A novel methodology to predict regulations using Deep Learning

Sergi Mas-Pujol
Computer Architecture Department
Universitat Politècnica de Calatunya
Barcelona, Spain
sergi.mas.pujol@upc.edu

Esther Salamí
Computer Architecture Department
Universitat Politècnica de Calatunya
Barcelona, Spain
esther.salami@upc.edu

Enric Pastor
Computer Architecture Department
Universitat Politècnica de Calatunya
Barcelona, Spain
enric@ac.upc.edu

***Abstract*—The current air traffic control system tries to allocate as many flights as possible in a scenario that is expected to be time-efficient, cost-efficient, and safe. To guaranty these safety conditions, it is performed a cyclic process known as Demand-Capacity Balancing. During this process, a specialized Air Traffic Controller analyses the situations where the demand is over the capacity to identify the required corrective actions. These corrective actions are mostly in the form of regulations, and they are necessary to avoid overload during the day of operation. The task of declaring a regulation is complicated, very time-consuming, and based on the Air Traffic Controller's experience. A massive amount of information must be considered simultaneously, together with a risk maturation process because of the uncertainty and granularity in the information. This paper proposes and evaluates two Deep Learning models able to mimic the current procedure's behavior, and therefore, helping the specialized Air Traffic Controller to automatically detect the imbalances that will require regulation. Both models, one based on Convolutional Neural Networks, and the second one based on Recurrent Neural Networks, have demonstrated the potential to predict regulations, with an accuracy of 81.45% and 80.73% respectively over the entire MUAC region in 30-minute intervals. This accuracy can be increased by up to 91% by developing specialized models for each airspace sector. Additionally, we performed an in-depth analysis of the most relevant features using *SHapley Additive exPlanations*.**

***Keywords*—Demand-Capacity Balancing, Regulations, Deep Learning**

I. Introduction

In the coming years, Air Navigation Service Providers (ANSPs) will have to handle and accommodate a continuously increasing traffic demand, in a scenario that is expected to be more time-efficient and cost-efficient [1]. Therefore, the most challenging problem facing the Air Traffic Management (ATM) will be to meet the airspace sectors' capacity with the growing demand. While at the same time, safety levels must be maintained or increased.

An airspace sector's capacity is directly related to the number of simultaneous flights an ATCO can safely manage [2], and the ANSPs establish it since they design the airspace's configuration and define the operative sectors during the day of operations (D0).

The process of ensuring that the demand is under the capacity, in all possible circumstances, is known as Demand-Capacity Balancing (DCB). The Air Traffic Flow and Capacity Management (ATFCM) role is in charge of this task, which starts months in advance to the day of operation. Typically, when the airlines, or flight operators, submit the initial flight plans. Therefore, it is a cyclic process whose primary goal is to ensure that no ATCO has managed a sector where the workload is above a predefined threshold.

First of all, automatic tools will report the locations (place and time) where it is detected a higher demand than the predefined capacity (imbalance). Second, this imbalance is studied manually by the ATFCM. Third, if regulation is declared, the most optimal DCB-measure is implemented.

Figure 1 shows an approximation of the most relevant factors at different time horizons concerning D0.

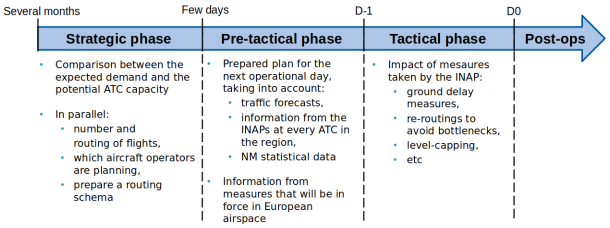

Figure 1. Main phases and available information depending on the time horizons with respect to D0.

From several days to 4 hours before the congested area event, the ATFCM must guarantee a non-excessive density of aircraft, mainly applying delays. Then, from 4 hours to 15 minutes before the entry of a flight in a congestion area, the Dynamic DCB process starts. In this case, using the available traffic forecast, more specific metrics can be used to control the workload of the ATCOs. Notice that no detecting them implies that there will be an airspace sector, during the day of operation, where there are not enough ATCOs to handle both the current and incoming traffic.

In this scenario, we propose to use a Deep Learning (DL) system to establish the relationships between the reported imbalances by the current system and the decision-making process executed by the ATFCM to regulate them. More precisely, we want to automatically detect those imbalances that have to be solved, and reduce the required cognitive resources (e.g., time spent per task, or required knowledge). Therefore, the ATFCM could cover a larger volume of traffic or focus

on other tasks. Moreover, it would allow other stakeholders involved in ATM to have more accurate information in advance and anticipate corrective actions.

Finally, to achieve the presented goal, we will consider both the *visual representation* of the information used and the corresponding *scalar values*. By visual representation, we refer to artificial images showing the airspace's configurations, such as the aircraft's coordinates, its flight level (FL), its flight phase (climbing, cruising or descending), and its heading. While by scalar variables, we refer to the exact numbers of flights inside the sectors, the number of incoming flights at different time horizons, or the number of flights in a specific phase.

II. Demand-Capacity Balancing

We have used to term *ATFCM* to refer to the specialized ACTO whose main tasks focus on DCB activities because it was the first acronym established in the European community. Nevertheless, several names are associated with this position/task: ATFCM, Flow Manager Position (FMP), and Integrated Network management & ATC Planning (INAP). Since INAP is is the most recent acronym, it is the one we will use in the rest of the paper.

DCB-measures need to be organized in such a way that common situation awareness (network information sharing) and Collaborative Decision Making (CDM) processes are maintained between all ATM actors, enabling the various organizations to continuously adjust their actions according to the most up-to-date DCB events, to optimize them as much as possible. Therefore, two different kinds of DCB-measures are available:

- DCB-measures at the Network Manager (NM) level consists of the assignation of ground delay, using the CASA algorithm [3], to solve the required imbalances at the early stages. They are known as **Regulations**,
- DCB-measures at sector level consists of the fine-tuning of the sector configuration (e.g., level-capping, grouping/splitting of sectors), or smaller modifications in the aircraft's trajectories (e.g., rerouting). They are taken a few hours before the flight enters the congested sector, and they are known as **Cherry-Picking Measures**.

Before going into more detail in the next subsection, Fig. 2 summarises the main steps done during the DCB process.

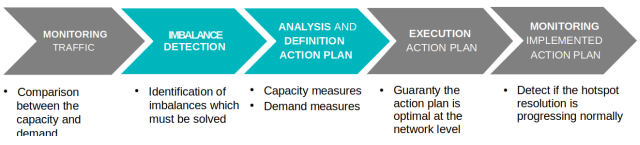

Figure 2. Main steps and metrics used to detect and solve capacity-demand imbalances.

A. Imbalance detection

During the imbalance detection step, the INAP can use many metrics. But, the most common are the expected flights entering to the sector in a given interval of time (Entry Count - EC), the number of flight inside the sector (Occupancy Count - OC), and a local analysis (additional metrics used to have a better intuition of the workload that will have the ATCO who is handling the traffic).

Some of the available metrics for the *local analysis* are the number of interactions between flights or the number of flights in a specific flight phase. On the other hand, a more complex metric is the overall complexity of a sector (a mathematical model that combines some of the primary metrics).

Therefore, to guaranty a safe level of workload and the application of the required measures, there is a set of predefined thresholds for some of these metrics [4]:

- The **Peak** threshold represents the maximum number of flights that can be handled, in a sector, at the same time. When the $count > peak$, it indicates a potential overload.
- The **Sustain** threshold represents an acceptable number of flights that can be handled in a sector under specific circumstance, and in particular, if the duration of the overload is not too long. When the $count > sustain$, and $count < peak$, it also indicates a potential overload.
- The **Overload duration** threshold represents the maximum duration beyond which an imbalance should be considered in case of $count > sustain$.

The previous thresholds refer to potential/possible overloads (when the ATCO's workload is too high). Nevertheless, the process of determining if each of these imbalances will require a DCB-measure is purely done by a human being through experience, knowledge, and skills.

As an illustration, the criteria to identify an overload could be: The occupancy count is above the *peak* threshold, or the occupancy count is continuously between *sustain* and *peak* for 20 minutes or longer. However, before declaring the start/end time of the overload, the INAP must perform a deep analysis of the traffic in that airspace sector to determine the following key parameters:

- Predictability: assessing data uncertainty, granularity, or integrity based on the flight status (e.g., planned, confirmed) to evaluate the quality/precision of the information.
- Complexity per flight: the contribution of each flight, or the contribution of a flow, to the overall complexity.

Finally, Fig. 3 shows how all the previous information is combined to identify imbalances that must be solved.

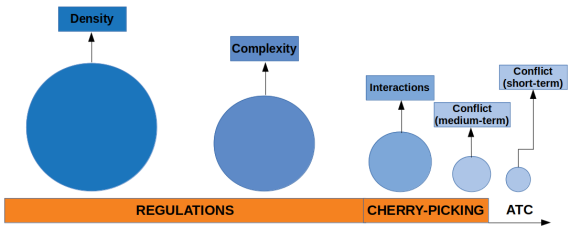

Figure 3. Metrics used according to the time horizon with respect to the D0. The size of the circles and the color-map is a visual representation of the uncertainty, and granularity in the information according to the time horizon.

Notice that according to the time horizon to D0, the INAP prioritizes some metrics over others. When the time horizon is broad (higher uncertainty), more generic metrics are considered, such as the density. While when the time horizon is close to D0 (less uncertainty), more accurate metrics can be used, such as the interactions between specific aircraft pairs.

III. State of the art

One important aspect to highlight is that a system based on DL works with the patterns that it discovers in the data. The algorithm learns only from the data we provide, so the success of the system depends mainly on the quality and quantity of the input data. If the data is not well selected, clean, and transformed, even the best algorithm will give us low-quality predictions. Therefore, we will start by identifying the possible *key features* of the problem. Next, we will focus on the research done for *overload detection and solving* using DL. And finally, due to we want to mimic a human cognitive capability, we will review the literature related to *AI applied to Collaborative Decision Making* for ATM.

For a more thorough view of the DL field, the reader may refer to [5], and [6].

A. Identification of the key features

Over the years, many different factors have been used to try to estimate the complexity of a given ATC sector or to quantify the controllers' workload.

In [7], the authors presented a multi-year, multi-organizational research initiative related to the measurement and prediction of sector-level complexity using Dynamic Density (DD). This concept is based on *traffic characteristics*, and *"it is a function based on the number of aircraft and their changing geometries for a given airspace sector" [7]*. Similarly, in [8], the authors presented a model to quantify the workload impact using traffic density, sector geometry, flow direction, and air-to-air conflict rates.

It is interesting to notice that, in the previous publications, only *traffic characteristics* were used, and only results for specific sectors were reported. However, in [9], the authors used Neural Networks (NNs) to predict the controllers' workload mainly focusing on *cognitive factors* (e.g number of keystrokes), or focusing on *physiological factors* (e.g heart rate or electrocardiogram). In this publication, the authors concluded that *"the NN method with complexity measures is successful in predicting controller workload" [9]*, nevertheless, they realized that the *physiological factors* are not the best factors for measuring the workload due to the job of an ATCO is primarily cognitive and information-intensive, rather than physical and labor-intense.

Finally, the most recent work-related to forecasting the air traffic controller workload is [10]. They compared several ML methods on the problem of learning a model of the ATCO's workload from historical data. They started analyzing the existing metrics of complexity and doing a Principal Component Analysis (PCA) to found the most representative factors. And after the analysis, their results showed that the most accurate method was the NN with an 82% of accuracy, and the most relevant factors for building airspace configuration prediction models are the following ones:

- The airspace volume of the considered ATC sector;
- The number of aircraft within the sector boundaries at time t;
- The incoming traffic flow within the next 15 minutes;
- The incoming traffic flow within a 1 hour time horizon;
- The average absolute vertical speed of the aircraft within the sector;
- The number of speed vector intersections with an angle greater than 20 degrees.

Notice that the previous metrics were not used to predict or detect congested sectors, they were developed to better estimate the complexity of a sector. In other words, our objective goes beyond complexity metrics, by trying to directly predict regulations.

B. Overload detection and solving using Deep Learning

To our knowledge, there is no previous research literature conducted for imbalance detection, or DCB, using supervised learning. Nevertheless, some approaches have been presented using reinforcement learning, where the imbalance detection and resolution problem was faced indirectly.

Focusing on DCB in the pre-tactical phase, in [11] the authors presented DART (Data-driven Aircraft Trajectory Prediction Research), which is composed by a data-driven trajectory prediction (individual trajectory prediction), and agent-based collaborative learning. Moreover, they took into account additional information such as calendar properties, weather, and aircraft characteristics. The aim here is to overcome the fact that existing ATM information is not accurate enough during this phase. Similarly, in [12], it was formalized the problem as a multi-agent Markov decision process (MDP) towards deciding flight delays to resolve DCB problems in ATM. The presented formulation allows agents to interact, and form their policies in coordination with others.

Another example of research done using reinforcement learning, and trying to resolve demand-capacity imbalances is [13]. In this case, it is interesting the fact they conclude that: *"Two important drawbacks of such prediction methods are that (a) they are limited to single trajectory predictions, and (b) their prediction horizon is a short time one. [13]"*.

Notice that these two last approaches imply a different paradigm of behavior, however, they can be of interest, as they can consider the uncertainty and the temporal evolution.

C. AI applied to Collaborative Decision Making

Advances in AI-integrated decision making support systems, or intelligent decision support systems (IDSS), are increasingly used to assist decision making in such areas as finance, healthcare, marketing, commerce, and cybersecurity.

Pioneer work was done in [14], where it was proposed and validated a new organization for ATC which allows ATCOs to stay active in the control and supervisory loop of the process

to maintain the present traffic safety level and to improve the global system performances.

In [15], the authors explored several approaches focusing on look-ahead reasoning whose main components are uncertainty and preferences. And similarly, in [16], it was presented an ATCO Psychological Model that considers two main components: the functional structure of the ATCO cognitive system, and the attentional resources needed. The authors reported a Root-Mean-Square Error (RMSE) of 0.76.

IV. Methodology

This paper focuses on identifying **Regulations** during the pre-tactical phase over the MAUC region. To detect/predict them, we have used two types of inputs: artificial images and scalar variables.

From the images, we want to extract the airspace configuration, the locations, and interactions between aircraft (overall situation) using Convolutional Neural Networks (CNN). On the other hand, we have extracted specific metrics such as the estimated workload or the number of flights entering the sector in the following minutes as scalar variables. In this case, we have used a Recurrent Neural Networks (RNN) to process them.

In both cases, we have generated the input samples using information/data from the AIRACs (detailed description of the aerospace configuration for a specific period) used in the R-NEST (model-based simulation tool). Concretely, we used the AIRACs from June, July, August, and September from 2019.

A. CNN-based model

CNNs are mainly used to process and extract features from static images (e.g., image classification), but we want an architecture able to handle images that evolve on time. Therefore, our samples will be composed of multiple images representing the airspace configuration at different consecutive time steps (see Fig. 4).

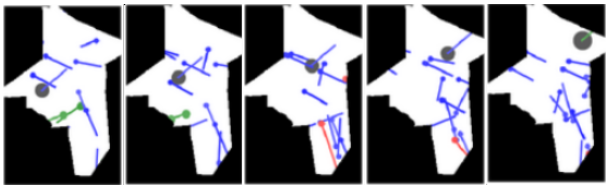

Figure 4. Example of an input sample composed by five time steps. The grey points show the path of a unique aircraft. The colors show the different flight phases (climbing in green, cruising in blue, and red descending). The points show the location of the aircraft, and their size reflects the FL. The lines show the heading, and the length represents the speed. Finally, the black mask expresses the shape of the airspace sector (in this case EDYYB3EH).

To process the previous sets of images, we have implemented an architecture equivalent to Fig. 5. This approach captures the temporal information since the images per set are processed in parallel.

Finally, they will be evaluated at the time steps level, but also at the interval level (see Sec. IV-C).

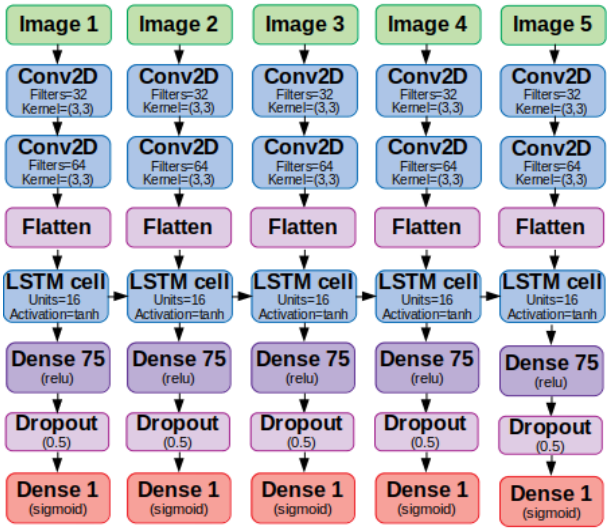

Figure 5. Architecture of the CNN used. *Conv2D* correspond with a 2D CNN layer, *Dense X* express a dense layer composed by *X* neurons, and rows show the connections between layers. Parallel processing has been done using the *TimeDistributed* function in the framework Keras. Notice that only five input time steps are shown for simplicity.

B. RNN-based model

An RNN is a class of NN where information travels in loops from layer to layer so that the state of the model is influenced by its previous states allowing it to exhibit temporal dynamics.

For this approach, the model will receive as input multiple time steps with a combination of the most basic scalar variables and those extracted in [10] after the PCA analysis:

- Timestamp (associated interval of the studied day);
- Capacity of the sector;
- Occupancy count;
- Entry count next 20 and 60 minutes;
- Expected workload;
- Number of conflicts;
- Number of flights at the different phases (climbing, cruising and descending)

Figure 6 is a graphical representation of the architecture used.

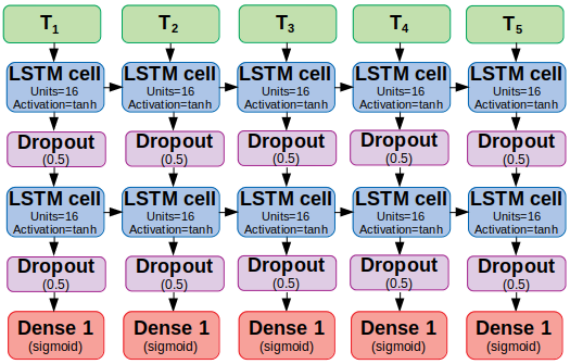

Figure 6. Architecture of the RNN used. T_X refers to the time step *X* in the input sample.

The output will be a prediction for each time step, as in the CNN model. Therefore, the model will be evaluated at the time steps level and for the entire interval (see Sec. IV-D).

Finally, we have decided to use a Long-Short Term Memory (LSTM) in the two previous models because they have shown better performance, for the problem we are facing, than Gated Recurrent Units (GRU) or purely RNN's cells.

C. Evaluation metrics

To evaluate the performance of these models, we will perform two analyses. The first one is based on analyzing the prediction per input time step (**time step classification metrics**). And the second one is based on analyzing how good are our models for the entire given interval (**interval classification metrics**)

Time step classification metrics

This first analysis consists of analyzing the accuracy, recall, and F1-score of the predictions done by the model at each input time step (for instance, every five minutes):

- **Accuracy**: The fraction of predictions our model got right.
  - $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- **Recall**: Attempts to answer the question of what proportion of the positive time steps were correctly identified.
  - $Recall = \frac{TP}{TP+FN}$
- **Precision**: Attempts to answer the question of what proportion of positive identifications were correctly identified.
  - $Precision = \frac{TP}{TP+FP}$
- **F1 score**: It is the harmonic mean of the precision and recall, or it can be interpreted as a weighted average of the precision and recall
  - $F1\ score = 2\frac{Precision*Recall}{Precision+Recall}$

TP refers to correct positive predictions, *TN* refers to correct negative predictions, *FP* refers to wrong positive predictions, and *FN* refers to wrong negative predictions.

Interval classification metrics

In the second analysis, we want to evaluate the model's performance predicting the overall situation, that is, to predict whether the entire given interval contains a regulation or no. The reason for this is because predicting exactly the starting and ending moment of regulations is a challenging task. Figure 7 is a graphical example of the explained issue.

Figure 7. In green ground truth sample, and in red a predicted sample. The symbol *X* shows the time steps which required a regulation. (Left) Samples per time step. (Right) The grouped samples.

Therefore, we propose an **Interval analysis** where we will group the labels from both the ground truth and the predictions to determine if each sample contains information from a regulated period or not. Then, we will perform the previously explained analysis of the accuracy, recall, precision, and F1 score. Positive samples will be the ones containing information from periods with regulations, and negative samples, the ones with information from none regulated periods.

Notice that this *Interval analysis* requires a threshold to determine how many positive time steps are required to determine that a sample belongs to a regulated period. This threshold will be set to one because we consider it more critical no detecting a regulation than a false-positive case.

Finally, to complement the analysis, we will perform a novel **Matching analysis**. In this case, we will compare the ground truth and the predictions allowing the model to have mismatches in only a few time steps, ensuring that the overall situation is coincident.

For instance, using samples of 30 minutes and a threshold of 85%, we will be detecting predictions with less than five incorrect time steps.

Finally, we will compute three additional metrics:

- **Perfect matching**: Proportion of predicted samples that exactly match the ground truth,
- **Strong matching:** Proportion of predicted samples that do not exactly match the ground truth, but they are above the threshold,
- **Weak matching:** Proportion of predicted samples under the specified threshold.

D. Model explainability

Explainable machine learning offers the potential to provide the stakeholders with insights into model behavior. Moreover, understanding the reasons behind predictions is crucial in assessing trust when we want to take action based on them.

It is often the case that deep neural networks are considered "black boxes". In response, various methods have recently been proposed to help users interpret the predictions of complex models.

SHapley Additive exPlanations (SHAP) [17] is a game theory approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the traditional Shapley values from game theory and their related extensions. It compares each neuron's activation and assigns contribution scores (input feature) by optionally giving separate consideration to positive and negative contributions.

In our case, we will perform the SHAP analysis on the RNN, aiming to identify which input features are more relevant for the trained model over the entire MUAC region. Therefore, we will represent the different features showing the relationship between how larger (or smaller) their values are, respect to the activation they generate.

V. Results

First, we will perform a statistical analysis of the data we are going to use. Second, we will present the results obtained usind the two proposed models: the CNN-based model that uses images as inputs and the RNN-based model that uses scalar variables as inputs. Finally, we will analyse the relevance of the input features in the RNN-based model using SHAP.

A. Input data

We have access to the data from 112 days (28 days per AIRAC and four AIRACs) with different types of regulations (see Table I). To label our samples, we will use the ones related to *en route* traffic and associated with demand regulations. However, purely capacity regulations were taken into account when showing this feature to the models.

TABLE I. Available regulations in the AIRACs. *C-ATC* is the only regulation associated to demand regulations for *en route* flights.

Regulations	Number of instances
C-ATC	132
S-ATC	18
M-Airspace	11
W-Weather	106
O-Others	4

Looking at the available data, the regulations we used to label our samples were tagged as *C-ATC*. A statistical analysis of the data showed that:

- 71 days had a demand regulation,
- 41 days had no demand regulations,
- The average duration of the regulations was 122.02 mins,
- And the average number of regulations per day were 2.5

Continuing with the statistical analysis, if we study the intervals of time with and without regulations, and we compare the *Occupancy Count* and the *Entry Count* with their associated thresholds, we can see that:

1) 7.4% of the minutes, from regulated periods, had an OC higher than the peak threshold, and 4.6% of the minutes had an OC between the sustained and the peak thresholds. On the other hand, for non-regulated periods, 7% of the minutes had an OC higher than the peak threshold, and 4.7% of the minutes had an OC between the sustained and the peak thresholds,
2) If we analyze the EC for the next 20 minutes for the regulated periods, 24.4% of them were above the peak threshold, and 10.3% of the minutes were between the sustained and the peak thresholds. For the no regulated periods, the analysis showed that 25.1% of the minutes were over the peak threshold, and 10.2% of them were between the sustained and the peak thresholds,
3) Finally, analyzing the EC for the next 60 minutes for the regulated periods, 25.1% of the minutes had an EC higher than the peak threshold, and 10.2% of them had an EC between the sustained and the peak thresholds. For no regulated periods, 38.9% of the cases had an EC above the peak threshold, and 7.9% of them were between the sustained and the peak thresholds.

Notice that we have only analyzed the scalar variables associated with the *OC* and the *EC* because they are the only ones with predefined thresholds (no all the metrics have an associated threshold, or we do not have access to all of them). However, the interesting part is that the proportion of samples in the previous analysis is very similar in both scenarios. The reason is that determining if an imbalance will become a regulation is not done just by looking at one of the above characteristics, but a combination of all of them is used. Moreover, it is necessary to consider the prolongation in time of these imbalances or their severity.

Finally, regarding the input samples, the random intervals of time we are going to use can contain time steps of the following three types:

- No regulation, from days without regulations;
- No regulation, from days with regulations;
- And regulation.

This fact will allow the models to detect in which precise moment there is an overload. For instance, if there was a regulation from 7:00 pm to 9:00 pm, and an input sample covers the interval from 6:45 pm to 7:15 pm, the model will only show a positive label for the time steps inside the regulated period (from 7:00 pm to 7:15 pm).

Finally, from the four available AIRACs, three have been used for training and the fourth for testing. Furthermore, we have discarded some samples in order to have a balanced dataset formed by the same number of positive and negative time steps (half of the negative samples were extracted from days without regulations and half of them from days with regulations). At the end, the dataset used consists of approximately 1200 30-minutes intervals (70% of them for training and 30% for testing).

B. CNN-based model

Generating the images and processing them requires a vast amount of computational resources. Therefore, we have used seven consecutive images (one every five minutes) to represent a given interval of 30 minutes.

We have tested several scenarios (e.g., more or fewer images, bigger or smaller intervals of time), and in all of the scenarios, we have obtained a similar performance. Therefore, we have decided to continue developing this one because it has good performance with a reasonable computational time.

Table II shows the *Time step analysis* using the presented scenario and trying to predict regulations over the entire MUAC region. Table III displays the values obtained performing the *Interval analysis* to visualize how good is this model detecting whether the complete given input interval will require regulation or not (overall situation). Finally, Table IV exhibits the results from the *Matching analysis*, allowing the model to have mismatches in less than 5 minutes.

TABLE II. **TIME STEPS ANALYSIS, CNN, AND MUAC REGION.** THE COLUMN *Train/Test* SHOWS THE NUMBER OF SETS OF SAMPLES USED.

Accuracy(%)	Recall(%)	Precision(%)	F1 score[0,1]	Train/Test
78.68	77.80	83.13	0.80	840/369

TABLE III. **INTERVAL ANALYSIS, CNN, AND MUAC REGION.**

Accuracy(%)	Recall(%)	Precision(%)	F1 score[0,1]
81.45	92.64	78.13	0.84

TABLE IV. **MATCHING ANALYSIS, CNN, AND MUAC REGION.** *Strong* REFERS TO SAMPLES WITH LESS THAN 1 MISMATCHES (5 MINUTES). *Valid predictions* IS THE COMBINATION OF BOTH THE *Perfect* AND *Strong*.

Perfect(%)	Strong(%)	Weak(%)	Valid predictions(%)
28.32	50.72	20.96	79.04

From this experiment, we can conclude that the CNN architecture has an accuracy close to an 80%, with high recall and precision. The main drawbacks of this model are the difficulty of taking into account the incoming traffic (EC) due to only information from previous images (from the same input set) can be used as incoming traffic, and the limited resolution (seven time steps per input set). Nevertheless, during the *Interval analysis*, the model reported a recall of 92% detecting, in most cases, the intervals that required regulation.

These promising results indicate that it could be used for the presented task and for more specialized problems, such as spatially identifying the region that must be regulated.

C. RNN-based model

The RNNs require much less computational resources during the training. Therefore, for a given interval of 30 minutes, we have decided to extract the selected scalar variables at every minute, having 30 inputs and 30 outputs per sample.

Table V shows the *Time step analysis*. Table VII exposes the results obtained computing the *Matching analysis*. And Table VI, presents the *Interval analysis*.

TABLE V. **TIME STEPS ANALYSIS, RNN, AND MUAC REGION.** THE COLUMN *Train/Test* SHOWS THE NUMBER OF SETS OF SAMPLES USED.

Accuracy(%)	Recall(%)	Precision(%)	F1 score[0,1]	Train/Test
76.68	86.23	79.57	0.81	1030/343

TABLE VI. **INTERVAL ANALYSIS, RNN, AND MUAC REGION.**

Accuracy(%)	Recall(%)	Precision(%)	F1 score[0,1]
80.73	100	75.87	0.86

TABLE VII. **MATCHING ANALYSIS, RNN, AND MUAC REGION.** *Strong* REFERS TO SAMPLES WITH LESS THAN 5 MISMATCHES (5 MINUTES).

Perfect(%)	Strong(%)	Weak(%)	Valid predictions(%)
49.46	39.89	10.65	89.35

Notice that, as in the previous case, several intervals of time have been tested, and this configuration showed excellent results with a reasonable amount of computational resources. Moreover, it allows us to have the two models (CNN & RNN) working with the same temporal window.

From the previous results, we can conclude that the RNN model presents a 10% increase in the recall, and an equivalent accuracy when analyzing the predictions per time steps. Furthermore, it can be seen a 10% improvement (from 80% in the CNN to 90% in RNN) in the *Matching analysis*, if we allow the model to have 5% of mismatches.

Finally, it is crucial to notice that using this model, we have achieved a recall equal to 100% in the *Interval analysis*, and therefore, we have detected all the intervals which contained a regulation. However, due to the very restrictive way of grouping samples within an interval, it is also predicted as positive 25% more time steps than necessary.

D. RNN-based model for specific airspace sectors

Due to the good results obtained in the previous experiment, we have decided to extend the analysis and figure out if our best model can work only using regulations from specifics sectors. The main reason for this experiment is to verify whether we can improve the model's performance by specializing it for a particular airspace sector.

However, this experiment only can be carried out for the top three sectors. We have to guarantee enough number of instances in the sector to have enough variety in the samples.

TABLE VIII. **TIME STEPS ANALYSIS, RNN, AND SPECIFIC SECTORS.** THE COLUMN *Train/Test* SHOWS THE NUMBER OF SETS OF SAMPLES USED.

Sector	Accuracy(%)	Recall(%)	Precision(%)	F1 score[0,1]	Train/Test
BOLN	90.95	98.11	85.51	0.94	274/119
B3EH	84.14	92.98	70.51	0.88	227/99
D6WH	80.04	88.82	79.61	0.84	237/107

TABLE IX. **INTERVAL ANALYSIS, RNN, AND SPECIFIC SECTORS.**

Sector	Accuracy(%)	Recall(%)	Precision(%)	F1 score[0,1]
BOLN	91.51	100	87.32	0.93
B3EH	83.54	100	73.47	0.84
D6WH	83.22	100	75.73	0.86

TABLE X. **MATCHING ANALYSIS, RNN, AND SPECIFIC SECTORS.** *Strong* REFERS TO SAMPLES WITH LESS THAN 5 MISMATCHES (5 MINUTES).

Sector	Perfect(%)	Strong(%)	Weak(%)	Valid pred.(%)
BOLN	67.22	28.57	4.21	95.79
B3EH	31.31	64.6	4.09	95.91
D6WH	54.37	33.75	11.88	88.12

The results show that these specialized models have an equivalent performance of the model that handles regulation from the entire MUAC region in the worse case. In the best case, it has improved the overall performance by 10%. However, more AIRACs are required to do proper validation.

E. Model explainability

As mentioned, we will apply the SHAP analysis on the RNN (see Fig. 8) to understand the reason behind the predictions and have a better intuition of its behavior. Notice that this study can not be applied to the CNN because of the input samples' nature.

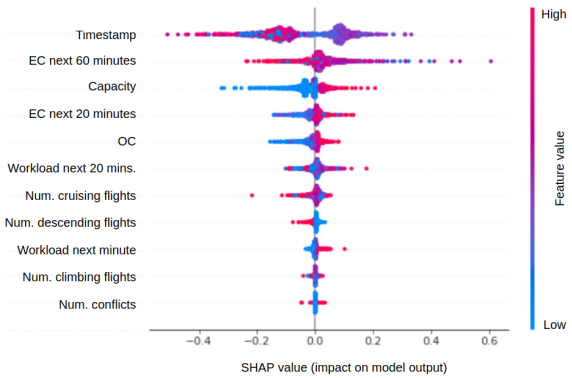

Figure 8. From top to bottom the image shows the more relevant features. The color map indicates how larger or smaller was the value of the input feature, and the location in the corresponding horizontal line represent the activation it generated. The zero in the *X-axis* represent no contribution to the prediction.

From the *Timestamp* feature, we can see that are detected more regulations at the early stages of the day, probably, because they want to avoid propagation of an overload along the day. On the other hand, regulations in the afternoon are less likely because of the traffic reduction. If we analyze the *Entry Count for the next 60 minutes*, it is surprising not to see a clear pattern of behavior. The reason could be that it is a really relevant feature, but in combination with another (similar for the *cruising flights)*. Remember, we have seen that information from multiple features is used to make the decisions. However, the *Entry Count for the next 20 minutes* shows the opposite trend. We can clearly see how larger countings are producing more activation. The *Capacity* can be seen as a crucial feature when it has a higher value. The higher the capacity, the larger the sector, and therefore, more aircraft and it is more likely they generate an overload. From the *Occupancy Count* and the *Workload every minute*, we can see that positive values have a higher activation. The *Number of descending aircraft* shows the opposite trend. Finally, the rest of the features do not appear to be decisive.

VI. Conclusion

We have proposed and evaluated two models capable of detecting situations that require regulation. Although the CNN-based model exhibits slightly higher accuracy than the RNN-based model (81.45% in front of 80.73% at the *interval level*), the RNN-based model achieved the maximum recall (100% in front of 92.64%). This means that it is detecting all the cases in which a regulation is needed, which in this particular scenario is preferable to high precision, that is, a low number of false positives. Moreover, accuracy can be increased by up to 91.51% developing specialized models for each airspace sector. On the other hand, if it is preferred more specific information, the *Time steps analysis* showed and accuracy of 90.95% and recall of 98.1% using specialized models.

Despite the good results obtained, further analysis of some of the hyperparameters is required to fine-tune the models, together with a deeper analysis of the false-positive cases.

Finally, due to the excellent performance of both models independently (CNN and RNN), we will study whether exist a hybrid model that could take advantage of each one and increase the precision while maintaining a high recall.

Acknowledgment

This work was funded EUROCONTROL under Ph.D. Research Contract No. 18-220569-C2 and by the Ministry of Economy, Industry, and Competitiveness of Spain under GrantNumber TRA2016-77012-R.

References

[1] EUROCONTROL, "Seven-year forecast september 2015, flight movements and service units 2015-2021," 2015.

[2] M. Ball, C. Barnhart, G. Nemhauser, and A. Odoni, "Air transportation: Irregular operations and control," *Handbooks in operations research and management science*, vol. 14, 2007.

[3] S. Niarchakou and J. Simón Selva. Atfcm operations manual–network operations handbook, 21.2017. [Online]. Available: www.eurocontrol.int/sites/default/files/content/documents/nm/network-operations/HANDBOOK/ATFCM-Operations-Manual-next.pdf.

[4] E. D. E. EUROCONTROL, DFS, "Air transport framework: The current situation," *SESAR Joint Undertaking*, 2015.

[5] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[6] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[7] P. Kopardekar and S. Magyarits, "Measurement and prediction of dynamic density," in *Proceedings of the 5th USA/Europe Air Traffic Management R & D Seminar*, vol. 139, 2003.

[8] J. D. Welch, J. W. Andrews, B. D. Martin, and B. Sridhar, "Macroscopic workload model for estimating en route sector capacity," in *Proc. of 7th USA/Europe ATM Research and Development Seminar, Barcelona, Spain*, 2007, p. 138.

[9] G. Chatterji and B. Sridhar, "Measures for air traffic controller workload prediction," in *1st AIAA, Aircraft, Technology Integration, and Operations Forum*, 2001, p. 5242.

[10] D. Gianazza, "Learning air traffic controller workload from past sector operations," 2017.

[11] E. C. Fernández, J. M. Cordero, G. Vouros, N. Pelekis, T. Kravaris, H. Georgiou, G. Fuchs, N. Andrienko, G. Andrienko, E. Casado *et al.*, "Dart: a machine-learning approach to trajectory prediction and demand-capacity balancing," *SESAR Innovation Days, Belgrade*, 2017.

[12] T. Kravaris, C. Spatharis, K. Blekas, G. A. Vouros, and J. M. C. Garcia, "Multiagent reinforcement learning methods for resolving demand-capacity imbalances," in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018.

[13] T. Kravaris, G. A. Vouros, C. Spatharis, K. Blekas, G. Chalkiadakis, and J. M. C. Garcia, "Learning policies for resolving demand-capacity imbalances during pre-tactical air traffic management," in *German Conference on Multiagent System Technologies*. Springer, 2017.

[14] S. Debernard, F. Vanderhaegen, and P. Millot, "An experimental investigation of dynamic allocation of tasks between air traffic controller and ai systems," in *Analysis, Design and Evaluation of Man–Machine Systems 1992*. Elsevier, 1993, pp. 95–100.

[15] J.-C. Pomerol, "Artificial intelligence and human decision making," *European Journal of Operational Research*, vol. 99, no. 1, 1997.

[16] P. L. de Frutos, R. R. Rodríguez, D. Z. Zhang, S. Zheng, J. J. Cañas, and E. Muñoz-de Escalona, "An air traffic controller's mental workload model for calculating and predicting demand and capacity balancing," in *International Symposium on Human Mental Workload: Models and Applications*. Springer, 2019, pp. 85–104.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.