IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 17, 2022, accepted March 2, 2022, date of publication March 10, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158313

Assessing Identifiability in Airport Delay Propagation Roles Through Deep Learning Classification

ILINKA IVANOSKA[1], LUISINA PASTORINO[2], AND MASSIMILIANO ZANIN[2]
[1]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia
[2]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain

Corresponding author: Massimiliano Zanin (massimiliano.zanin@gmail.com)

This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Agreement 851255, and in part by the Agencia Estatal de Investigación (AEI, MCI, Spain) and Fondo Europeo de Desarrollo Regional (FEDER, UE) through the Maria de Maeztu Program for units of Excellence in Research and Development under Grant MDM-2017-0711.

ABSTRACT Delays in air transport can be seen as the result of two independent contributions, respectively stemming from the local dynamics of each airport and from a global propagation process; yet, assessing the relative importance of these two aspects in the final behaviour of the system is a challenging task. We here propose the use of the score obtained in a classification task, performed over vectors representing the profiles of delays at each airport, as a way of assessing their identifiability. We show how Deep Learning models are able to recognise airports with high precision, thus suggesting that delays are defined more by the characteristics of each airport than by the global network effects. This identifiability is higher for large and highly connected airports, constant through years, but modulated by season and geographical location. We finally discuss some operational implications of this approach.

INDEX TERMS Air transport, airport identifiability, delays, deep learning.

I. INTRODUCTION

Air transport delays, and specifically the phenomenon of delay propagation, is one of the most important research topics in air transport management, due to delays' profound implications in the cost-efficiency [1] and safety of the system [2], as well as their contribution to the negative impact of air transport on the environment [3]. Research works on air transport delays can be divided in two (partially overlapping) groups, depending on whether they focus on understanding the mechanisms behind delay propagation, or on predicting their appearance. The first group can further be divided among those works in which the propagation is modelled by means of large-scale simulations, as for instance in [4]–[8]; and those that rely on historical data to extract some properties of the propagation, as e.g. in [9]–[12]. With respect to the second group, the use of data mining and machine learning models, and specifically on tasks related to the prediction of delays, has a long history, possibly fostered by the operational relevance of such models. To illustrate, such models have extensively been used in the last two decades to predict delays [13]–[17]. Narrower models have also been proposed, focusing on identifying and estimating the impact of some operational elements on delays, e.g. of adverse weather events [18]–[21].

In more recent years a new trend is emerging: the use of Deep Learning (DL) models [22], [23], i.e. machine learning models not requiring an *a priori* definition of features, to predict the occurrence and magnitude of delays. The idea is thus to train a model using data that are not strongly pre-processed; on the contrary, the definition and selection of high-level features is performed in an automatic way by the model. The possibly first application of DL to delay prediction was proposed by Kim and co-authors, in which the sequences of departure and arrival flight delays of an airport were predicted using a Long Short-Term Memory network architecture, using input features like the delay of previous flights and the weather condition [24]. Numerous

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Luo.

new studies have followed this initial work, mainly focusing on increasing the spectrum of information fed in the models: from micro-scale meteorological conditions [25]–[28], reasons of previous delays [29], airline and flights connection structure [26], [28], [30], [31], airport crowdedness [26], [32], to aircraft trajectories [33] and airspace structure [32]. The interested reader can refer to [34] for a review on the use of data analysis in the study of air transport delays.

In this study we propose a bridge between both groups, by analysing the role of airports in the delay propagation process through the use of data mining, and specifically of Deep Learning, models. The starting hypothesis is that, given time series representing the observed dynamics of a set of elements (here, airports), the score of a classification task aimed at distinguishing them can be used as a metric of their dissimilarity - or, in other words, of their respective *identifiability*. In other words, if the classifier model is able to successfully discriminate between the two elements, it can then be concluded that there are differences between them, even though the exact nature of such differences is not readily available. Note that this is conceptually different from predicting the magnitude of future delays, as commonly done through DL models; and is instead closer to the problem of identity recognition [35]. The choice of DL models, as opposed to classical machine learning ones, is justified by the fact that *i*) they are extremely sensitive, i.e. they are able to detect even subtle and complex differences between data sets; and *ii*) their main drawback, i.e. their *black-box* nature, is not a limitation for our objective. While the use of DL models for identifying relationships and couplings between pairs of time series is not new (see [36]–[38]), to the best of our knowledge this is the first time such problem is tackled through the concept of identifiability.

We specifically analyse the delay profiles of the top-30 European airports from 2015 to 2018, where the delay profile is here defined as the normalised average hourly delay observed during one day of operation. By performing a classification task on pairs of airports, we are able to define their *identifiability*, i.e. a metric describing how uniquely are delays distributed across the day for each airport in the set. Two alternative scenarios can then emerge: airports can have similar (or identical) profiles, i.e. low identifiability; or, on the other hand, unique ones and hence a high identifiability. In the first case, this would imply that delays are a *global* property of the system, i.e. that their appearance is independent on the considered airport, or that their generation is driven by some systemic properties of air transport. On the other hand, unique delay profiles imply that delays are a *local* property, more the result of the dynamics and rules of each airport than of the global system.

The results here presented suggest an intermediate and complex situation. Generally speaking, airports are highly identifiable, i.e. their delay profiles are unique. Delays thus seem to be a local property; or, at least, characteristic dynamics at each airport dominate over global network effects. This identifiability is nevertheless not homogeneous and, on one hand, increases with connectivity, i.e. large and highly connected airports are more unique; and, on the other hand, reduces with geographical proximity, such that near airports tend to share similar profiles. We further analyse how this identifiability changed over time, and how it is affected by the winter and summer seasons.

The remainder of the work is organised as follows. Sec. II presents a simplified synthetic model of the identifiability of a set of networking elements, aimed at clarifying the interpretation of subsequent results. Sec. III then introduces the main materials and methods of this work, including the used flight data set, the Deep Learning models used in the classification, and the optimisation of the classification parameters. Results are presented in Sec. IV, including the study of the identifiability of airports, their geographical distribution, and their evolution through time. Conclusions on these results and ideas for future works are finally drawn in Sec. V.

II. A SYNTHETIC MODEL OF IDENTIFIABILITY

Let us suppose a system composed of three coupled elements a, b and c (in the case here considered, these would be three airports), each one described by an observable metric through time (here, average hourly delay) - see left part of Fig. 1 for a graphical representation. In each realisation of the system, 24 values are observed, thus yielding three time series $x_a(t)$, $x_b(t)$ and $x_c(t)$, with $t = 1, \ldots, 24$; note that each realisation is assumed to be independent from the other ones. These time series are further defined as the sum of three components:

$$x_a(t) = v_a(t) + \mathcal{N}(0, 0.1) + \gamma \left[x_b(t) + x_c(t)\right], \quad (1)$$

$$x_b(t) = v_b(t) + \mathcal{N}(0, 0.1) + \gamma \left[x_a(t) + x_c(t)\right], \quad (2)$$

$$x_c(t) = v_c(t) + \mathcal{N}(0, 0.1) + \gamma \left[x_a(t) + x_b(t)\right]. \quad (3)$$

The first part (i.e. v_a, v_b and v_c) represents a constant dynamic observed at each element independently of the specific realisation, or, in other words, what is characteristic of that element. To this, a random component is added, in the form of random numbers drawn from a normal distribution $\mathcal{N}$ of zero mean and 0.1 standard deviation - note that, without loss of generality, vectors v are expected to be defined between zero and one, hence this noise is 10% of them. This random component thus represents the variability that is observed across different realisations - or across different days. Increasing (or decreasing) this noise only changes the final classification score, as a large noise will mask the underlying dynamics; but the conclusions drawn from the model are independent from the selected noise level. Finally, each element receives contributions from the other two, through a coupling constant γ. While, for the sake of simplicity, this parameter is considered equal for all pairs of interactions, a more realistic scenario should include six parameters $\gamma_{a,b}$, $\gamma_{b,a}$, $\gamma_{a,c}$, $\gamma_{c,a}$, $\gamma_{b,c}$, and $\gamma_{c,b}$; results should then be represented in a six-dimensional space. In synthesis, the behaviour of each element is given by a characteristic

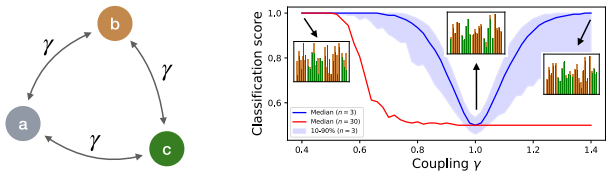

FIGURE 1. Synthetic model exemplifying the concept of identifiability. (Left) Graphical depiction of a system composed of three elements, pairwise coupled through a coupling constant γ; see main text for definitions. (Right) Median (solid blue line) and 10 – 90 percentile band (blue shaded region) of the obtained classification score as a function of γ, for a system composed of three elements. The solid red line represents the results for a larger system of 30 elements. Insets report three examples of the time series x_b and x_c, for $\gamma = 0.4$, 1.0 and 1.4. Results correspond to 10^3 independent realisations, using a ResNet deep learning classifier.

and (almost identical) repetitive dynamic, plus a tuneable contribution from the two neighbours.

Once a value of γ has been fixed, 100 random realisations have been obtained, thus obtaining 100 vectors $x(t)$ for each element. Note that this is equivalent to a data set comprising the average hourly delays of three airports along 100 days, as will be considered in the next section; still, results here reported only correspond to the previously described synthetic model, and not to real data. Vectors of elements b and c were then classified, using the Residual Network (ResNet) Deep Learning algorithm that will be described in detail in Sec. III-C. The final classification score, measured through the accuracy metric, has finally been calculated. Note that this classification score represents the identifiability of elements b and c, as these two elements can correctly be classified only if they are different in a characteristic way. As discussed in the introduction, the use of Deep Learning models is justified by the fact that they yield the best classification scores, i.e. they are able to detect the subtlest differences between the elements. Also note that a full identifiability analysis, as the one performed in the following sections, would require a similar classification task between all pairs of elements; the analysis is here limited to a single pair for the sake of clarity.

Fig. 1, right panel and blue line, reports the median of the classification score as a function of the coupling γ. For low values of the coupling constant γ, the behaviour of each element is dominated by its own characteristic dynamic v; as $v_b(t) \neq v_c(t)$, it is in general possible to distinguish x_b from x_c, and the classification yields a score close to 1. On the other hand, values of γ approaching 1 imply that the behaviour of each element is virtually indistinguishable from those of the other ones, as all dynamics are mixed, reaching what known in statistical physics as a mean field; hence no classification is possible. Finally, a perfect classification is recovered for $\gamma > 1.2$. In this latter case, the dynamics of each airport is dominated by the dynamics of the remaining ones; in other words, airport b is mainly defined by $v_a + v_c$, airport c by $v_a + v_b$, and hence $x_b \neq x_c$.

When the system is expanded to include 30 airports, similar results are still observed, see the red line in the same panel. The larger number of elements, and hence of connections, nevertheless acts as a global noise, which reduces the coupling required to loose the identifiability of the elements. It further becomes impossible to identify elements for $\gamma > 1$, as now the dynamics of each airport becomes the result of the sum of a large number of independent contributions, i.e. it effectively becomes unpredictable. Most importantly, it can be appreciated that a high classification score (i.e. a high identifiability) is still obtained for low values of γ; in other words, the size of the system does not affect the fact that elements are identified as long as they are not strongly coupled.

This simple model illustrates how the identifiability of a set of elements coupled together is the result of two contributions: how unique the dynamic of each element is, and how tightly they are coupled together. While such unambiguous conclusions cannot in principle be drawn when studying a real system, as more elements may affect the identifiability of elements, this synthetic model still suggests some interesting ideas. High classification scores will always imply high identifiability, as elements have dynamics that can be recognised; and this is usually the result of the unique dynamic of each element dominating over a global network signal. On the other hand, low classification scores (low identifiability) suggest that pairs of elements are coupled together in a tight way, such that they have a shared dynamics. It is additionally worth noting that in a real system the coupling $\gamma_{i,j}$ can be different for each pair of elements (i, j); in this case, elements can be identifiable even when having similar internal dynamics, provided their coupling patterns differ enough. For the sake of completeness, a final possibility can also emerge, not relevant for the present study: elements may not be identifiable when they are both disconnected and lacking an individual dynamic, i.e. $v_i = v_j$ and $\gamma_{i,j} = 0$ for all i and j.

III. MATERIALS AND METHODS

A. OPERATIONS AND DELAY DATA

Data about air transport operations have been extracted from the EUROCONTROL's R&D Data Archive, a public repository of historical flights made available for research purposes, and freely accessible at https://www.eurocontrol.int/dashboard/rnd-data-archive. It includes information about all commercial flights operating in and over Europe, completed with flight plans, radar data, and associated airspace structure. While it is limited to four months (i.e. March, June, September and December) of four years (2015-2018), it provides a good starting point to analyse both the structure of operations in Europe and the corresponding evolution.

In this study, we considered the information associated with the 30 largest airports in Europe, ranked according to their number of passengers. Table 1 reports the full list, and information about the number of landings and delayed flights. Additionally, Table 2 describes the evolution through time of some basic network metrics.

TABLE 1. Information on the 30 airports considered in this study, including their 4-letters ICAO code, number of landing flights, and number and percentage of flights delayed more than 10 minutes.

Name	ICAO	# flights	# delayed	% delayed	Name	ICAO	# flights	# delayed	% delayed
Heathrow Airport	EGLL	314,256	210,634	67.03%	Oslo Airport	ENGM	163,165	26,653	16.33%
Paris Charles de Gaulle Airport	LFPG	317,427	115,032	36.24%	Manchester Airport	EGCC	127,424	50,034	39.27%
Amsterdam Airport Schiphol	EHAM	322,436	147,448	45.73%	London Stansted Airport	EGSS	117,152	47,775	40.78%
Frankfurt Airport	EDDF	315,483	93,742	29.71%	Vienna International Airport	LOWW	160,290	61,393	38.30%
Adolfo Suárez Madrid-Barajas Airport	LEMD	255,238	85,266	33.41%	Stockholm Arlanda Airport	ESSA	159,099	38,698	24.32%
Josep Tarradellas Barcelona-El Prat Airport	LEBL	207,449	70,089	33.79%	Brussels Airport	EBBR	148,967	55,020	36.93%
Munich Airport	EDDM	261,189	67,268	25.75%	Milan Malpensa Airport	LIMC	116,939	37,491	32.06%
Gatwick Airport	EGKK	187,035	92,456	49.43%	Düsseldorf Airport	EDDL	141,298	39,122	27.69%
Rome-Fiumicino International Airport	LIRF	204,317	47,262	23.13%	Athens International Airport Eleftherios Venizelos	LGAV	122,253	23,297	19.06%
Paris Orly Airport	LFPO	156,009	48,817	31.29%	Berlin Tegel Airport	EDDT	119,595	29,654	24.79%
Dublin Airport	EIDW	142,772	35,877	25.13%	Málaga Airport	LEMG	79,231	25,333	31.97%
Zurich Airport	LSZH	167,403	59,574	35.59%	Warsaw Chopin Airport	EPWA	105,709	12,072	11.42%
Copenhagen Airport	EKCH	173,942	28,262	16.25%	Geneva Airport	LSGG	114,486	27,147	23.71%
Palma de Mallorca Airport	LEPA	128,784	32,538	25.27%	Hamburg Airport	EDDH	97,619	20,018	20.51%
Humberto Delgado Airport	LPPT	127,311	46,709	36.69%	Václav Havel Airport Prague	LKPR	91,072	20,877	22.92%

TABLE 2. Evolution of the data available in this work. Starting from the third, each column reports: the total number of flights included in the data set; the number of flights landing in the 30 considered airports; the busiest airport, in terms of number of landings (the number in parenthesis); and the second busiest airport.

Year	Month	# flights	# flights in airports	Busiest airport	Second busiest airport
2015	March	697,365	287,289	EGLL (19,713)	LFPG (18,875)
2015	June	850,196	334,147	EDDF (20,865)	LFPG (20,705)
2015	September	859,814	337,847	EDDF (20,760)	LFPG (20,710)
2015	December	660,745	270,925	EGLL (18,550)	LFPG (18,387)
2016	March	716,256	292,827	EGLL (19,502)	LFPG (18,824)
2016	June	862,947	342,380	EHAM (21,418)	LFPG (20,439)
2016	September	890,970	350,575	EHAM (21,775)	LFPG (20,901)
2016	December	695,250	283,487	LFPG (18,541)	EGLL (18,390)
2017	March	746,984	306,292	EHAM (19,892)	EGLL (19,623)
2017	June	910,906	353,950	EHAM (22,251)	LFPG (20,698)
2017	September	935,520	359,369	EHAM (22,338)	EDDF (21,426)
2017	December	713,504	281,931	EGLL (18,771)	LFPG (18,598)
2018	March	775,674	311,154	EDDF (20,375)	EHAM (20,240)
2018	June	951,437	361,765	EDDF (22,309)	EHAM (21,996)
2018	September	962,115	368,506	EDDF (22,701)	EHAM (22,346)
2018	December	750,835	302,906	LFPG (19,449)	EHAM (19,264)

For each flight landing at these 30 airports, the corresponding delay has been calculated as the difference between the actual (from the ATFM-updated flight plan) and the planned (according to the last filed flight plan) landing times. Afterwards, for each airport, flights have been grouped according to the actual landing hour, and the average delay per hour has been calculated. Subsequently, these time series have been split in windows of 24 hours, i.e. of the average hourly delay per airport per day.

In this work we are interested in the profile of delays, i.e. their distribution throughout the day, as opposed to the absolute value. In other words, we are interested in seeing if an airport suffered from delays homogeneously throughout the day or at some specific hours, and not in the value of the average delay at a given time. For that, average delays have been transformed through a Z-Score. Mathematically, given a time series $(x_1, x_2, \ldots, x_{24})$, it is transformed according to:

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}, \tag{4}$$

with $\bar{x}$ being the average and σ_x the standard deviation. Therefore, a value of z_i above (below) zero indicates that the

average delay at time i was higher (respectively, lower) than what is observed throughout the same day.

B. WEATHER DATA

Each airport of Tab. 1 has further been characterised by the average climatic conditions that are expected in winter and summer; this information will be used in Sec. IV-B to explain changes in identifiability between these two operational seasons. Towards this aim, we have downloaded several average climate indicators for each airport from the corresponding city's Wikipedia page. These include: the average minimum winter temperature, as the average of the average minimum temperature observed in March and December (i.e. the two winter months available in the data set); the average minimum summer temperature, corresponding to June and September; the drop in temperature between summer and winter, i.e. the difference between the two previous values; and the average number of rainy days in winter (again, average between March and December).

C. CLASSIFICATION MODELS

Deep learning can be defined as a set of machine learning algorithms that progressively extract higher-level features from the raw input, usually with the objective of performing a supervised classification [22], [23]. Compared to standard machine learning classification models, deep learning presents the advantage of not assuming nor requiring *a priori* structures in the data, and of not requiring a pre-processing of features; in other words, features are automatically extracted from data, without human intervention. This results in a drastically higher efficiency, especially in complex problems for which features are difficult to be defined. On the other hand, this also implies high computational costs, and usually the need of dedicated hardware - as, e.g., general purpose graphics processing unit (GPGPU).

Here we consider a subset of deep learning algorithms designed to classify time series; in other words, given a set of time series, each one associated with a label, the objective is to assign the correct label to a new time series presented to the algorithm. While this is not one of the main focuses of deep learning, several models have been developed, usually evolutions of models designed for image classification - see [39] for a full review. More specifically, the following five models have been used:

- *Multi Layer Perceptron (MLP)*. One of the most traditional and simplest form of neural networks, it is composed of a set of nodes organised in layers, each one receiving information from the previous layer and responding through a nonlinear activation function. Even though it does not encode temporal information, the MLP model has been proposed as a baseline architecture for classifying time series [40]. The network here considered is composed of 4 layers, each one fully connected to the outputs of the previous one, and with the final layer being a *softmax* classifier. The activation function is the well-known rectifier linear unit (ReLU) [41].
- *Convolutional Neural Network (CNN)*. Convolutional networks are specialised versions of MLP, in which the matrix multiplication is substituted by a convolution operation [42]. Their advantages include a space (or, in the case of time series, time) invariance [43], and a reduced tendency to overfitting. We here consider a simple convolutional model, composed of two convolutional layers followed by a final *sigmoid* classifier.
- *Residual network (ResNet)*. Residual networks are inspired by the way pyramidal cells are organised in the cerebral cortex; specifically, the connections between layers are not sequential, but instead some layers can be skipped (creating shortcuts or jumps). This presents the advantage of a simpler structure, and consequently of a reduced training cost [44]. The networks here considered are composed of 11 layers, the first 9 of them being convolutional, followed by a global average pooling layer that averages the time series across the time dimension, and by a final softmax classifier, as proposed in [40].
- *Fully Convolutional neural Network (FCN)*. FCNs are networks in which only convolution operations can be performed; in other words, they are equivalent to CNNs without fully connected layers [45]. The model is composed of three convolutional blocks, each one performing a convolution, a batch normalisation and a final activation. As a last step, the result of the third convolutional block is fed to a *softmax* classifier [40].
- *Multi Channel Deep Convolutional Neural Network (MCDCNN)*. This model is based on a modified CNN, in which the convolutions are applied independently (in parallel) on each dimension (or channel) of the input multivariate time series [46], [47].

The five models have been implemented in Python 3.8.5 using the libraries TensorFlow [48] and Keras [49]. In each iteration of the classification problem, a random half of the available time series has been used for training, and the remaining half for evaluating the model, being thus equivalent to a two-fold cross-validation. Each model performance is finally measured through the corresponding accuracy score; note that other complementary metrics, as e.g. recall or F-score, are here redundant due to the use of a perfectly balanced data set.

D. INTERPRETING THE CLASSIFICATION SCORE

The output of any classification task described in this work is measured in terms of the accuracy, i.e. the number of correctly labeled instances divided by the total number of instances. As only problems involving two classes and with the same number of instances in each class are here considered, the accuracy is expected to be included between 0.5 and 1.0. The former case indicates that only half of the instances have been correctly labeled, which is also what expected if labels were

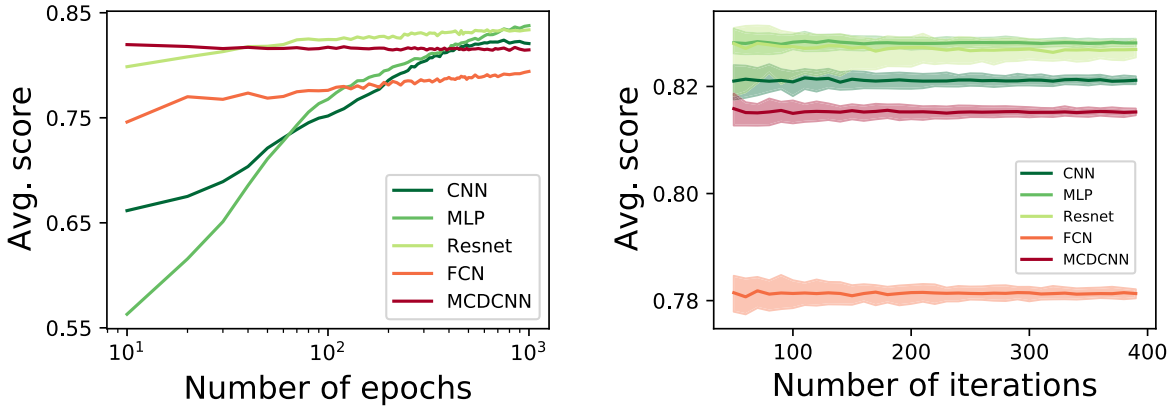

FIGURE 2. Tuning the parameters of the classification models. (Left) Evolution of the average score as a function of the number of epochs in the training, for the classification of the pair of airports Dublin Airport, DW, and Brussels Airport, EBBR. (Right) Evolution of the average classification score (solid lines) and of the 10 – 90 confidence band as a function of the number of random iterations. The number of epochs in the second case was set to 600 for CNN and MLP, and 200 for Resnet, FCN and MCDCNN.

assigned at random; in other words, the model is not able to detect any useful pattern in the data. On the other hand, an accuracy of 1.0 points towards a perfect classification; from the point of view of the model, the two groups of instances are clearly different.

It is straightforward to interpret this classification score as a metric of identifiability. Specifically, if the delay profiles of two airports can perfectly be classified, i.e. with an accuracy of 1.0, they then are substantially different, and it is possible to construct a model able to identify the airport corresponding to each delay profile without errors. It is worth noting that such identifiability metric is a lower bound of the real identifiability. On one hand, the fact that a given classification model is not able to identify differences between two groups does not imply that the two groups are equal - they may actually be identifiable by more advanced and complex models. On the other hand, a high classification accuracy implies that there are differences, at least as large as those identified by the considered model. The use of Deep Learning classification models, i.e. of the most accurate models presently available, guarantees that this lower bound is as close as possible to the real identifiability value.

The five models here considered are different in terms of their internal structure, and of what features in the time series they are able to detect; as such, they may perform differently when classifying different pairs of airports. In other words, it may happen that one model works well classifying two airports, but its efficiency may lower for another pair, as the features that were important in the first case are not so in the second. As the objective here is to assess the identifiability of airport delays, and not the presence of a specific feature, each classification is executed with all five models, for then retaining only the highest score.

E. PARAMETERS TUNING

There are two parameters that have to be set for each model, and that may strongly affect the outcome of the classification. The first one is the number of epochs, i.e. the number of times the training is performed over all available data. This number controls a trade-off: the larger the number of epochs, the more accurate is the classification, albeit at an increased computational cost. Additionally, performing additional trainings beyond a certain level usually does not improve the results.

Secondly, one has to note that the training is a stochastic process: the original data set is split between training and validation at random; and the initial state of the neural network is also random. As a consequence, it is customary to repeat the whole training and validation process several times, and to finally average the result. The second parameter is thus the number of iterations, i.e. executions of the full training and evaluation cycle with random initial conditions. The larger this value, the more stable is the final result, yet again at the price of a larger computational cost.

Fig. 2 reports the results of tuning these two parameters. Specifically, we have considered one pair of airports (Dublin Airport, DW, and Brussels Airport, EBBR), and performed the corresponding classification varying these parameters. The right panel of Fig. 2 indicates that 400 iterations is a value high enough to obtain very stable results with all models. On the other hand, the left panel presents a more complex situation. Some models, as e.g. MCDCNN, only require a few epochs; others, specifically CNN and MLP, only saturate for thousands of epochs. As a compromise between precision and computational cost, the following number of epochs have been selected: 600 for CNN and MLP, and 200 for Resnet, FCN and MCDCNN.

F. TOPOLOGICAL METRICS

In order to evaluate how the identifiability of each airport is modulated by its connectivity inside the network, the following six metrics have been calculated for each one of them. These are based on a network analysis [50], [51], in which each airport is represented by a node, and nodes are

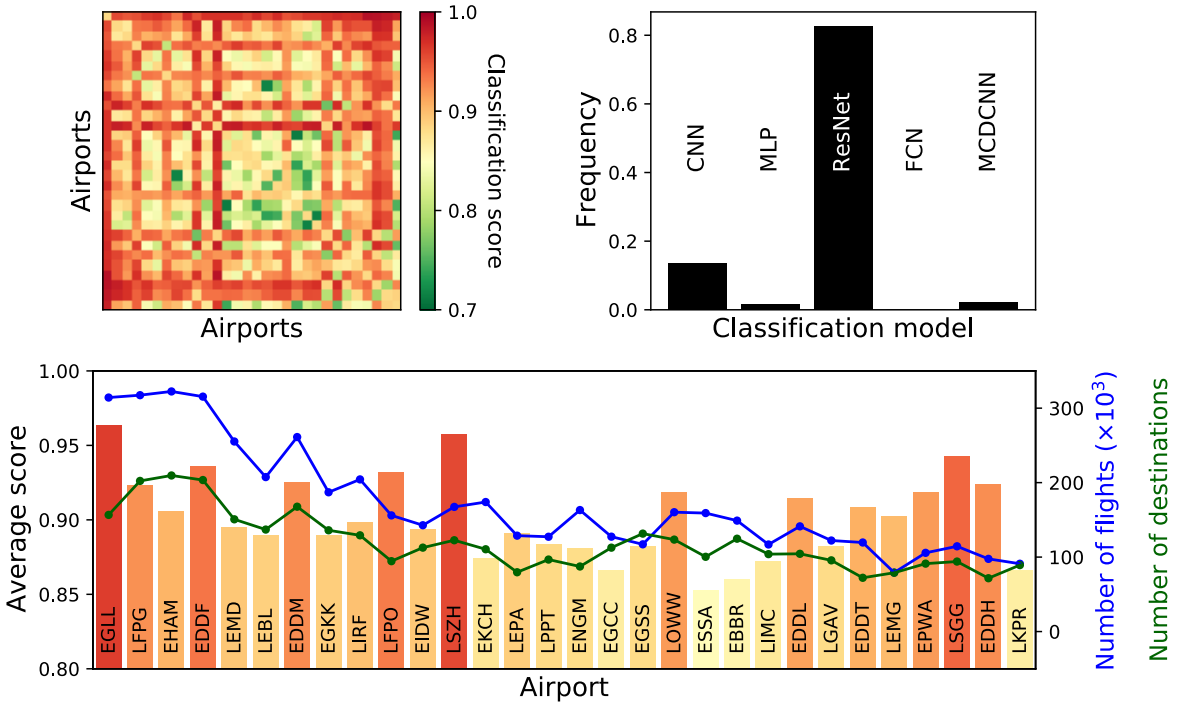

FIGURE 3. Identifiability of airports' delay profiles. (Top left) Classification score of each pair of airports; airports are sorted according to Table 1, and also as in the bottom panel. (Top right) Fraction of times each classification model yielded the best classification score. (Bottom) Average classification score of each airport, left Y axis; and corresponding number of flights and connected airports, right Y axis.

pairwise connected by links whenever a direct flight exists between the corresponding airports [52].

Number of flights. Total number of flights landed at the considered airport, as reported in the available data set, irrespectively of their origin.

Number of destinations. Number of unique airports which the considered airport is connected to, according to all available flights.

Clustering coefficient. The local clustering coefficient of a node quantifies the tendency of the neighbours of a node to form a clique, i.e. a complete graph [53]. More specifically, it is defined as the proportion of the number of links between the nodes within the neighbourhood of the considered airport, divided by the number of links that could possibly exist between them.

Weight centrality. Measure of the centrality (i.e. importance) of each airport, and defined as the sum of all flights in it landing, divided by the total number of flights.

Eigenvector centrality. Measure of node centrality, according to which the centrality of a node is proportional to the sum of the centralities of nodes to it connected.

Closeness centrality. Node centrality measure calculated as the reciprocal of the sum of the length of the shortest paths between the considered node and all other nodes in the network.

Note that, in order to simplify the comparison between different metrics, all three centralities have been normalised between zero and one, with one being the centrality of the most central airport.

IV. RESULTS

A. IDENTIFIABILITY OF AIRPORTS

We start by analysing the identifiability of airports in Fig. 3, understood as the score of the classification of hourly delay profile. Firstly, the top left panel reports the score obtained for all couples of airports. It can be appreciated that most pairs of airports yield a very good classification, in all cases above 70%. This is well above what obtained for the same classification problem when the data of each airport are randomly shuffled in order to destroy any characteristic pattern (average score of $61.3 \pm 9.3\%$), thus confirming the statistically significance of these results. Additionally, the top right panel reports the fraction of times each model yields the highest classification score; ResNet is the clear winner, followed by CNN. The bottom panel of Fig. 3, left Y axis, depicts the average classification score for each airport, i.e. the classification score averaged over all pairs including that airport. All airports are highly identifiable, and especially London Heathrow (EGLL), Zurich (LSZH) and Geneva (LSGG). In order to exclude that these results could be biased by the use of a pairwise classification task, scores are also reported for a task in which the profiles of one airport are compared to a random selection of profiles

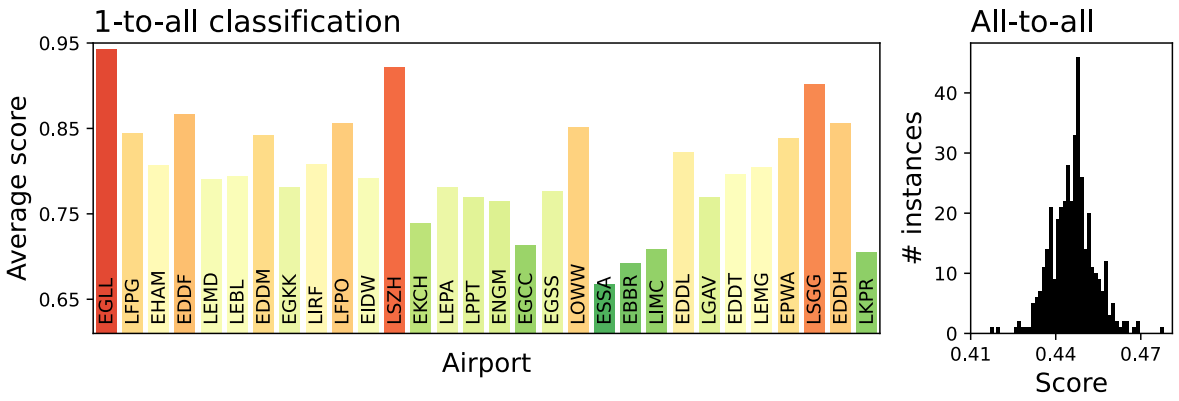

FIGURE 4. Global identifiability of airports' delay profiles. (Left) Classification score obtained when profiles of one airport are compared against a random selection of profiles of all other airports. (Right) Histogram of the classification score obtained by considering all the airports at the same time, i.e. a 30-classes classification problem.

of all other airports (Fig. 4, left panel); and a task in which all airports are considered at the same time, thus corresponding to a 30-classes classification (Fig. 4, right panel). While scores are generally lower, as is to be expected due to the higher complexity of these tasks, airports with the highest scores coincide with those of Fig. 3. These 1-to-all classification tasks can also be used to extract class-discriminative vectors, i.e. vectors representing what parts of the delay profiles are more characteristic of each airport, see Appendix A. Additionally, the 30-classes classification yields a score (0.446 ± 0.0076) significantly higher than what would be expected if results were random ($1/30 \approx 0.033$).

In order to understand whether such identifiability is associated to some airport properties, the right Y axis of Fig. 3 reports the corresponding number of flights and number of connected airports. This is further explored in Fig. 5, depicting scatter plots of the classification score of each airport as a function of the six metrics defined in Sec. III-F, the corresponding linear fits, and the resulting ρ and p-values. While the effect is not very strong (and seldom statistically significant for $\alpha = 0.05$), large and highly connected airports are more identifiable; in other words, small airports seem to have more common patterns of delays, while those of large airports are more unique. Still, a linear model using the six metrics to predict the classification score only reaches an $R^2 = 0.377$; this indicates that the metrics are highly correlated, and that the identifiability of each airport only marginally depends on them.

We afterwards study the structure created by such pair-wise identifiability, with the objective to understand if there are clusters (or communities) of highly similar airports. For that, the pair-wise identifiability (as depicted in Fig. 3 top left panel) has been transformed into a similarity using the standard metric $s_{i,j} = \sqrt{(1 - I_{i,j})/2}$, $I_{i,j}$ denoting the identifiability of airports i and j. Afterwards, those similarities have been interpreted as weights of the network links, in which nodes represent airports and links the pair-wise similarity. Finally, the celebrated Louvain algorithm for community detection [54] has been applied. The results of this community analysis are presented in Table 3; additionally, the three largest communities are also plotted on the European map in Fig. 6. Most notably, communities are not randomly distributed in space, but instead seem to be related to geographical regions - e.g., the largest one (red) to the central Europe, the second one (yellow) to UK and Belgium, etc. Prima facie, this may seem to be associated to time zones, such that two airports may have the same delay profile because most delays occur at the same (local) time. Nevertheless, this is disproved by several cases, e.g. Brussels, which does not share time zone with UK, or Prague, which shares time zone with Germany.

Another possibility is the existence of medium-scale weather patterns, that create disruptions across large regions and hence force the delay profiles of several airports to some common dynamics. This would explain why, for instance, a similar pattern is observed in UK and northern Europe. The fact that not all airports in those regions are included can be explained by considering their different role in air transport; to illustrate, and following the previous example, London Heathrow, Paris Charles de Gaulle and Amsterdam Schiphol may not be included in the yellow community due to their significantly higher traffic, which results in specific delay patters - as also illustrated in Fig. 5.

B. INVARIANCE OF IDENTIFIABILITY THROUGH TIME

One natural question is whether airport delay profiles are stable through time, or, on the contrary, have changed throughout the years. In order to assess this point, we here perform a classification in which the delay profiles of an airport are organised in two groups, one for years 2015 and 2016, and a second one for years 2017 and 2018. A high classification score would thus indicate that delays have changed through the four years. The result, depicted in Fig. 7, indicates that this is not the case, and that most airports have maintained a similar delay profile.

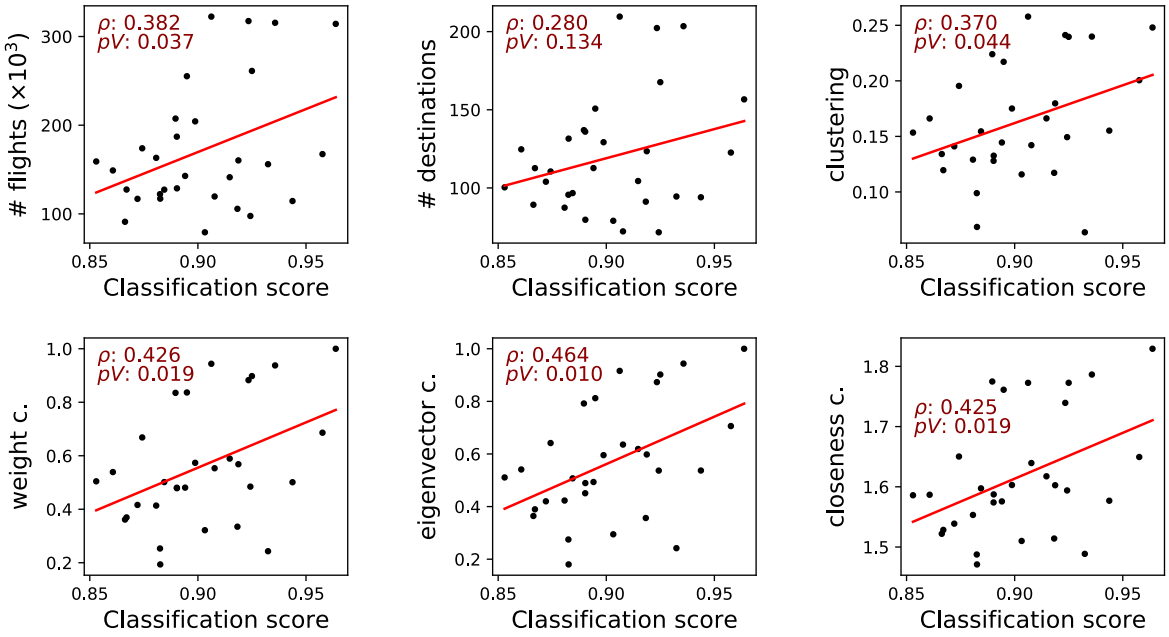

FIGURE 5. Scatter plots of the identifiability of each airport, represented by its average classification score, *vs.* the six metrics defined in Sec. III-F. The red lines represent the best linear fit; and each panel further includes the ρ and the *p*-value of each fit.

TABLE 3. List of communities, obtained from the structure of similarities between them, and ordered by size. The color in parenthesis for the three largest communities corresponds to their color in Fig. 6.

Community #	Airports	Community #	Airports
1 (red)	LFPO, LSZH, EDDL, EDDT, LSGG, EDDH	8	EGLL
2 (yellow)	EGKK, EGCC, EGSS, EBBR	9	LEMD
3 (green)	EKCH, ENGM, ESSA, LKPR	10	LIRF
4	LFPG, EHAM, LIMC	11	EIDW
5	LEPA, LGAV, EPWA	12	LPPT
6	EDDF, EDDM	13	LOWW
7	LEBL, LEMG		

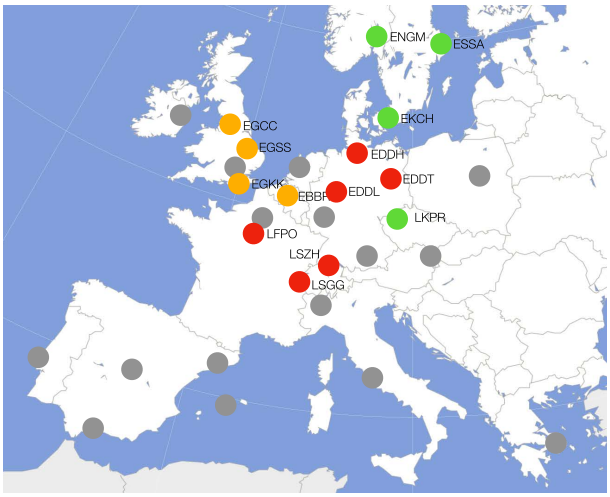

FIGURE 6. Map of the spatial location of the three largest communities, as listed in Table 3. Airports included in the analysis, but not belonging to these three communities, are marked in grey.

As in the previous section, we analysed whether this time-dependent identifiability can be explained by metrics representing the dynamics of airports. The difference is that, here, metrics must represent variations between the two groups of years; for that, given a metric m, its variation is calculated as $\Delta m = \log_2 m_{2015-2016}/m_{2017-2018}$. Therefore, positive values of Δm indicate an increase in m, and the opposite for negative values. An analysis of the correlation between this identifiability and the variation of the six previously considered airport metrics does not yield significant results: number of flights, $\rho = 0.026$ and p-value $= 0.892$; number of destinations, $\rho = -0.195$ and p-value $= 0.303$; clustering, $\rho = -0.071$ and p-value $= 0.707$; weight centrality, $\rho = -0.014$ and p-value $= 0.940$; eigenvector centrality, $\rho = 0.021$ and p-value $= 0.912$; and closeness centrality, $\rho = 0.003$ and p-value $= 0.998$.

We further analysed whether the delay profiles of each airport changes between summer and winter, through a classification task involving days in March and December on one hand, and June and September on the other. Results, reported in Fig. 8, indicate that there is a generally strong difference in the delays of different seasons; yet, such difference is not driven by changes in the number of

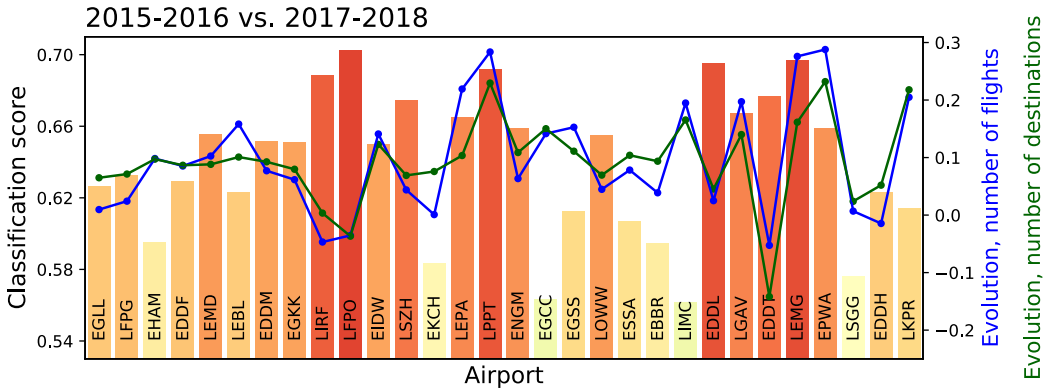

FIGURE 7. Identifiability of airports through time. Each bar (left Y axis) represents the classification score between the delay profiles of each airport, with the two groups defined as days in years 2015 – 2016 and 2017 – 2018. The left Y axes, and the blue and green lines, depict the evolution in number of flights and connected airports, respectively.

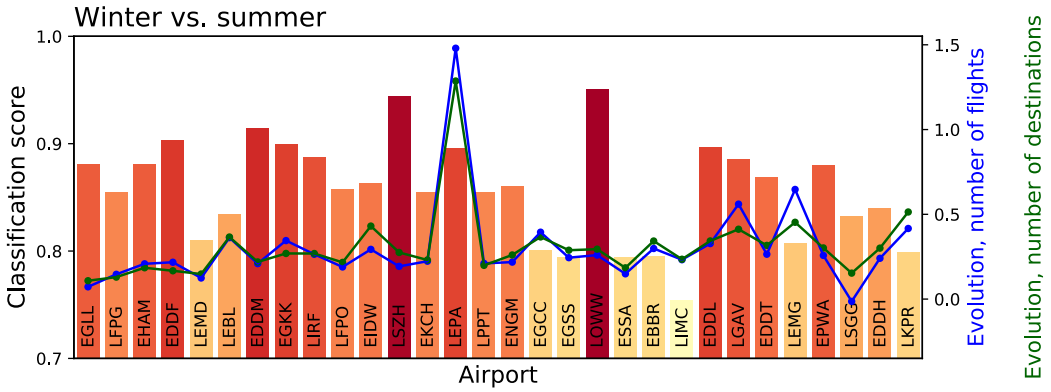

FIGURE 8. Identifiability of airports across seasons. Each bar (left Y axis) represents the classification score between the delay profiles of each airport, with the two groups defined as days in March/December and June/September. The left Y axes, and the blue and green lines, depict the evolution in number of flights and connected airports, respectively.

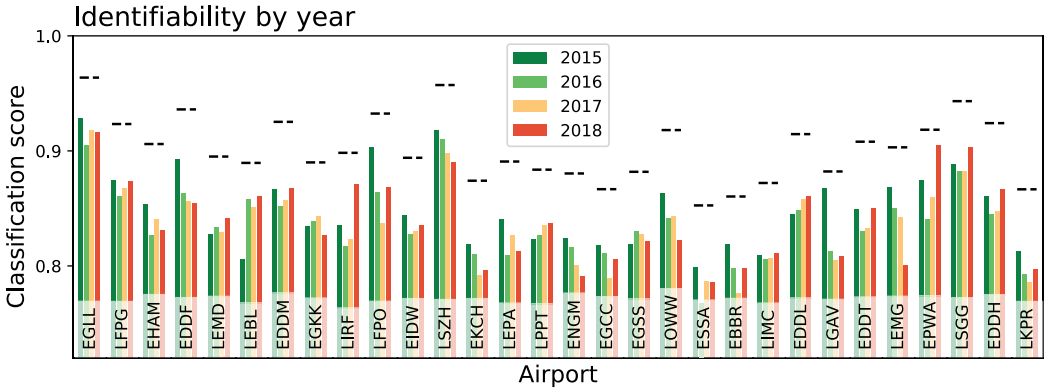

FIGURE 9. Evolution of the identifiability of each airport throughout the four years. Each bar indicates the average classification score obtained by using only the data of one year. The horizontal dashed lines depict the global average classification score for each airport, as reported in Fig. 3 bottom panel.

flights or of destinations (blue and green lines). We further checked whether such difference may be due to weather patterns, using the weather data described in Sec. III-B, and by performing a linear regression between the score of the summer/winter classification and the weather values. No statistically significant relation was found with the drop in minimum temperature between summer and winter ($\rho = 0.0452$, p-value $= 0.8157$), average minimum winter temperature ($\rho = -0.0591$, p-value $= 0.7605$), and average number of rainy days in winter ($\rho = 0.2700$, p-value $= 0.1917$).

As a final issue, we have studied how the identifiability of each airport, as depicted in Fig. 3 bottom panel, has evolved through time. For this, for each pairs of airports, four

different classification tasks have been performed, each one using data of one single year. The result, as shown in Fig. 9, is thus four scores for each airport, each one quantifying the identifiability of the airport's delays in one given year. Note that this is complementary to what presented in Fig. 7, as there the focus was the identifiability of each airport from itself in two time intervals, and here is of each airport from the other ones. It can be appreciated that variations across years are minimal, with no clear trend. Also, the fact that the identifiability by year is always lower than the global one (the latter represented by the horizontal dashed lines) is easily explained by the fact that each classification task can rely on one forth of the data.

V. DISCUSSION AND CONCLUSION

In this work we have leveraged on the possibilities offered by Deep Learning to create an identifiability index for the top-30 European airports, i.e. a metric describing how uniquely distributed are average delays throughout the day. The main obtained result is that airports are actually highly identifiable and unique, with pair-wise classification scores almost reaching a 100% precision - see Fig. 3. While it is tempting to conclude that delays are a local phenomenon, i.e. only driven by local dynamics and operational constrains, several caveats have to be highlighted.

First of all, high identifiability does not preclude a propagation process, i.e. the diffusion of delays throughout the system through secondary (or reactionary) delays. Both aspects are compatible if received delays are different enough across different airport, as seen in Fig. 1. They are also compatible if one allows for received delays to be processed (or dealt with) by each airport through different (i.e. local) mechanisms. To illustrate, two airports may receive the same amount of delays from other airports; yet, the resulting dynamics can be different due to, e.g., different equipments and operational buffers. Therefore, what here presented does not contradict the large body of literature studying the propagation of delays [7], [11], [55]–[59]; instead, it suggests that propagation itself cannot be considered as a homogeneous phenomenon.

Secondly, this global identifiability is not homogeneous, but is instead modulated by different factors. On one hand, it is positively (albeit weakly) correlated with the traffic volume and connectivity of each airport - see Fig. 5. Thus, being strongly integrated in the network, i.e. receiving delays from multiple and heterogeneous sources, makes an airport more unique. On the other hand, airports located in some geographical regions seem to be more similar (see Fig. 6), possibly because their operations are modulated by common weather patterns.

Thirdly, when all flights throughout one year are considered, airport identifiability did not substantially change between the first and last year analysed. On the other hand, delay profiles were different (hence identifiable) between the winter and summer seasons of the same airport. This suggests that the change in network connectivity between winter and summer produces different, albeit characteristic, delay patterns in each season.

Irrespectively on how this identifiability originates, the fact that airports are characterised by fairly unique delay profiles leads to important conclusions. Delays (at least at the aggregated level here considered) are mostly defined by their past, as opposed to other external factors. In other words, it has previously been shown that predicting the delay of an individual flight requires taking into account aspects like local weather patters, and airport and airspace crowdedness [34]. Results here presented instead show that the average delay per hour can be estimated by only knowing the delay evolution of the same airport in previous days, as this evolution will be unique to that airport, and hence different from those of other airports. Weather and other elements, while useful to reduce the prediction error, are therefore not the main features. Most notably, this also applies to network effects: while the delay of individual flights can only correctly be predicted by taking into account the status (e.g. delays and crowdedness) of other airports and of the system as a whole [26], [32], [33], [60], the aggregated dynamics at an airport can be described without it. To make a parallelism with statistical physics, macroscopic observables of a system (e.g. the pressure or the temperature of a gas) can be studied without explicit knowledge of the individual elements composing it (e.g. position and velocity of all gas' particles).

From an operational perspective, this implies that, firstly, delays can be predicted, and hence acted upon. This stems from the fact that, if delays were unpredictable and were randomly changing across different days, they would not form consistent and identifiable profiles. This is of course in agreement with the large body of Literature dealing with delay prediction [13]–[17], [34]. Secondly, and more importantly, that these predictions and interventions aimed at reducing delays must be local in nature, as the delay profile at one airport (and hence, the evolution of delays through time) encodes most of the information about such delay dynamics. In other words, average delays at a given airport are largely predictable without looking at the network status. These results also suggest that airport identifiability could be used as a metric to measure how efficiently delays are managed throughout the system, and how intensely is the local dynamics of airports dictated by the global propagation.

As a final point, several limitations of the present study have to be highlighted. The calculation of delays has been performed with the data available in the EUROCONTROL's R&D Data Archive. While it provides a complete and official source of information about all flights operating in Europe, the planned time of landing corresponds to the last filed flight plan, and not to the original intentions of the airlines. This may result in a bias (specifically, a decrease) in the estimated delay at landing. Also, delays have been calculated per flight, i.e. disregarding the number of passengers that were affected and how these delays may disrupt multi-leg trips [61]–[64]; in other words, only the magnitude of delays, and not their

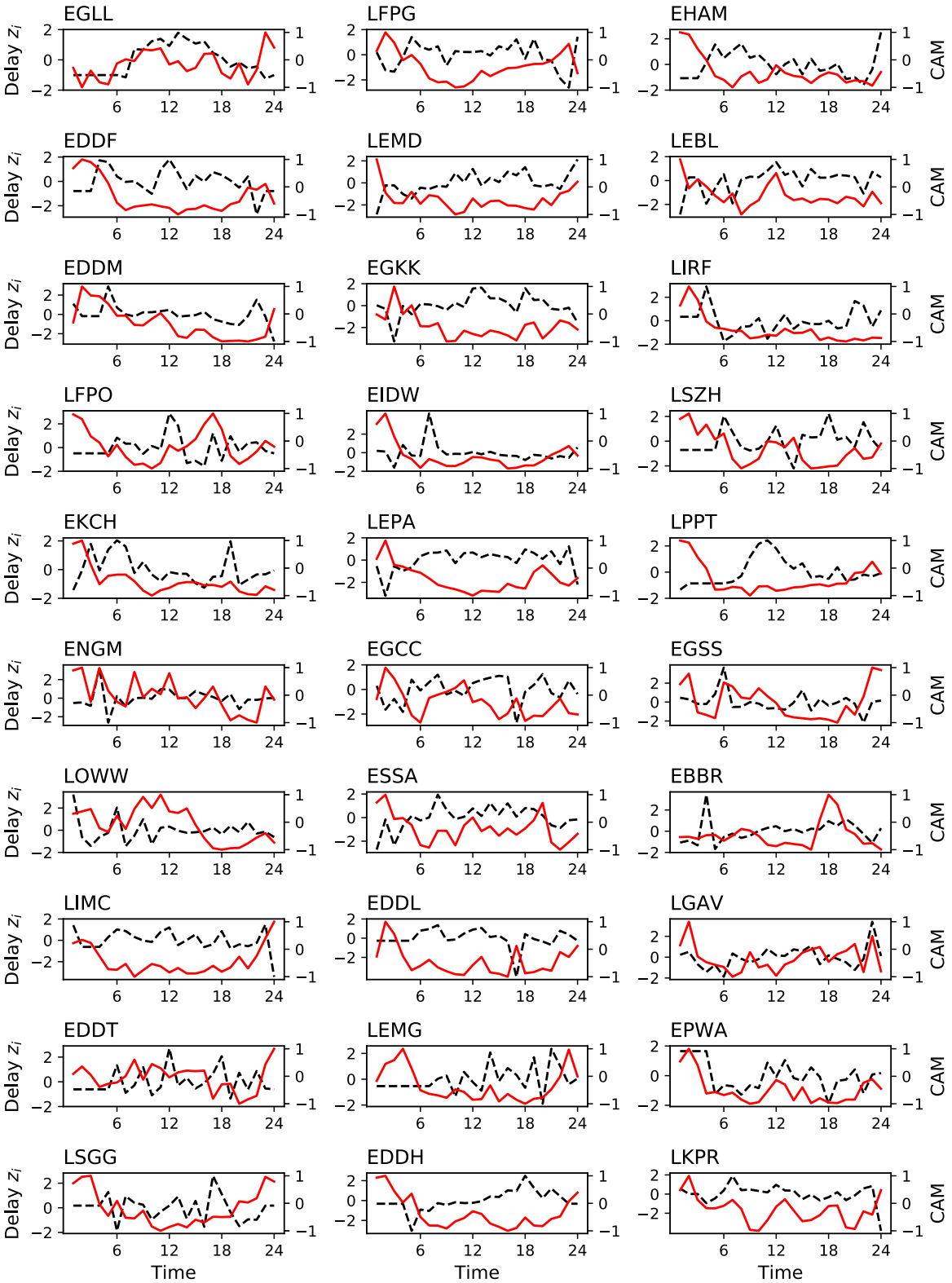

FIGURE 10. Graphical representation of the class-discriminative vector (solid red lines, right Y axes) for each airport, as obtained by the Gradient-weighted Class Activation Mapping (Grad-CAM) approach [74]. Dashed black lines (left Y axes) represent examples of the delay profiles of each airport.

importance, has here been considered. Regarding available data, limited information about the status of the airports and of the system in general has been included in the analysis; on the other hand, using data like flight connectivities, airspace structure and crowdedness may help explaining the origin of the identifiability, as previously shown for delay prediction [26], [34]. While delay propagation is a global process involving multiple airports simultaneously, such process as here been represented as a set of pairwise interactions. This is customary in network science, and specifically in the reconstruction of functional networks [65], [66]. Still, higher dimensional options have recently been proposed [67], [68], which could in principle yield more complete representations of the propagation dynamics. It has to finally be noted that, while the strength of the interactions was represented by the parameter γ in the model, it is not possible to derive a real interaction strength from the available data. While it may be tempting to state that, given a pair of airports, their associated γ should be inversely proportional to the corresponding pairwise identifiability, this would not take into account that the airport dynamics is in reality not constant, and it may be changed by random events or different traffic patterns. While outside the scope of this work, a more precise identification of all interaction γs may be possible through the use of recent results on the dynamics of coupled systems, e.g. [65], [69]–[71].

APPENDIX A
VISUALISATION OF AIRPORT CHARACTERISTIC PROFILES

In spite of the unavoidable black-box nature of Deep Learning models, several attempts have been made in recent years to extract intuitive and understandable components from them, in order to allow the practitioner to understand why a given model predicted what it predicted - see [72], [73] for reviews on the topic. We here applied the Gradient-weighted Class Activation Mapping (Grad-CAM) approach [74] to the 1-to-all classification problems presented in Fig. 4. In synthesis, a ResNet model is trained to recognise the delay profiles of one airport against a pool of profiles of all other airports; the Grad-CAM approach is then used to retrieve a class-discriminative vector, i.e. a vector representing how important is the value observed at a given time of the day for recognising such airport. The vectors for the 30 airports here considered are depicted in Fig. 10 (red solid lines, right Y axes), alongside an example of the delay profile (dashed black lines, left Y axes).

REFERENCES

[1] A. J. Cook and G. Tanner, "European airline delay cost reference values," Univ. Westminster, London, U.K., Tech. Rep. 09-112277-C, 2011.
[2] D. Duytschaever, "The development and implementation of the EUROCONTROL central air traffic flow management unit (CFMU)," *J. Navigat.*, vol. 46, no. 3, pp. 343–352, Sep. 1993.
[3] S. Carlier, I. De Lépinay, J.-C. Hustache, and F. Jelinek, "Environmental impact of air traffic flow management delays," in *Proc. 7th USA/Eur. Air Traffic Manage. Res. Develop. Seminar (ATM)*, vol. 2, 2007, p. 16.
[4] K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany, "A model for projecting flight delays during irregular operation conditions," *J. Air Transp. Manage.*, vol. 10, no. 6, pp. 385–394, Nov. 2004.
[5] M. Janić, "Modeling the large scale disruptions of an airline network," *J. Transp. Eng.*, vol. 131, no. 4, pp. 249–260, Apr. 2005.
[6] M. Jetzki, "The propagation of air transport delays in Europe," Ph.D. dissertation, Dept. Airport Air Transp. Res., RWTH Aachen Univ., Aachen, Germany, 2009.
[7] P. Fleurquin, J. J. Ramasco, and V. M. Eguiluz, "Systemic delay propagation in the US airport network," *Sci. Rep.*, vol. 3, no. 1, pp. 1–6, Dec. 2013.
[8] K. Gopalakrishnan and H. Balakrishnan, "A comparative analysis of models for predicting delays in air traffic networks," in *Proc. 12th USA/Eur. Air Traffic Manage. Res. Develop. Seminar (ATM)*, 2017, pp. 1–11.
[9] M. Zanin, "Can we neglect the multi-layer structure of functional networks?" *Phys. A, Statist. Mech. Appl.*, vol. 430, pp. 184–192, Jul. 2015.
[10] S. Belkoura and M. Zanin, "Phase changes in delay propagation networks," in *Proc. 7th Int. Conf. Res. Air Transp. (ICRAT)*, 2016, pp. 1–8.
[11] M. Zanin, S. Belkoura, and Y. Zhu, "Network analysis of Chinese air transport delay propagation," *Chin. J. Aeronaut.*, vol. 30, no. 2, pp. 491–499, 2017.
[12] P. Mazzarisi, S. Zaoli, F. Lillo, L. Delgado, and G. Gurtner, "New centrality and causality metrics assessing air traffic network interactions," *J. Air Transp. Manage.*, vol. 85, Jun. 2020, Art. no. 101801.
[13] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen, "Estimation of delay propagation in the national aviation system using Bayesian networks," in *Proc. 6th USA/Eur. Air Traffic Manage. Res. Develop. Seminar*, 2005, pp. 1–11.
[14] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transp. Res. C, Emerg. Technol.*, vol. 44, pp. 231–241, Jul. 2014.
[15] K. R. Chandramouleeswaran, D. Krzemien, K. Burns, and H. T. Tran, "Machine learning prediction of airport delays in the US air transportation network," in *Proc. Aviation Technol., Integr., Oper. Conf.*, 2018, p. 3672.
[16] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao, "Flight delay prediction based on aviation big data and machine learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 140–150, Jan. 2020.
[17] P. Monmousseau, A. Marzuoli, E. Feron, and D. Delahaye, "Predicting and analyzing US air traffic delays using passenger-centric data-sources," in *Proc. 13th USA/Eur. Air Traffic Manage. Res. Develop. Seminar (ATM)*, 2019, Paper 59. [Online]. Available: https://www.atmseminar.org/13th-seminar/papers-and-presentations/
[18] B. Sridhar and N. Y. Chen, "Short-term national airspace system delay prediction using weather impacted traffic index," *J. Guid., Control, Dyn.*, vol. 32, no. 2, pp. 657–662, 2009.
[19] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," in *Proc. IEEE/AIAA 35th Digit. Avionics Syst. Conf. (DASC)*, Sep. 2016, pp. 1–6.
[20] M. Zanin, Y. Zhu, R. Yan, P. Dong, X. Sun, and S. Wandelt, "Characterization and prediction of air transport delays in China," *Appl. Sci.*, vol. 10, no. 18, p. 6165, 2020.
[21] Z. Chen, Y. Wang, and L. Zhou, "Predicting weather-induced delays of high-speed rail and aviation in China," *Transp. Policy*, vol. 101, pp. 1–13, Feb. 2021.
[22] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, Feb. 2015.
[24] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, "A deep learning approach to flight delay prediction," in *Proc. IEEE/AIAA 35th Digit. Avionics Syst. Conf. (DASC)*, Sep. 2016, pp. 1–6.
[25] V. Venkatesh, A. Arya, P. Agarwal, S. Lakshmi, and S. Balana, "Iterative machine and deep learning approach for aviation delay prediction," in *Proc. 4th IEEE Uttar Pradesh Sect. Int. Conf. Elect., Comput. Electron. (UPCON)*, Oct. 2017, pp. 562–567.
[26] B. Yu, Z. Guo, S. Asian, H. Wang, and G. Chen, "Flight delay prediction for commercial air transport: A deep learning approach," *Transp. Res. E, Logistics Transp. Rev.*, vol. 125, pp. 203–221, Mar. 2019.
[27] J. Qu, T. Zhao, M. Ye, J. Li, and C. Liu, "Flight delay prediction using deep convolutional neural network based on fusion of meteorological data," *Neural Process. Lett.*, vol. 52, no. 2, pp. 1461–1484, 2020.
[28] Y. Jiang, Y. Liu, D. Liu, and H. Song, "Applying machine learning to aviation big data for flight delay prediction," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput., Int. Conf. Pervasive Intell. Comput., Int. Conf. Cloud Big Data Comput., Int. Conf. Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Aug. 2020, pp. 665–672.

[29] M. F. Yazdi, S. R. Kamel, S. J. M. Chabok, and M. Kheirabadi, "Flight delay prediction based on deep learning and Levenberg-Marquart algorithm," *J. Big Data*, vol. 7, no. 1, pp. 1–28, Dec. 2020.

[30] W. Zeng, J. Li, Z. Quan, and X. Lu, "A deep graph-embedded LSTM neural network approach for airport delay prediction," *J. Adv. Transp.*, vol. 2021, Mar. 2021, Art. no. 6638130.

[31] K. Cai, Y. Li, Y.-P. Fang, and Y. Zhu, "A deep learning approach for flight delay prediction through time-evolving graphs," *IEEE Trans. Intell. Transp. Syst.*, early access Aug. 12, 2021, doi: 10.1109/TITS.2021.3103502.

[32] Y. Ai, W. Pan, C. Yang, D. Wu, and J. Tang, "A deep learning approach to predict the spatial and temporal distribution of flight delay in network," *J. Intell. Fuzzy Syst.*, vol. 37, no. 5, pp. 6029–6037, 2019.

[33] N. Takeichi, R. Kaida, A. Shimomura, and T. Yamauchi, "Prediction of delay due to air traffic control by machine learning," in *Proc. AIAA Modeling Simulation Technol. Conf.*, Jan. 2017, p. 1323.

[34] L. Carvalho, A. Sternberg, L. M. Gonçalves, A. B. Cruz, J. A. Soares, D. Brandão, D. Carvalho, and E. Ogasawara, "On the relevance of data science for flight delay research: A systematic review," *Transp. Rev.*, vol. 41, no. 4, pp. 499–528, Jul. 2021.

[35] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[36] Z. Zhang, Y. Zhao, J. Liu, S. Wang, R. Tao, R. Xin, and J. Zhang, "A general deep learning framework for network reconstruction and dynamics learning," *Appl. Netw. Sci.*, vol. 4, no. 1, pp. 1–17, 2019.

[37] R. Rossi, A. Murari, and P. Gaudio, "On the potential of time delay neural networks to detect indirect coupling between time series," *Entropy*, vol. 22, no. 5, p. 584, 2020.

[38] X. Gao, W. Zhu, Q. Yang, D. Zeng, L. Deng, Q. Chen, and M. Cheng, "Time delay estimation from the time series for optical chaos systems using deep learning," *Opt. Exp.*, vol. 29, no. 5, pp. 7904–7915, 2021.

[39] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Müller, "Deep learning for time series classification: A review," *Data Mining Knowl. Discovery*, vol. 33, no. 4, pp. 917–963, Jul. 2019.

[40] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 1–8.

[42] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.

[43] W. Zhang, J. Tanida, K. Itoh, and Y. Ichioka, "Shift-invariant pattern recognition neural network and its optical architecture," in *Proc. Annu. Conf. Jpn. Soc. Appl. Phys.*, 1988, pp. 2147–2151.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[45] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[46] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *Proc. Int. Conf. Web-Age Inf. Manage.* Cham, Switzerland: Springer, 2014, pp. 298–310.

[47] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Exploiting multi-channels deep convolutional neural networks for multivariate time series classification," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 96–112, Feb. 2016.

[48] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[49] A. Gulli and S. Pal, *Deep Learning With Keras*. Birmingham, U.K.: Packt, 2017.

[50] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, Mar. 2001.

[51] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Phys. Rep.*, vol. 424, nos. 4–5, pp. 175–308, 2007.

[52] M. Zanin and F. Lillo, "Modelling the air transport with complex networks: A short review," *Eur. Phys. J. Special Topics*, vol. 215, no. 1, pp. 5–21, Jan. 2013.

[53] D. J. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[54] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, 2008, Art. no. P10008.

[55] R. Beatty, R. Hsu, L. Berry, and J. Rome, "Preliminary evaluation of flight delay propagation through an airline schedule," *Air Traffic Control Quart.*, vol. 7, no. 4, pp. 259–270, 1999.

[56] S. AhmadBeygi, A. Cohn, Y. Guan, and P. Belobaba, "Analysis of the potential for delay propagation in passenger airline networks," *J. Air Transp. Manage.*, vol. 14, no. 5, pp. 221–236, Sep. 2008.

[57] J.-T. Wong and S.-C. Tsai, "A survival model for flight delay propagation," *J. Air Transp. Manage.*, vol. 23, pp. 5–11, Aug. 2012.

[58] N. Kafle and B. Zou, "Modeling flight delay propagation: A new analytical-econometric approach," *Transp. Res. B, Methodol.*, vol. 93, pp. 520–542, Nov. 2016.

[59] C.-L. Wu and K. Law, "Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model," *Transp. Res. E, Logistics Transp. Rev.*, vol. 122, pp. 62–77, Feb. 2019.

[60] A. Rodríguez-Sanz, F. G. Comendador, R. A. Valdés, J. Pérez-Castán, R. B. Montes, and S. C. Serrano, "Assessment of airport arrival congestion and delay: Prediction and reliability," *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 255–283, Jan. 2019.

[61] S. Bratu and C. Barnhart, "An analysis of passenger delays using flight operations and passenger booking data," *Air Traffic Control Quart.*, vol. 13, no. 1, pp. 1–27, 2005.

[62] A. Cook, G. Tanner, and M. Zanin, "Towards superior air transport performance metrics–imperatives and methods," *J. Aerosp. Oper.*, vol. 2, nos. 1–2, pp. 3–19, 2013.

[63] C. Barnhart, D. Fearing, and V. Vaze, "Modeling passenger travel and delays in the national air transportation system," *Oper. Res.*, vol. 62, no. 3, pp. 580–601, 2014.

[64] A. Voltes-Dorta, H. Rodríguez-Déniz, and P. Suau-Sanchez, "Vulnerability of the European air transport network to major airport closures from the perspective of passenger delays: Ranking the most critical airports," *Transp. Res. A, Policy Pract.*, vol. 96, pp. 119–145, Feb. 2017.

[65] H. Tam, E. S. Ching, and P.-Y. Lai, "Reconstructing networks from dynamics with correlated noise," *Phys. A, Stat. Mech. Appl.*, vol. 502, pp. 106–122, Jul. 2018.

[66] T. P. Peixoto, "Network reconstruction and community detection from dynamics," *Phys. Rev. Lett.*, vol. 123, no. 12, Sep. 2019, Art. no. 128301.

[67] V. Salnikov, D. Cassese, and R. Lambiotte, "Simplicial complexes and complex systems," *Eur. J. Phys.*, vol. 40, no. 1, 2018, Art. no. 014001.

[68] O. T. Courtney and G. Bianconi, "Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 93, no. 6, 2016, Art. no. 062311.

[69] G. Tirabassi, R. Sevilla-Escoboza, J. M. Buldú, and C. Masoller, "Inferring the connectivity of coupled oscillators from time-series statistical similarity analysis," *Sci. Rep.*, vol. 5, no. 1, pp. 1–14, 2015.

[70] Z.-K. Gao, Y.-X. Yang, W.-D. Dang, Q. Cai, Z. Wang, N. Marwan, S. Boccaletti, and J. Kurths, "Reconstructing multi-mode networks from multivariate time series," *Europhys. Lett.*, vol. 119, no. 5, Sep. 2017, Art. no. 50008.

[71] E. Peluso, T. Craciunescu, and A. Murari, "A refinement of recurrence analysis to determine the time delay of causality in presence of external perturbations," *Entropy*, vol. 22, no. 8, p. 865, 2020.

[72] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–6.

[73] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham, Switzerland: Springer, 2018, pp. 19–36.

[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

ILINKA IVANOSKA was born in Skopje, North Macedonia, in 1986. She graduated at the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Skopje, in 2009. She received the master's degree and the Ph.D. degree in computer science from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, in 2013 and 2021, respectively.

She is currently an Assistant Professor with the Faculty of Computer Science and Engineering, Institute for Intelligent Systems, Ss. Cyril and Methodius University. Her current work in teaching includes courses, such as algorithms and data structures, object-oriented analysis and design, discrete mathematics, information systems, and intelligent information systems. Areas of her scientific research interests include machine learning, complex networks, and bioinformatics.

LUISINA PASTORINO was born in Argentina, in 1990. She graduated in industrial engineering from the National University of Rosario, Argentina, in 2016. She received the master's degree in big data analysis in economics and business from the University of the Balearic Islands, Spain, in 2021. She is currently pursuing the Ph.D. degree with the Institute for Cross-Disciplinary Physics and Complex Systems, Palma de Mallorca, Spain.

She focuses on the study of the propagation of delays in the air transport system from an information processing perspective within the ARCTIC project.

MASSIMILIANO ZANIN was born in Verona, Italy, in 1982. He received the Ph.D. degree in computer engineering from the Universidade Nova de Lisboa, Portugal, in 2014.

He is currently a Researcher with the Institute for Cross-Disciplinary Physics and Complex Systems, Palma de Mallorca, Spain. His main topics of interests include complex networks and data science, both from a theoretical perspective and through their application to several real-world problems. Throughout his career, he has published more than 80 articles in international journals and more than 50 contributions in conferences, reaching an H-index of 25. He has participated in several European competitive research projects, and he is currently a PI of the ERC StG ARCTIC, on the analysis and modelling of delay propagation in air transport.