airport's location

and meteorological conditions have a significant impact on its performance. Therefore, we aim to address these issues with our machine learning approach and focus on two core features during the model development: the model must be impact-based by linking effects to their causes. The model must be adaptive by enabling an airport-specific assessment. Accordingly, in our approach, we transform the weather categorization into an *inverse problem* and derive the triggering causes (weather) from the observed airport performance (output) by applying the following steps.

(1) *data preprocessing* of flight schedules/ weather data (Sec. 3.2)
- *feature selection*, choice of Performance Indicator (PI), Key Performance Indicator (KPI)
- *clustering*, class creation of impact data

(2) *machine learning model* as mapping function (Sec. 3.3)
- *model creation*, parameterization and setup
- *model training*, application of models to data

(3) *evaluation* of the machine learning model (Sec. 3.4)
- *validation*, cross-validation using prediction results
- *extraction*, knowledge about input importance

Due to a large amount of data, non-linear time series, and interdependencies, self-learning algorithms are used. We assume that these algorithms offer opportunities for independent, complex solutions to similar problems. In this case, an Artificial Neural Network (ANN) serves as an adaptive intermediate model to process weather and airport performance data (see Fig. 6 - the numbers correspond to the three model steps explained above). The selection of features in this dataset can be done algorithmically or by using expert knowledge (Sec. 3.2.1). The same applies to the classification of impacts on airport performance (Sec. 3.2.2). The neural network itself is determined by its parameters, its structure, and the range of data available (Sec. 3.3).

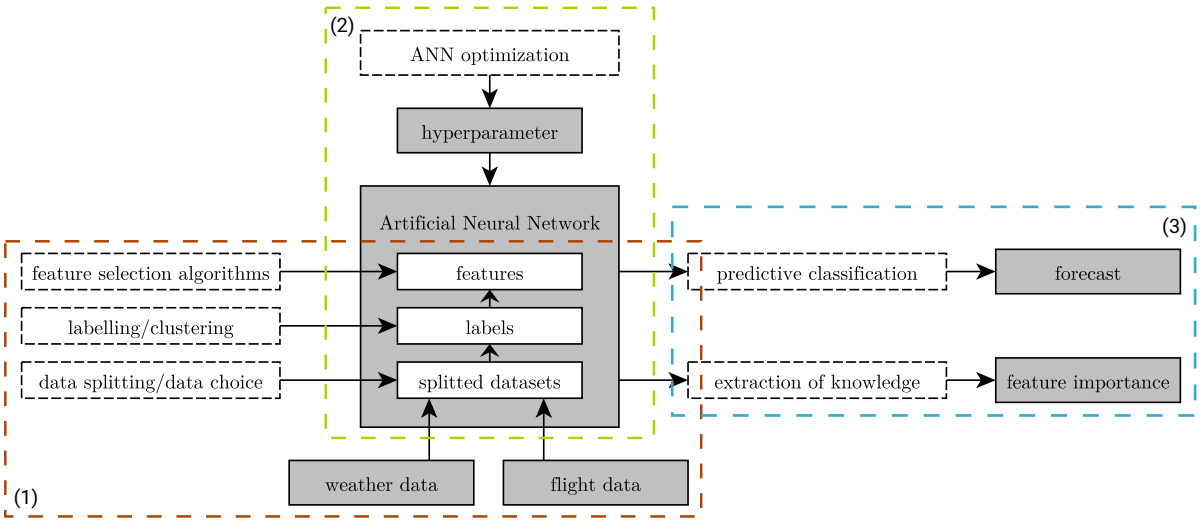

Figure 6: Outline of the process of data classification. The ANN is algorithmically optimized as the central model for determining the prediction and importance of the features. The data are selected and prepared by expert knowledge or algorithms.

3.1. Classification with Neural Networks

Classification is about predicting a label and regression is about predicting a quantity. Classification predictive modeling provides an estimate for a mapping function f from an input X to an output Y. The output variables are termed *labels* or *categories*. The mapping function predicts the class or category for a given observation. Classification problems can be solved by a variety of methods within and outside machine learning. They all have advantages and disadvantages, and their applicability depends on the particular application. An example is the paradigm of Support Vector Machines (Cortes and Vapnik, 1995). Neural networks also offer the opportunity to classify non-linear, multivariate data. Thus, it is possible to build adaptive decision support that delivers complex but fast outputs to specific input sets. This is why we focus on neural networks in our application.

There are two main neural network approaches that are suitable for time series classification and have been proven in various applications. These are Convolutional Neural Network (CNN) (Goodfellow et al., 2016) and Recurrent Neural Network (RNN) (Hagen et al., 2014), especially their sub-type Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). CNN are recommended to detect short-term correlations that have a natural order, while RNN and LSTM perform better in deriving long-term repeated interdependencies. The reason for this is that RNN could take advantage of the time correlation between measurements, and CNN is better at learning deep traits contained in recursive patterns (Wang et al., 2018). The benefit of using LSTM for sequence classification is that they can learn from the raw time series data directly. In turn, they do not require domain expertise to manually engineer input features. The model can learn an internal representation of the time series data and achieve comparable performance to models that are fitted on a version of the dataset with engineered features.

We consider the following definitions for the machine learning application in our approach.

- The (actual) air traffic system is to be understood as dynamic system[1] (Reitmann, 2013). It is defined by a state space M, which is initially the $\mathbb{R}^n$ or an open subset thereof, and a one-parameter family of figures $\varphi^t : M \to M$, where the parameter t comes from $\mathbb{R}$ (continuous time). The total time set corresponds to Γ and includes all time points t which were used to measure the ATM (sub)system. φ^t are those figures which cause a time-dependent change of the system (and thus of the characteristics) and which are to be examined with the help of the ANN.

- An ANN is a directed graph $G = (U, C)$, whose nodes $u \in U$ are neurons and whose edges $c \in C$ are connections (Kruse et al., 2015).

- The set U of nodes is partitioned into the amount U_{in} of input neurons, the amount U_{out} of the output neurons, and the amount U_{hidden} of hidden neurons.

- The base set Λ denotes a set of examples which are used for the learning and assessment process of an ANN.

- The indicators or characteristics over which the data were aggregated are called *features* in the following. Let Ξ be the set of features available in the data. The PI/KPI used as features in the simulation form the set $\widetilde{\Xi}$, where $\widetilde{\Xi} \subset \Xi$. The total amount of existing features PI/KPI form Λ.

[1] according to which a dynamic system is a mathematical object for the description of the time evolution of physical, biological or other real existing systems.

- As training data $\Lambda_{train} \subseteq \Lambda$ refers to the data used to adjust the network. Validation data $\Lambda_{valid} \subseteq \Lambda$, on the other hand, are excluded from training and are used for the unbiased evaluation of the network capability, while hyperparameters of the network can still be varied. The test data set $\Lambda_{test} \subseteq \Lambda$ is used to evaluate the model after final training, based on the Λ_{train} used.
- A (fixed) learning task L_{fixed} for an ANN with n input neurons, i.e. $U_{in} = \{u_1, ..., u_n\}$, and m output neurons, i.e. $U_{out} = \{v_1, ..., v_m\}$, is a set $L_{fixed} = \{l_1, ..., l_r\}$ $l = (x^{(l)}, o^{(l)})$, each consisting of an input vector $x^{(l)} = (ext_{u_1}^{(l)}, ...ext_{u_n}^{(l)})$ and an output vector $o^{(l)} = (o_{v_1}^{(l)}, ..., o_{v_m}^{(l)})$ (Kruse et al., 2015).

3.2. Data Preparation and Clustering

To enable the process of classification, the target data streams must be labeled. The label creation can be done algorithmically or added with expert knowledge. The labels of the target variables should represent impact categories of the severity of the respective effect on airport performance. Using the general example of flight delays, it is expected that a deviation from the flight plan by -5 min to 15 min (*on-time*) will not have a significant impact on the airport performance (Eurocontrol, 2007). Further categories can be found that capture the severity of increasing deviations (delays). This, in turn, is an individual value that refers to the specific airport and demand/capacity conditions there. This form of label creation requires local expertise. It should also be noted that the exclusive consideration of delay is a one-dimensional target label. In this case, the creation of intervals is recommended[2].

A predicted delay is continuously carried along in the input vectors for the ANN. The other characteristics used for the forecast (traffic demand, weather) can be used since their future is given by flight plans and weather forecasts. This form of data reconstruction transfers the multilevel prediction into formal modeling. If the time series of the attributes are scaled differently due to their diversity, many machine learning algorithms can benefit from rescaling to a uniform scale (*normalization*). Attributes are scaled in the range between 0 and 1, which is useful for optimization procedures in the core of machine learning algorithms (e.g., gradient descent). At the same time *standardization* is applied, a technique for converting attributes with a Gaussian distribution and different μ and σ to a standard Gaussian distribution with values of $\mu = 0$ and $\sigma = 1$. This is suitable for techniques that assume a Gaussian distribution in the input variables (e.g., linear regression).

3.2.1. Feature Selection

The selection of the features is an essential part of the modeling and mapping process of an ANN. Only information stored in the data that is reflected by the selected features can be learned. This can lead to the fact that the ANN calculates very good results, which, however are based on correlations, which are not present in the real system[3]. Feature selection is a process that selects the features in the data that contribute most to the predictive variable or output. Irrelevant features in the data can affect the accuracy of many models. Feature selection offers three advantages: reduced overfitting (less redundant data means and fewer decision-making options), improved accuracy (less misleading data), and reduced training time (less data means algorithms train faster).

[2] In the course of an extended experiment, a two-dimensional classification was applied using a further variable (Reitmann et al., 2019). Here, the classification procedure was turned into a clustering based on k-means and fuzzy c-means algorithms.

[3] For example, an ANN can correlate aircraft color and taxiing speed as long as they are the two selected characteristics, but this correlation can be rejected by a causality check.

There are two basic ways to approach the problem of feature selection (see Tab. 4). First, expert knowledge from the real system is transferred, and a selection of specific features can be made based on previous knowledge of the system and its data representation. Thus, a reduction of the weather features from METAR and TAF messages can be made from empirical values about meteorological conditions. Second, analytical methods are used to examine the data basis and test possible correlations. For multivariate approaches, exploratory analysis offers a wide range of options to investigate correlations, causalities, and visual representations. Properties of the individual time series, which are important for modeling steps, are also mapped. In addition, methods are known which either independently select features (Recursive Feature Elimination (RFE) (Kuhn and Johnson, 2013)) or reduce the data basis (Principle Component Analysis (PCA) (Bishop, 2007)). The latter is not recommended for the prediction of certain values, since here a reduction to linear combinations of the original data set is performed.

Table 4: Problem specific and analytical methods for feature selection. Problem specific selection is suitable if sufficient knowledge of local conditions at an airport is available.

Application form	Method	Comments
problem-specific	expert knowledge	prior knowledge of relevance of the PI/KPI
problem-specific	explorative analysis	statistically provable/detectable influences
analytical	variance threshold	selection according to the limit for σ^2
analytical	k-best	selection according to univariate statistical tests
analytical	RFE	recursive observation of decreasing $\widetilde{\Xi}$

3.2.2. Data Splitting and Data Choice

After determining which attributes of the dataset will be used as features to work on the problem, it is necessary to define which data will be used for training (Λ_{train}), validation (Λ_{valid}), and testing (Λ_{test}) of the ANN model. In the training, the actual adjustment of the parameterization of the network takes place; the validation can be used to compare different network configurations with each other; the test evaluates the mapping capability from the input to the output. The learning tasks L presented in Sec. 3.1 can be derived from different perspectives. For example, it is possible to form a generally valid ANN, which takes the learning tasks L without any restriction from all procedures. However, it is also possible to use a L from a base set of a specific procedure and to apply it only to that procedure. Tab. 5 exhibits the assumptions we made in our approach.

Table 5: Data choice criteria depends on the area of application.

Scope	Division criterion	Example or note
LGW	all data	without restrictions
	limit separation	split by METAR indicator or KPI
specific use cases	airport specific	according to exploratory analysis
	time-specific	$\Lambda^{weekday}, \Lambda^{weekend}$

Considering that one day is the smallest indivisible unit (and thus represents one L each), 365 days a year can be categorized according to different criteria. For example, a simple limit separation is possible, which refers to a specific weather factor or an operative KPI. It is also possible to bundle data that does not refer to the weather but to the dynamics of the ATFM at the respective airport (e.g., consideration of special traffic peaks, exceptional situations, differentiation between weekdays and weekends). In return, prior knowledge about the airport can be applied, which is incorporated into airport-specific data filtering.

Identification of similar days is complex, especially for the application in ATFM. In this paper, similarity is understood as a similar temporal course of a KPI or weather value. Methods for the type and form of data set separation are as diverse as they are problem-specific. Similar days of operations can be defined by capacity and demand data and are subject to different clustering approaches. Using a PCA as authoritative clustering method (Gorripaty, 2017), a comparison of days by traffic demand seems to be useful but inappropriate concerning the possibilities of the dataset. In time series analysis, Dynamic Time Warping (DTW) (Müller, 2007) is an important algorithm for measuring the similarity between two temporal sequences. Fig. 7 depicts the comparison of Arrival (ARR) and Departure (DEP) movements at LGW.

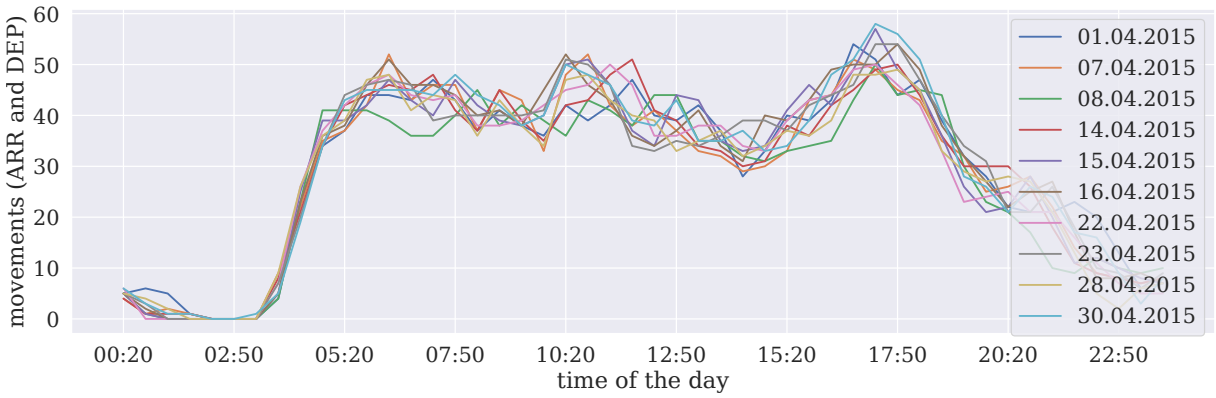

Figure 7: Comparison of movements (ARR + DEP) at the airport LGW by DTW in April 2015 for ten similar days (applying a maximum Euclidean distance of 130 as difference of the demand sequences over 24 hours).

Similarities in traffic demand could be found with the DTW, even if there were variations within the individual demand sequences. Sequences of similar delay courses can be identified. In the DTW, the Euclidean distance is used as a cost function. The comparative analysis is limited to the traffic demand (Gorripaty, 2017), which can be derived from the flight plan data. The concentration on traffic demand refers to the idea that the system excitation is compared, not the respective system behavior indicated by the flight delay. An ANN training based on similar data can lead to a generalization in which the network increasingly concludes that the output is uniform regardless of the type of input.

3.3. Machine Learning Model

The topology of the ANN consists of layers. Creating the ANN model structure means finding values for the number of layers of each type and the number of neurons in each of these layers. To ensure the generalization capability of the network, the number of neurons should be kept as low as possible. If there is a large surplus of neurons (and nodes within the network), the ANN becomes a memory that can retrieve the training set well but does not work well with data outside the training set (Lawrence et al., 2001).

Each of the ANN used has exactly one input layer consisting of a set of neurons U_{input} determined by the shape of the training data. In particular, the number of neurons that comprise this layer is equal to the number of features of the data. Some ANN configurations add a node for a bias term. Parallel to the input layer, each ANN also has exactly one output layer. The number of neurons U_{output} is determined entirely by the chosen model configuration. Suppose the ANN is a regressor or classifier. In that case, the output layer has one node (univariate prediction) or several nodes (multivariate prediction), which depends on the dimension of the output vector. Besides the input

and output layer, the number of hidden layers can be defined by the following basic rules (Heaton, 2008).

- 0 - ANN is capable of representing linearly separable functions or decisions.
- 1 - ANN can approximate any function that contains a continuous mapping from one finite space to another.
- 2 - ANN can represent any decision limit for any precision with rational activation functions and approximate any smooth mapping to any precision.

A hidden level is sufficient for the vast majority of problems. The situations where performance improves with a second (or more) hidden layer are rare. Looking back at the finite state space M and the characteristics of the figure φ allow the specification of a hidden layer. Finally, the question about the appropriate number of hidden neurons U_{hidden} forming the hidden layers arises. There are some empirically derived rules of thumb (especially Heaton (2008) should be mentioned here), of which - based on their positive results in experiments - are frequently used:

$$U_{output} \leq U_{hidden} \leq U_{input}. \tag{1}$$

From empirical values from (Hagen et al., 2014) the following quantification rule for the number of hidden neurons U_{hidden} can be derived, which specifically refers to the amount of data:

$$U_{hidden} = \frac{\#\Lambda}{\alpha * (U_{input} + U_{output})}. \tag{2}$$

The power of the training data $\#\Lambda$ refers to the learning patterns l and not to the number of learning tasks L, and α is a scaling factor between 2 and 10. This sensitizes the ANN to generalization and prevents overfitting.

3.3.1. Hyperparameter

In addition to the creation of the model structure, the initialization includes the initial parameterization of the ANN (Sutskever et al., 2013). To control the efficiency of the learning process, hyperparameters have to be set, such as a learning rate η and the batch size. The time needed to train and test a model may depend on the choice of the respective hyperparameter. For ANN, hyperparameter optimization describes the problem of selecting a set of optimal hyperparameters for a learning algorithm to form a model. The same type of ANN paradigm may require different constraints, weights, or learning rates to generalize different data patterns (e.g., the same RNN paradigm may require different parameters for various airports). To help the model solve the problem best, hyperparameter optimization finds a tuple of hyperparameters that defines an optimal model for minimizing a predefined loss function for given independent data. The objective function takes a tuple of hyperparameters and returns the corresponding loss. The cross-validation is used to estimate the generalization power (Kuhn and Johnson, 2013; James et al., 2013). The *grid search* is a technique for optimizing hyperparameters in the model (Claesen and Moor, 2015; Bergstra and Bengio, 2012). In our approach, we focus on the following set of hyperparameters:

- batch value and n_{epochs} (adjustment of the l sequence length and the number of epochs),
- selection of the optimizer,
- η (adjustment of the learning rate),

- U_{hidden} (adaptation of the network structure, especially the number of hidden neurons), and
- dropout rate (reduces overfitting).

The training of the used ANN is based on the gradient descent method (see Ruder (2016)). There are three variants of this approach, which differ in how much data is used to calculate the gradient of the objective function (Batch Gradient Descent (BGD), Stochastic Gradient Descent (SGD), Mini-batch Gradient Descent). While with BGD an update is performed after the complete data set, with SGD this is done for each training pattern l. The mini-batch gradient descent combines both methods. Consequently, the SGD optimizer is used with variable batch size and can thus be variably adapted to the conditions of the data set via further optimizers. We focus on optimizer *Adam* (Kingma and Ba, 2014), an extension of SGD.

3.4. Extraction of Input Influences

The extraction of *feature importance* provides a highly compressed, global insight into the behavior of the ANN. The importance measurement automatically considers all interactions with other features so that by changing the features, the interdependencies also have an influence. This means that the importance measurement must include both the main feature effect and interaction effects on the model performance. To determine the relevance of individual features in an ANN-based system, various methods are described in the literature (Garson, 1991; Olden et al., 2004; Gevrey et al., 2003). One hurdle is that the structure of more complex ANN, i.e., mainly of RNN structures, can no longer be described by methods that use the parameters of the network. Therefore the model must be stimulated with test signals and the output must be examined (this corresponds to the basic procedure of sensitivity analysis). Thus, the method described here thus represents an extension of robustness testing and leads to an evaluation of the variable influence in an ANN.

A distinction can be made between methods that are dedicated to the direct parameterization of the ANN and those that record the ANN as unknown in itself (*black box*) (Olden et al., 2004). Parameter-bound methods are limited to simple Multi Layer Perceptron (MLP) and are initially not considered because of the primary use of RNN. Therefore, algorithms based on the idea of sensitivity analyses are used, which act superordinate and observe and analyze the reaction of the system's outputs through various stimuli. Methods that refer to the weightings within the network and are therefore applicable to MLP are *Garson's algorithm* and the *connection weights method*. Garson derives hidden-output connection weights in components associated with each input neuron using the absolute values of the connection weights (see Garson (1991); Gevrey et al. (2003)). The *connection weights method* calculates the product of the raw, input-hidden, and hidden-output connection weights between each input neuron and output neuron, and sums the products over all hidden neurons (Olden and Jackson, 2002). Using *partial derivatives* of the ANN output in relation to the input neurons, a solution for determining the importance of inputs can also be derived for feed-forward paradigms (Dimopoulos et al., 1999).

For our research presented, we apply the *Permutation Importance* method (Breiman, 2001; Altmann et al., 2010) because this is an appropriate heuristic for recurrent structures with low computational effort and avoids the hurdles mentioned before. Permutation importance is one form of extracting knowledge of different kinds of neural networks, also of complex recurrent structures. As we not only use feed-forward networks, which could be explained by using parameter-based methods, we need a heuristic approach, which is computationally cheap and applicable to all kinds of neural networks. As a method to extract feature importance, permutation importance is one approach, which matches our requirements. We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature (implemented with MLxtend

python library (Raschka, 2018)). Since this is calculated after an ANN has been adjusted, neither the ANN nor the predictions obtained for a given Λ input quantity are changed. Instead, individual columns of validation data are randomly shuffled, with the remaining data remaining constant. Influencing the accuracy of the prediction determines the relevance of the feature. Random reordering of a single column should result in less accurate predictions since the resulting data will no longer match the sequences observed in M. Model accuracy suffers particularly when a column of a feature is shuffled, which has a high relevance to the ANN in predictions. For example, this is likely the case for traffic demand, where mixing will lead to inaccurate predictions. The process of feature extraction works as follows:

1. Preservation of the trained ANN.
2. Merging the values in a single column of a feature, making predictions with the resulting data set. Compare these predictions and the former predictions to calculate how much the loss function has suffered from the mixing. This performance degradation measures the importance of the feature.
3. Restoring the data to the original order (undoing the randomization of step 2.) Repeat step 2 with the next column in Λ until the meaning of each column has been calculated.

4. Model Application

We have implemented the neural networks described above in *Python 3.8.3* using the open-source deep learning library *Keras 2.4.0* (Chollet et al., 2015) (frontend) with open-source framework *TensorFlow 2.2.0* (Abadi et al., 2016) (backend), *Scikit-learn 0.23.1* (Pedregosa et al., 2011) (additional machine learning modules) and *Scipy 1.4.1* (Jones et al., 01) (routines for numerical integration and optimization). Training and testing were performed on GPU (NVIDIA Geforce 980 TI) using CUDA as a parallel computing platform and application programming interface. Similar experiments have also been carried out for regression applications, in which RNN models were used to predict delays (cf. Reitmann and Schultz (2018)).

Our investigations focus on the correlation between traffic demand, flight delays, and the impact of weather conditions at LGW. Four scenarios were defined to cover different aspects (see Tab. 6). The scenarios differ mainly in the choice of different weather conditions for filtering the data for training and testing (Λ_{train}, Λ_{test}). Thus, scenario B is additionally divided into two sub-scenarios to train and test *good* and *bad* weather days separately (using a decision value of 1.5 ATMAP weather score (Eurocontrol, 2011)), and scenario C consider days of similar traffic demand.

Table 6: Basic definitions of the scenarios within the applications.

Scenario	Description	Trained days Λ_{train} and predicted days Λ_{test}
A	no restrictions	Λ
B1	weather split	$\Lambda^{\text{bad weather}}$
B2	weather split	$\Lambda^{\text{good weather}}$
C	similar demand	similar traffic demand (DTW)

4.1. Data preparation

The first step in the application is to prepare the raw data for use in machine learning approaches. This includes the selection of features for input and output and the classification of output. As with clustering, there is also the option of solving the selection of features algorithmically or relying on expert knowledge. Feature selection for the following applications is made with expert knowledge.

All weather input features are numerically accessible factors of the METAR dataset (see Tab. 7). Airport performance is decisively determined by the relationship between demand and capacity, with weather events having a significant impact on airport capacity. An imbalance between the traffic demand (scheduled movements) and actual movements results in delays, which are added as an airport performance indicator. To provide an overview of the data, figure 8 comprises the mean values for arrival and departure delays of the observation period at LGW.

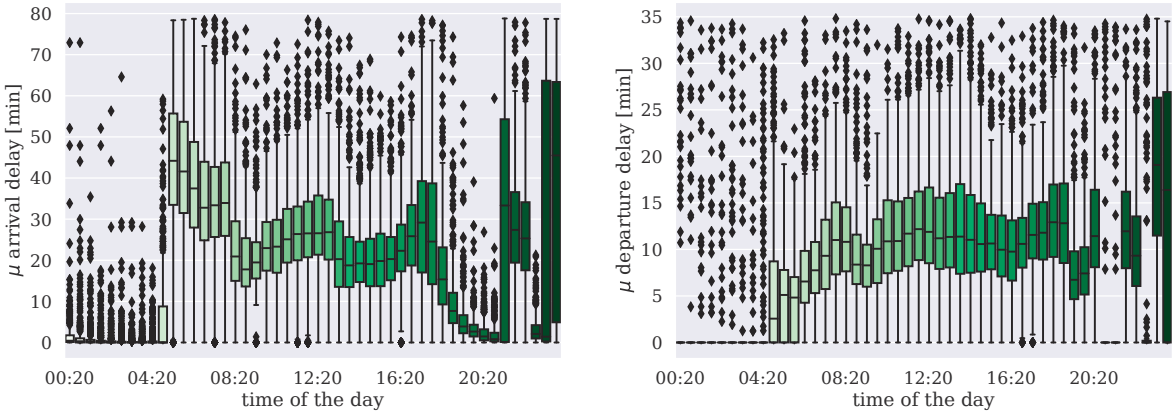

Figure 8: Boxplots of the average mean values (μ) for ARR (left) and DEP delays (right) for LGW in the 2014/2015 observation period.

Particularly noticeable are the large number of delay outliers in the early morning hours until about 04:20. In the same period, however, the demand is at a low level, so that mainly external backgrounds for the delay have to be assumed here, which could further complicate an appropriate modeling.

4.1.1. Feature selection

Qualitative descriptions of weather conditions have been removed from the recorded weather features at LGW and only raw data from the METAR messages were considered. In this way, a link to the output value is present, ensuring that only weather factors with appropriate relevance within Γ are considered. Variance threshold (see Sec. 3.2.1) is used for the algorithmic feature selection. Finally, the input features used in $\widetilde{\Xi}$ are shown in Tab. 7. The main goal of this parameter reduction is to avoid overfitting and improve the accuracy of the ANN. The operational conditions at the airport are described by the air traffic demand ($n_{arrivals\_scheduled}$, $n_{departures\_scheduled}$). Overall delay is used as output in multi-class classification and ARR/DEP delay is used for binary classification.

Table 7: Selection of feature sets $\widetilde{\Xi}$ to specify METAR and ATFM input data.

Set	Input features
METAR $\widetilde{\Xi}$	wind direction [°], wind speed, visibility, temperature, rain
ATFM $\widetilde{\Xi}$	traffic demand, mix of ARR-/DEP movements

Although the expert-based ATMAP approach also incorporates weather features, such as gusts or cloud cover, these features were not found to be significant for the LGW use case. Statistical analysis shows that only in 0.2% of the observations cloud ceiling information is not covered by information conducted from the indication of rain and the visibility ($>$ 1.5km). Furthermore, 93% of gusts occur

always in combination with rain, which means that an essential correlation is already ensured with the use of the rain indicator. At this point, our feature selection exhibits its potential to provide a focused solution for specific airports. However, it must also be said that the ATMAP algorithm was an attempt at a general, quantified assessment of weather phenomena concerning average operations at European airports. An appropriate comparison of these two different approaches is not really useful due to the different levels of detail and scope of application.

Both delay and the deviation of $n_{flights\_actual}$ from $n_{flights\_scheduled}$ can be used as output. It should be noted that the METAR data is provided at 30 min time intervals and the output values need to be mapped to these slots since a uniform database is indispensable for the learning process. Either the METAR data is mapped to a single flight event, or the events are aggregated to the 30 min slots of the weather data. Since we consider the constant 30 min time slots as an advantage in the learning process, we have decided on the second variant. As an aggregated value, the average absolute delay as the deviation from the scheduled time to the actual time is used. The deviation $\Delta flights = n_{flights\_scheduled} - n_{flights\_actual}$ is calculated absolutely per 30 min time slot.

For improved prediction and general estimation of effects within the ATFM, the prediction of delay values is transferred to a predictive classification, which is why the target values must be classified. The basic class reaches from -5 to 15 minutes, which is defined as *punctual* or *on-time* (Eurocontrol, 2007). This case is a binary classification[4]. The classifications over more than two classes are derived based on the distribution functions of the delay values. For the Γ period under consideration this is exhibited in Fig. 9 (left). It can be seen that for DEP^{Γ} the basic class $[-5, 15]$ covers a large part of the data. Thus $22,422$ usable data tuples are available, of which $18,365$ can be assigned to the basic class, which corresponds to 81.9% of the values.

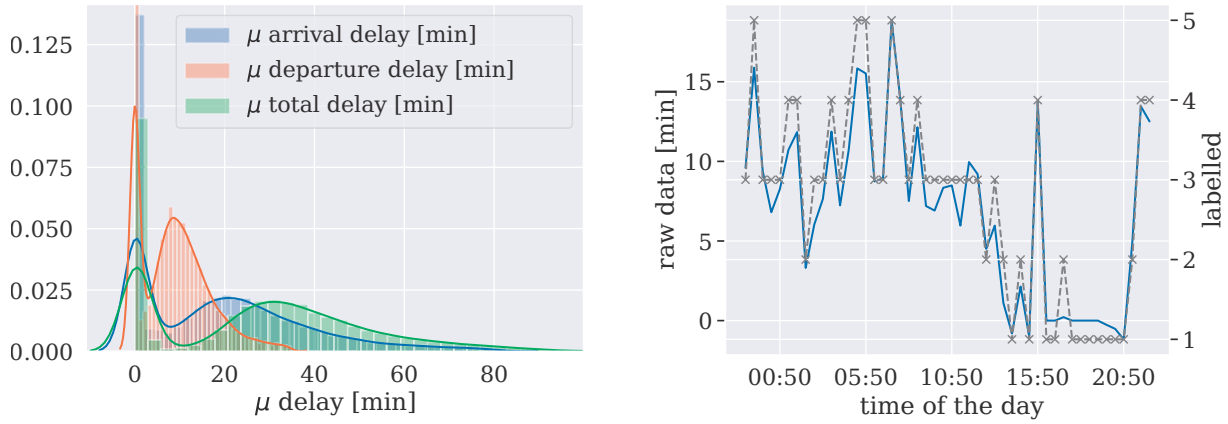

Figure 9: Distributions of average delay at LGW 2014/2015 (left) and exemplary delay categorization (blue - raw data, grey - categorized) (right).

It is desirable in the course of classification that the thickness of the categories does not differ significantly. For this reason, the thickness specified in Tab. 8 is used, which makes the experiment more granular, but takes into account the characteristics of the data set for ARR and DEP movements. For DEP this division corresponds to a 5 minute division within the basic class, as well as two boundary classes (≤ -5 and > 15), for ARR the interval steps vary. A granular classification is expected to be successful for DEP delay, while the binary classification supports the already difficult ARR classification. Both granular interval approaches still allow for the consideration of the basic

[4]The basic class is defined as punctual, all delay values outside as unpunctual. It is not considered whether the delay value is below the lower limit or above the upper limit of the interval, which results in two classes.

class (for ARR this corresponds to the labels 1 and 2, for DEP 1 to 4). The smallest interval in each case id $(-5, 0]$ and contains only delay values of 0 since negative delays are defined as zero delays in the LGW dataset. Fig. 9 (right) exhibits an exemplary section of classified data from 18.02.2014 to visualize the approximation of the original data by the applied classification. It can be seen that if the base class $(-5, 15)$ was used exclusively - with three exceptions - a uniform class value would be assigned. This corresponds to the overall picture of the data set and underlines the necessity of a differentiated division of class boundaries.

Table 8: Categorization intervals and thicknesses of the data sets for the airport LGW in the limited observation period $\Gamma = [04/2014, 08/2015]$ for multi-class (left) and binary classification (right). It should be mentioned that there is an imbalance between ARR and DEP. During the study period, 1056 more ARR than DEP movements were recorded.

[min]	$\lvert DEP^{\Gamma}\rvert$	Label	$\lvert ARR^{\Gamma}\rvert$	**[min]**
$(-5, 0]$	5,663	1	3,916	$(-5, 0]$
$(0, 5]$	2,067	2	5,389	$(0, 15]$
$(5, 10]$	5,866	3	4,600	$(15, 25]$
$(10, 15]$	4,769	4	4,560	$(25, 40]$
$(15, \infty)$	4,057	5	2,850	$(40, \infty)$
	22,422	$\sum$	23,478	

[min]	$\lvert DEP^{\Gamma}\rvert$	$\lvert ARR^{\Gamma}\rvert$	Label
$(-5, 15]$	18,365	11,098	0
$(15, \infty)$	4,057	12,380	1
$\sum$	22,422	23,478	

4.2. Neural network setup

The structure of ANN for classification differs essentially from models of regression (Reitmann and Schultz, 2018). This is also because other paradigms (CNN) and hybrid structures are used in our experiments. Thus, CNN-based networks have proven to be advantageous for classification. However, concerning the data basis and the results achieved in previous experiments, the use of recurrent structures is advantageous for the interdependencies of the characteristics. For this reason, besides the pure CNN and LSTM, the hybrid structures of a CNN-LSTM and a Convolutional LSTM (ConvLSTM) are used in the following. Tab. 9 comprises the structure of the ANN for the classification (left), as well as the set sets of hyperparameters (right), from which optimal combinations for the networks are derived using grid search.

Table 9: Structure of the ANN (left) and hyperparameter for the grid search (right) of the classification.

	LSTM	Conv1D	ConvLSTM2D	Dropout	MaxPooling1D	Flatten	Dense
LSTM	x			x			x
CNN		x		x	x	x	x
CNN - LSTM	x	x		x	x	x	x
ConvLSTM			x	x		x	x

Hyperparameter	Range of values
batch value	[20,40,60,80,100]
n_{epochs}	[10,100]
optimizer	[Adam,Adamax]
learning rate η	[0.001, 0.01, 0.1]
momentum ϱ	[0.0, 0.2, 0.4, 0.6, 0.8, 0.9]
U_{hidden}	[10, 100, 500, 1,000]
dropout rate	[0.0, 0.2, 0.4, 0.6, 0.8, 0.9]

The CNN-based folded networks consist of Conv1D and MaxPooling1D layers arranged in a stack with the required depth. Conv1D layers interpret snapshots of the data basis, and the pooling layers condense/abstract this interpretation. This is how the ANNs from Tab. 9 (left) transform the input data into its own vector or matrix interpretations. MaxPooling bundles these interpretations and reduces the output. The flattening layer takes the reduced output and transforms it into a vector

that leads to the output via via a dense layer. While the CNN-LSTM is a combination of two separate layers, the ConvLSTM consists of a special ConvLSTM2D layer. This is a recurrent layer, just like the LSTM, but internal matrix multiplications are exchanged with convolution operations.

4.3. Model fitting and evaluation

The ANN for the classification consists of a complex structure with several layers and a high number of hidden neurons (U_{hidden}). *Adam* and its variant *Adamax* are the best performing optimizers in the current context. The results of the grid search for the hyperparameter optimization are shown in Tab. 10.

Table 10: Hyperparameter for the ANN application in the classification.

	ANN	**Hyperparameter**
config I	LSTM	Batch = 40, $n_{epochs} = 50$, optimizer = Adamax, f_{act} = softmax, dropout = 0.4, $U_{hidden} = 200$
config II	CNN	Filters= 64, Kernel= 3, MaxPool= 2, $U^{dense}_{hidden} = 10$
config III	CNN-LSTM	$U^{lstm}_{hidden} = 100$, Filters= 64, Kernel= 3, MaxPool= 2, $U^{dense}_{hidden} = 100$
config IV	ConvLSTM	Filters= 64, Kernel= 3, Batch = 40, $n_{epochs} = 50$, f_{act} = ReLU, optimizer = Adamax, dropout = 0.4, $U_{hidden} = 500$

The sparse categorical cross-entropy (multi-class) and the binary cross-entropy (binary classes) are used as cost functions for the classification. Tab. 11 contains the accuracy values for training and validation data for the classified output data. The classification is applied specifically to scenario B (weather dependency) and refers to the delay of ARR and DEP movements. The best results from Tab. 11 are achieved by using the CNN, thus confirming the assumption that this paradigm is appropriate for classification. Nevertheless, the hybrid structures that also contain CNN components are subject to the pure LSTM (with the exception of DEP in B2), which overall performs only slightly worse than the CNN on departure delays.

Table 11: Accuracy of the classification of LSTM, CNN, CNN-LSTM, and ConvLSTM for scenarios B1 (bad weather) and B2 (good weather), differentiated by ARR and DEP on the basis of Λ_{valid}.

	Scenario B1		**Scenario B2**		**B1 (binary)**	**B2 (binary)**
	ARR	*DEP*	*ARR*	*DEP*	*ARR*	*ARR*
LSTM	48.9%	89.6%	39.1%	82.1%	88.3%	96.1%
CNN	55.0%	96.9%	43.3%	90.1%	90.0%	96.8%
CNN-LSTM	30.9%	74.3%	28.2%	72.6%	83.2%	83.4%
ConvLSTM	45.8%	88.2%	41.3%	87.4%	87.2%	91.5%

It should be noted that the validation accuracies from Tab. 11 include the percentage of the correctly determined classes (label) for a subsequent time step. This is used to estimate the ANN about the assignment of delay classes in general but is not an estimate of the predictive classification up to this point. A specific investigation of a prediction of the classes over several time steps is part of the following section.

4.4. Model summary

Tab. 12 comprises the forecast quality of the predictive classification for scenario B2 (good weather). The percentage values refer to the proportion of correctly determined delay classes in the respective time slots. Again, the figures refer to the μ of repeated forecasts, with σ not specified.

Table 12: Accuracy of the classification of the LSTM for scenario B2 (good weather) on the basis of Λ_{test}.

Delay	Error	Accuracy_{t+1h}	...	Accuracy_{t+6h}
ARR	same class	37.4%		25.8%
	$\triangle \pm 1$ class	69.6%		58.8%
$\text{ARR}_{\text{binary}}$	same class	95.3%		72.1%
DEP	same class	63.2%		50.7%
	$\triangle \pm 1$ class	93.0%		89.9%

It can be seen that the forecast confirms the positive validation results of Tab. 11. Especially for ARR, the reduction of the output information can improve the overall results. Thus, the binary classification can be used to determine with 95.3% accuracy within a forecast period of one hour whether arrivals are on time or not according to weather and demand. For a forecast period of 6 hours, this accuracy is still 72.1%. For the multi-class forecast, the accuracy decreases so that at the maximum forecast time of 6 hours with 58.8% accuracy, the ARR delay can be predicted with a deviation of up to one class. In the case of DEP, the delay prediction after 6 hours is accurate to 89.9% - also with a maximum deviation of one class, but smaller class widths (5 minutes, cf. Tab. 8). Nevertheless, it was possible to qualitatively predict ARR delays even based on a less suitable data set.

The results from Tab. 11 show that there are several appropriate model solutions for different prerequisites, data, and configurations. The decision to use a particular depends largely on the available data and the level of detail of the investigation. Our results show that a transfer of weather events and performance indicators to a classified delay is useful. It should be mentioned here that extracting knowledge from the neural network can provide added value. Suitable approaches exist for RNNs and CNNs. Even though delays measured at airports are a product of multiple dependent inputs, our results highlight that mapping delays and weather conditions through machine learning approaches can lead to appropriate prediction. However, it remains a challenge to determine the contribution of each meteorological component, as is the case with the expert-based ATMAP approach.

The trained network can be used as an adaptive decision support tool for operators, considering local airport environment conditions such as weather data and traffic demand. While the traffic demand is anchored in the flight plan by scheduled/estimated times for arrivals and departures and offers a wide forecast horizon, the weather forecast is currently limited by the time horizon of the TAF reports. This amounts to a period of 6 hrs and means a forecast horizon of 12 time slots with a width of 30 min. This is a multi-step rather than a single-step prediction, which means that the real value does not update the value predicted by the model over the entire forecast horizon. Incorrectly predicted labels are carried along accordingly. Thus, the input values used to determine the delay label are related to the window size.

A brief, representative example (Tab. 13 and Fig. 10) refers to flight plan data at LGW on 3^{rd} September 2014. A 6-step prediction (3 hours) is performed after a start pulse of 50 slots. The values of Tab. 13 represent the labels of the applied intervals.

The ground truth comprises the real data of LGW (labeled). The numbers in the lines of the applied ANN types represent the predicted label by each network. Falsely predicted labels are underlined. The advantage of neural networks, especially recurrent paradigms, is the integration of parallel or past knowledge. Thus, errors can occur in a prediction step, but they do not lead to a continuation of these erroneous predictions. Tab. 13 shows that LSTM, CNN-LSTM, and ConvLSTM each have one or two prediction errors.

Fig. 10 depicts the underlying labeled dataset and a comparison to the actual labels observed

Table 13: Label prediction for 3^{rd} September 2014.

	Time slot					
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$
ground truth	4	4	1	2	2	2
LSTM	4	4	1	*1*	2	2
CNN	4	4	1	2	2	2
CNN-LSTM	4	*2*	*2*	2	2	2
ConvLSTM	4	4	1	2	2	*1*
	Label					

at LGW. The data describe the day already presented in Fig. 5 (left), having 2 hours of fog in the early morning, with the data already labeled and assigned to the slots of the weather data.

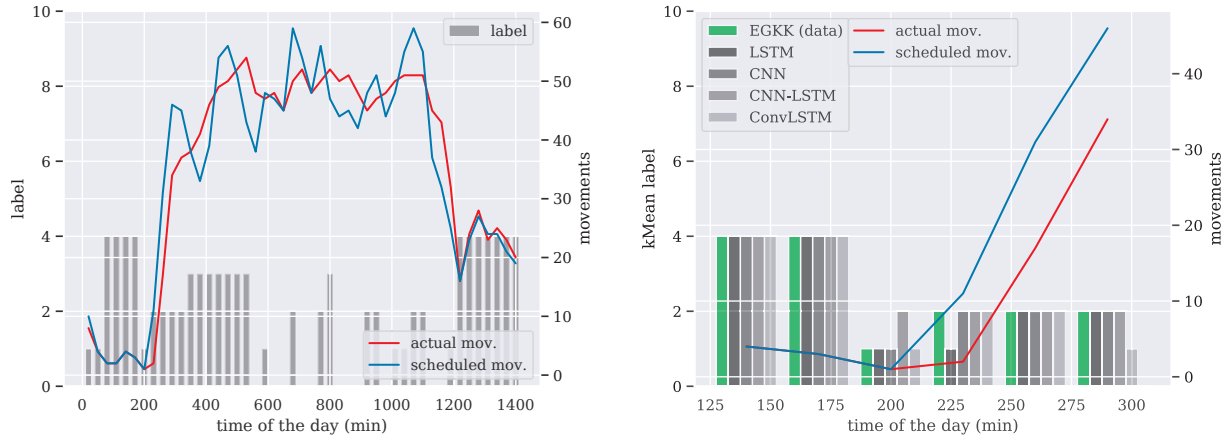

Figure 10: 3^{rd} September 2014, labelled and slotted (left), 6 slot prediction of Tab. 13 (right).

4.5. Operational application - evaluation of input importance

For operational implementation of our approach, appropriate weather forecasts must be available. Depending on the geographical location, different forecast periods must also be taken into account here. In some circumstances, responding to weather events 24 hours in advance may not appear to be an appropriate strategy to efficiently manage the capacity of an airport. The pre-tactical phase (24 - 3 h before operations) seems to be more suitable to react to forecasted local weather conditions especially when a weather phenomenon has already been active for several days.

Operators can benefit from our modeling approach in the tactical phase of the operation (from 3 h before the operation). In an airport with an implemented A-CDM environment, all stakeholders could consider the predictions to derive appropriate adjustments. Specific control measures could then be jointly determined on the basis of the actual data in order to adjust air traffic demand to the given airport capacity. This includes the allocation of operational slots (time windows) for upcoming flights. Flights can be re-routed to counteract congestion and use alternative flight profiles.

Extracting knowledge about dependencies is quite a challenging task. That's why we developed an adaptive model, which learns relationships, and we apply permutation importance to bring them out. There are two possibilities for the extended knowledge extraction from the trained ANN. Starting with the determination of the feature influences, an image can be obtained, which provides the operators with the relevance for considering certain weather features. This is done exemplary

for the scenarios B1 and B2 at LGW (Fig. 11). Depending on the weather conditions (e.g., good weather (ATMAP score ≤ 1.5) and bad weather (ATMAP score > 1.5), cf. Schultz et al. (2018)), different features are important to improve the situational awareness of the operators. Scheduled Number of Departures (SND) and Scheduled Number of Arrivals (SNA) are used to reflect the traffic conditions at the airport.

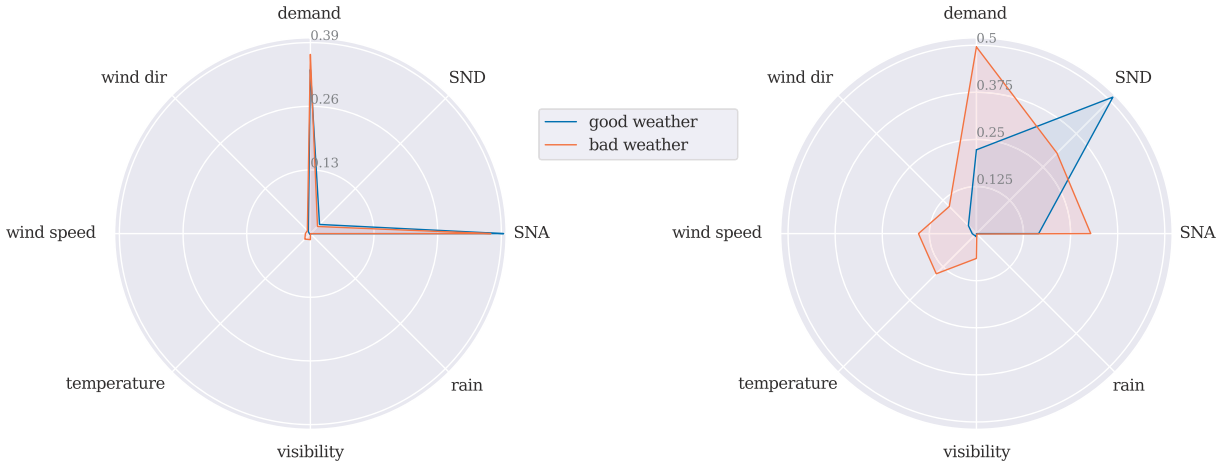

Figure 11: Relative importance of the features used for ARR (left) and DEP delays (right) at LGW.

The picture for LGW shows that the integration of weather data is possible to a great extent. On the one hand, this is due to a granular observation by METAR, on the other hand to the high sensitivity of the airport to meteorological events. For DEP delays a focus on wind and temperature can be identified, while rain has no significant influence.

5. Conclusion

We investigated a quantification of the influence of meteorological conditions on the individual airport performance at London–Gatwick Airport (LGW) using machine learning approaches. Different neural networks were used and combined to process the corresponding data foundation in a target-oriented way. The accuracy of the trained networks was compared and the networks were exemplarily applied.

Our contribution differs from the mechanism introduced by the expert-based ATMAP algorithm (Eurocontrol, 2011) because we follow an effect-to-cause relation. Based on higher-level, aggregate performance measures at the airport (e.g., delayed operations), machine learning methods were used to infer the underlying weather data. The aggregation was done by unsupervised learning and classification by supervised learning. In particular, the pure paradigms of LSTM and CNN show reasonable results and could provide weather-related decision support for future airport operations. Both the data and the ANN were processed and adjusted accordingly to provide optimal results for the model application. The structure of the networks was implemented logically, the parameterization was determined algorithmically, and the learning process was validated.

We have succeeded in creating models that can make valid predictions of the delay in a time horizon of 6 hours by classifying the quality of the predictions. The models actively consider local weather conditions at the airport and form forecasts that consider both operational and meteorological factors. We have also succeeded in extracting the importance of each input factor in the forecast model by applying appropriate algorithms, thus making a significant contribution to the application of complex ANN in time series forecasting. Our model is limited to the assumption we

make, such as neglecting unexpected network delays that are not covered by regular patterns over the day of operations. Therefor, we assume that a more comprehensive, network-based approach could identify additional correlations.

The application presented in Sec. 4.5 provides a first example of a potential use case of the developed model. In further investigations we want to find out how the trained knowledge can be used to derive decision support for local operators (e.g., airline, airport, ground handling agencies, air navigation service providers) can be derived from the trained knowledge. In the context of a collaborative decision-making at the airport, an optimal adaptation of the actual traffic to the expected meteorological conditions would help minimize the overall delay. The effects of individual METAR components should also be quantified, but this should be considered a separate research task due to the complex interactions between weather components.

We want to continue our research towards networks of connected airports, as arrival delays trigger reaction delays and propagate through the network. That is why we will study airports and airport clusters in more detail with respect to their specific weather dependencies.

Glossary

References

Eurocontrol, Performance Review Report – An Assessment of Air Traffic Management in Europe during the Calendar Year 2019, Technical Report, Performance Review Commission, 2020.

C. Bongiorno, G. Gurtner, F. Lillo, R. N. Mantegna, S. Micciché, Statistical characterization of deviations from planned flight trajectories in air traffic management, Journal of Air Transport Management 58 (2017) 152–163. doi:10.1016/j.jairtraman.2016.10.009.

J. Bronsvoort, G. McDonald, R. Porteous, E. Gutt, Study of aircraft derived temporal prediction accuracy using FANS, in: Proceedings of the 13th ATRS World Conference, 2009.

E. Mueller, G. Chatterji, Analysis of aircraft arrival and departure delay, in: Proceedings of the AIAA ATIO Conference, 2002.

Eurocontrol, Performance Review Report – An Assessment of Air Traffic Management in Europe During the Calendar Year 2014, 2015, 2016, Technical Report, Performance Review Commission, 2017.

Eurocontrol, CODA Digest All-Causes Delay and Cancellations to Air Transport in Europe – 2016, Technical Report, CODA, 2016.

M. Tielrooij, M. C. Borst, M. van Paassen, M. Mulder, Predicting arrival time uncertainty from actual flight information, in: 11$^{\text{th}}$ USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), 2015.

S. Helm, S. Loth, M. Schultz, Advancing Total Airport Management- An Introduction of Performance Based Management in the Airport Context, in: 19th Air Transport Research Society World Conference, Singapore, 2015.

J. Rosenow, M. Schultz, COUPLING OF TURNAROUND AND TRAJECTORY OPTIMIZATION BASED ON DELAY COST, in: 2018 Winter Simulation Conference (WSC), 2018, pp. 2273–2284. doi:10.1109/WSC.2018.8632250, iSSN: 1558-4305.

H. Ali, Y. Guleria, S. Alam, M. Schultz, A Passenger-Centric Model for Reducing Missed Connections at Low Cost Airports With Gates Reassignment, IEEE Access 7 (2019) 179429–179444. doi:10.1109/ACCESS.2019.2953769, conference Name: IEEE Access.

A. Cook, G. Tanner, P. Enaud, Quantifying airline delay costs - the balance between strategic and tactical costs, in: 14$^{\text{th}}$ Air Transport Research Society (ATRS) World Conference, 2010.

M. G. Sohoni, S. Erat, Can time buffers lead to delays? the role of operational flexibility, SSRN (2015). doi:10.2139/ssrn.2572801.

L. Cappelleras, Additional Asma Time Performance Indicator Document, 00-06 Eurocontrol/PRU, 2015.

M. Schultz, J. Rosenow, X. Olive, A-CDM lite: situation awareness and decision making for small airports based on ADS-B data, in: 9th Eurocontrol SESAR Innovation Days, 2019.

Eurocontrol Airport CDM Team, Airport CDM implementation manual, ver. 5, 2017.

Eurocontrol Network Manager, Advanced atc twr implementation guide, 2020.

M. Schultz, S. Lorenz, R. Schmitz, L. Delgado, Weather impact on airport performance, Aerospace 5 (2018). doi:10.3390/aerospace5040109.

Performance Review Commission, ATM Airport Performance (ATMAP) Framework, 2009.

I. Gerdes, A. Temme, M. Schultz, Dynamic airspace sectorisation for flight-centric operations, Transportation Resarch Part C: Emerging Technologies 95 (2018) 460–480.

T. Standfuss, I. Gerdes, A. Temme, M. Schultz, Dynamic airspace optimisation, CEAS Aeronatuical 9 (3) (2018) 517–531.

J. Rosenow, M. Lindner, H. Fricke, Impact of climate costs on airline network and trajectory optimization: a parametric study, CEAS Aeronautical Journal 8 (2) (2017) 371–384.

M. Niklaß, B. Lührs, V. Grewe, K. Dahlmann, T. Luchkova, F. Linke, V. Gollnick, Potential to reduce the climate impact of aviation by climate restricted airspaces, Transport Policy (2017) In Press.

B. F. Santos, M. M. E. C. Wormer, T. A. O. Achola, R. Curran, Airline delay management problem with airport capacity constraints and priority decisions, Journal of Air Transport Management 63 (2017) 34–44. doi:10.1016/j.jairtraman.2017.05.003.

Ingrid Gerdes, A. Temme, M. Schultz, Dynamic Airspace Sectorization using Controller Task Load, Delft, Netherland, 2016.

M. Kaiser, J. Rosenow, H. Fricke, M. Schultz, Tradeoff between optimum altitude and contrail layer to ensure maximum ecological en-route performance using the enhanced trajectory prediction model, in: 2nd International Conference on Application and Theory of Automation in Command and Control Systems (ATACCS), 2012.

S. Carlier, I. de Lépinay, J. Hustache, F. Jelinek, Environmental impact of air traffic flow management delays, in: 7th USA/Europe Air Traffic Management Research and Development Seminar (ATM2007), 2007.

I. Gerdes, A. Temme, M. Schultz, From free-route air traffic to an adapted dynamic main-flow system, Transportation Research Part C: Emerging Technologies 115 (2020) 102633. doi:10.1016/j.trc.2020.102633.

J. Rosenow, H. Fricke, T. Luchkova, M. Schultz, Impact of optimised trajectories on air traffic flow management, The Aeronautical Journal 123 (2019) 157–173.

L. Zhou, Z. Chen, Measuring the performance of airport resilience to severe weather events, Transportation Research Part D: Transport and Environment 83 (2020) 102362.

S. A. Markolf, C. Hoehne, A. Fraser, M. V. Chester, B. S. Underwood, Transportation resilience to climate change and extreme weather events – Beyond risk and robustness, Transport Policy 74 (2019) 174–186. doi:10.1016/j.tranpol.2018.11.003.

M. Taszarek, S. Kendzierski, N. Pilguj, Hazardous weather affecting European airports: Climatological estimates of situations with limited visibility, thunderstorm, low-level wind shear and snowfall from ERA5, Weather and Climate Extremes 28 (2020) 100243. doi:10.1016/j.wace.2020.100243.

R. Burbidge, Adapting European Airports to a Changing Climate, Transportation Research Procedia 14 (2016) 14–23. doi:10.1016/j.trpro.2016.05.036.

I. Stamos, E. Mitsakis, J. M. Salanova, G. Aifadopoulou, Impact assessment of extreme weather events on transport networks: A data-driven approach, Transportation Research Part D: Transport and Environment 34 (2015) 168–178. doi:10.1016/j.trd.2014.11.002.

A. Islami, M. Sun, S. Chaimatanan, D. Delahaye, Optimization of military missions impact on civilian 4D trajectories, in: ENRI International Workshop on ATM/CNS (EIWAC 2017), 2017.

T. Luchkova, R. Vujasinovic, A. Lau, M. Schultz, Analysis of impacts an eruption of volcano Stromboli could have on european air traffic, in: USA/Europe ATM R&D Seminar (11th ATM Seminar), 2015.

M. Kreuz, T. Luchkova, M. Schultz, Effect of restricted airspace on the ATM system, in: WCTR Conference, 2016.

J. Rosenow, H. Fricke, M. Schultz, Air traffic simulation with 4D multi-criteria optimized trajectories, in: Winter Simulation Conference, 2017, p. 2589–2600.

J. Rosenow, H. Fricke, T. Luchkova, M. Schultz, Minimizing contrail formation by rerouting around dynamic ice-supersaturated regions, Aeronautics and Aerospace Open Access Journal 2 (3) (2018) 105–111.

S. Kaltenhaeuser, F. Morlang, T. Luchkova, J. Hampe, M. Sippel, Facilitating sustainable commercial space transportation through an efficient integration into air traffic management, New Space 5 (4) (2017) 244–256.

E. Sunil, J. Hoekstra, J. Ellerbroek, F. Bussink, D. Nieuwenhuisen, A. Vidosavljevic, S. Kern, Metropolis: Relating airspace structure and capacity for extreme traffic densities, in: USA/Europe ATM R&D Seminar (11th ATM Seminar), 2015.

M. Schultz, I. Gerdes, T. Standfuß, A. Temme, Future Airspace Design by Dynamic Sectorization, in: Electronic Navigation Research Institute (Ed.), Air Traffic Management and Systems III, Lecture Notes in Electrical Engineering, Springer, Singapore, 2019, pp. 19–34.

A. Montlaur, L. Delgado, Flight and passenger delay assignment optimization strategies, Transportation Research Part C: Emerging Technologies 81 (2017) 99–117. doi:10.1016/j.trc.2017.05.011.

L. Delgado, X. Prats, Operating cost based cruise speed reduction for ground delay programs: Effect of scope length, Transportation Research Part C: Emerging Technologies 48 (2014) 437–452. doi:10.1016/j.trc.2014.09.015.

G. Buxi, M. Hansen, Generating day-of-operation probabilistic capacity scenarios from weather forecasts, Transportation Research Part C: Emerging Technologies 33 (2013) 153–166. doi:10.1016/j.trc.2012.12.006.

B. Campanelli, P. Fleurquin, A. Arranz, I. Etxebarria, C. Ciruelos, V. M. Eguíluz, J. J. Ramasco, Comparing the modeling of delay propagation in the US and European air traffic networks, Journal of Air Transport Management 56 (2016) 12–18. doi:10.1016/j.jairtraman.2016.03.017.

N. Ivanov, F. Netjasov, R. Jovanović, S. Starita, A. Strauss, Air traffic flow management slot allocation to minimize propagated delay and improve airport slot adherence, Transportation Research Part A: Policy and Practice 95 (2017) 183–197. doi:10.1016/j.tra.2016.11.010.

B. Baspinar, N. K. Ure, E. Koyuncu, G. Inalhan, Analysis of Delay Characteristics of European Air Traffic through a Data-Driven Airport-Centric Queuing Network Model, IFAC-PapersOnLine 49 (2016) 359–364. doi:10.1016/j.ifacol.2016.07.060.

M. N. Postorino, L. Mantecchini, C. Malandri, F. Paganelli, A methodological framework to evaluate the impact of disruptions on airport turnaround operations: A case study, Case Studies on Transport Policy 8 (2020) 429–439. doi:10.1016/j.cstp.2020.03.007.

S. Borsky, C. Unterberger, Bad weather and flight delays: The impact of sudden and slow onset weather events, Economics of Transportation 18 (2019) 10–26. doi:10.1016/j.ecotra.2019.02.002.

M. Schultz, H. Fricke, J. M. T.Kunze, J. L. Leonés, C. Grabow, J. D. Prins, M. Wimmer, P. Kappertz, Uncertainty Handling and Trajectory Synchronization for the Automated Arrival Management, in: 2nd Eurocontrol SESAR Innovation Days, Braunschweig, 2012.

S. Förster, M. Schultz, H. Fricke, Probabilistic prediction of separation buffer to compensate for the closing effect on final approach, Aerospace 8 (2021). doi:10.3390/aerospace8020029.

R. Merkert, L. Mangia, Management of airports in extreme winter conditions—some lessons from analysing the efficiency of Norwegian airports, Research in Transportation Business & Management 4 (2012) 53–60. doi:10.1016/j.rtbm.2012.06.004.

A. Cook, L. Delgado, G. Tanner, S. Cristóbal, Measuring the cost of resilience, Journal of Air Transport Management 56 (2016) 38–47. doi:10.1016/j.jairtraman.2016.02.007.

S.-L. Proag, V. Proag, The Cost Benefit Analysis of Providing Resilience, Procedia Economics and Finance 18 (2014) 361–368. doi:10.1016/S2212-5671(14)00951-4.

A. Cook, G. Tanner, V. Williams, G. Meise, Dynamic cost indexing – Managing airline delay costs, Journal of Air Transport Management 15 (2009) 26–35. doi:10.1016/j.jairtraman.2008.07.001.

R. Arnaldo Scarpel, L. C. Pelicioni, A data analytics approach for anticipating congested days at the São Paulo International Airport, Journal of Air Transport Management 72 (2018) 1–10. doi:10.1016/j.jairtraman.2018.07.002.

R. Henriques, I. Feiteira, Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport, Procedia Computer Science 138 (2018) 638–645. doi:10.1016/j.procs.2018.10.085.

P. Rozas Larraondo, I. Inza, J. A. Lozano, A system for airport weather forecasting based on circular regression trees, Environmental Modelling & Software 100 (2018) 24–32. doi:10.1016/j.envsoft.2017.11.004.

M. Schultz, S. Reitmann, S. Alam, Classification of Weather Impacts on Airport Operations, in: 2019 Winter Simulation Conference (WSC), 2019, pp. 500–511. doi:10.1109/WSC40007.2019.9004915, iSSN: 1558-4305.

S. Reitmann, M. Schultz, Computation of Air Traffic Flow Management Performance with Long Short-Term Memories Considering Weather Impact, in: V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2018, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 532–541.

F. Herrema, R. Curran, S. Hartjes, M. Ellejmi, S. Bancroft, M. Schultz, A machine learning model to predict runway exit at Vienna airport, Transportation Research Part E: Logistics and Transportation Review 131 (2019) 329–342. doi:10.1016/j.tre.2019.10.002.

H. Ming, M. Wei, M. Wang, L. Gao, L. Chen, X. Wang, Analysis of fog at Xianyang Airport based on multi-source ground-based detection data, Atmospheric Research 220 (2019) 34–45. doi:10.1016/j.atmosres.2019.01.012.

S. Fernández-González, P. Bolgiani, J. Fernández-Villares, P. González, A. García-Gil, J. C. Suárez, A. Merino, Forecasting of poor visibility episodes in the vicinity of Tenerife Norte Airport, Atmospheric Research 223 (2019) 49–59. doi:10.1016/j.atmosres.2019.03.012.

A. G. Salman, Y. Heryadi, E. Abdurahman, W. Suparta, Single Layer & Multi-layer Long Short-Term Memory (LSTM) Model with Intermediate Variables for Weather Forecasting, Procedia Computer Science 135 (2018) 89–98. doi:10.1016/j.procs.2018.08.153.

M. Bagamanova, M. M. Mota, A multi-objective optimization with a delay-aware component for airport stand allocation, Journal of Air Transport Management 83 (2020) 101757. doi:10.1016/j.jairtraman.2019.101757.

D. Serhan, S. W. Yoon, S. H. Chung, Dynamic reconfiguration of terminal airspace during convective weather: Robust optimization and conditional value-at-risk approaches, Computers & Industrial Engineering 132 (2019) 333–347. doi:10.1016/j.cie.2019.04.010.

Eurocontrol, Algorithm to describe weather conditions at European airports, Technical Report, Performance Review Unit, 2011.

S. O'Flynn, Airport Capacity Assessment Methodology - ACAM Manual, Technical Report 1.1, Network Manager, Eurocontrol, 2016.

M. Schultz, H. Fricke, Managing passenger handling at airport terminal, in: 9th USA/Europe Air Traffic Management Research and Development Seminar (ATM2011), 2011.

J. Evler, M. Schultz, H. Fricke, A. Cook, Development of stochastic delay cost functions, in: 10th Eurocontrol SESAR Innovation Days, 2020.

F. A. Administration, Advisory Circular 00-45 – Aviation Weather Services, Technical Report, Federal Aviation Administration, 2016.

C. Cortes, V. Vapnik, Support-Vector Networks, in: Machine Learning, 1995, pp. 273–297.

I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, Cambridge, Massachusetts, 2016.

M. Hagen, H. Demuth, M. Beale, O. D. Jesus, Neural Network Design, 2. ed., 2014.

S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780.

J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep Learning for Sensor-based Activity Recognition: A Survey, Pattern Recognition Letters (2018). ArXiv: 1707.03502.

V. Reitmann, Reguläre und chaotische Dynamik, Springer-Verlag, 2013.

R. Kruse, C. Borgelt, C. Braune, F. Klawonn, C. Moewes, M. Steinbrecher, Computational Intelligence: Eine methodische Einführung in Künstliche Neuronale Netze, Evolutionäre Algorithmen, Fuzzy-Systeme und Bayes-Netze, Computational Intelligence, 2 ed., Springer Vieweg, 2015. URL: https://www.springer.com/de/book/9783658109035.

Eurocontrol, A Matter of Time: Air Traffic Delay in Europe, Technical Report, 2007.

S. Reitmann, M. Schultz, S. Alam, Advanced quantification of weather impact on air traffic management, Air Traffic Management Research and Development Seminar (ATM2019), Wien, Österreich, 2019.

M. Kuhn, K. Johnson, Applied predictive modeling, volume 26, Springer, 2013.

C. M. Bishop, Pattern Recognition and Machine Learning, 1st ed. 2006. corr. 2nd printing 2011 ed., Springer, New York, 2007.

S. Gorripaty, Finding Similar Days for Air Traffic Management, Ph.D. thesis, UC Berkeley, 2017.

M. Müller, Dynamic Time Warping, in: Information Retrieval for Music and Motion, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 69–84.

S. Lawrence, C. Giles, A. Tsoi, What size neural network gives optimal generalization? convergence properties of backpropagation (2001).

J. Heaton, Introduction to Neural Networks for Java, 2. ed., Heaton Research, Inc., 2008.

I. Sutskever, J. Martens, G. E. Dahl, G. E. Hinton, On the importance of initialization and momentum in deep learning., in: ICML (3), volume 28 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2013, pp. 1139–1147.

G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, volume 112, Springer, 2013.

M. Claesen, B. D. Moor, Hyperparameter search in machine learning., CoRR abs/1502.02127 (2015).

J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

S. Ruder, An overview of gradient descent optimization algorithms, arXiv:1609.04747 [cs] (2016). URL: http://arxiv.org/abs/1609.04747, arXiv: 1609.04747.

D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs] (2014). URL: http://arxiv.org/abs/1412.6980, arXiv: 1412.6980.

G. D. Garson, Interpreting neural-network connection weights, Artif. Intell. Expert 6 (1991) 47–51.

J. D. Olden, M. K. Joy, R. G. Death, An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, Ecological Modelling 178 (2004) 389–397.

M. Gevrey, I. Dimopoulos, S. Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecological modelling 160 (2003) 249–264.

J. Olden, D. Jackson, Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks, Ecological Modelling 154 (2002) 135–150. doi:10.1016/S0304-3800(02)00064-9.

I. Dimopoulos, J. Chronopoulos, A. Chronopoulou-Sereli, S. Lek, Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece), Ecological Modelling 120 (1999) 157–165. doi:10.1016/S0304-3800(99)00099-X.

L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347.

S. Raschka, Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack, The Journal of Open Source Software 3 (2018). URL: http://joss.theoj.org/papers/10.21105/joss.00638. doi:10.21105/joss.00638.

F. Chollet, et al., Keras, https://keras.io, 2015.

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (2011) 2825–2830.

E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open source scientific tools for Python, 2001–. URL: http://www.scipy.org/, [Online; accessed 26.09.2020].