Identifying Atmospheric Nucleation Events Using Machine Learning

Álvaro Silva-Silva, Javier Andrade-Garda, Daniel Garabato, Sonia Suárez-Garaboa, Elisabeth Alonso-Blanco, and Francisco J. Gómez-Moreno

Universidade da Coruña, Fac. Informática, Elviña 15071 A Coruña, Spain.
Universidade da Coruña, CITIC, Dep. Ciencias da Computación e Tecnoloxías da Información, Fac. Informática, Elviña 15071 A Coruña, Spain.
Universidade da Coruña, CITIC, Grupo ISLA, Dep. Ciencias da Computación e Tecnoloxías da Información, Fac. Informática, Elviña 15071 A Coruña, Spain.
Universidade da Coruña, Grupo ISLA, Dep. Ciencias da Computación e Tecnoloxías da Información, Fac. Informática, Elviña 15071 A Coruña, Spain.
Departamento de Medio Ambiente, CIEMAT, Avenida Complutense 40, 28040 Madrid, Spain
Correspondence: alvaro.silva@udc.es

DOI: *https://doi.org/10.17979/spu.23.c13*

Abstract: Atmospheric nucleation (*New Particle Formation*) is not only a key process in aerosol dynamics, but it also assists in regulating the planet's radiative balance. Accurate detection is essential to understand its implications for both climate and public health. However, manually identifying these events from particle size distributions is a slow and tedious process.
This work studies the feasibility of a proposal based on machine learning and computer vision to classify nucleation events from images (*surface plots*) of particle distribution time series. To this end, different preprocessing configurations are explored and both classical models and deep neural networks are tested. Preliminary results show promising performance, highlighting the system's ability to identify positive events with high sensitivity, suggesting a possible future integration into atmospheric monitoring platforms.

1 Introduction

The atmosphere is a dynamic system in which multiple physical and chemical processes take place to regulate the planet's equilibrium. Among these, atmospheric nucleation, or *New Particle Formation* , stands out as a fundamental mechanism in the formation of aerosols (Kulmala et al. (2004)). These nucleations generate ultrafine particles that, as they grow through condensation and coagulation, can participate in cloud formation and modify the Earth's radiative balance(Twomey (1977)). In addition to their climatic importance, the presence of ultrafine particles affects air quality and public health, as their inhalation is associated with cardiovascular and respiratory risks(Downward et al. (2018)). Despite their relevance, the detection and characterization of nucleations remains a challenge.

In this context, machine learning and computer vision techniques represent an innovative opportunity. Images contain a wealth of information that can be exploited to identify patterns associated with atmospheric processes.

Their analysis using *deep learning* algorithms allows complex phenomena to be classified based on visual representations, in many cases overcoming the limitations of traditional

approaches (Dal Maso et al. (2005)).

This study focuses on this line of research, identification of nucleation events based on automated image analysis. To this end, neural networks and optimization strategies are used to extract relevant information from visual data, with the aim of detecting the days on which this phenomenon occurs.

2 Objectives

The overall objective of this work is to design and evaluate a model for classifying atmospheric nucleation events based on the analysis of images derived from particle size distributions.

Specifically, the following objectives are proposed:

1. Preprocess particle distribution data to generate graphical representations suitable for automated analysis.

2. Implement and compare different machine learning models, including classical methods and deep neural networks.

3. Evaluate the performance of the models in terms of accuracy, precision, recall, and F1-score, prioritizing sensitivity for the detection of positive events.

3 Data

The data used in this study comes from a Scanning Mobility Particle Sizer (SMPS), an instrument that can determine both the particle sizes and the concentration of each fraction in the air at a given moment. These records were provided by CIEMAT (Center for Energy, Environmental and Technological Research), a Spanish public research organization that develops projects in the fields of energy, the environment, and technology. These data were obtained in the ACTRIS station located at the CIEMAT site, Madrid.

These files account for different particle sizes throughout the day, providing such measurements over a time-series with a 5-minute frequency that allows for detailed monitoring of the evolution of particle size distributions.

In general terms, the available data corresponding to campaigns carried out during summers, when nucleation episodes are more frequent. In addition, CIEMAT provided surface plots for all days where positive events were manually detected and classified (e.g., Figure 1), which clearly show the appearance of nucleation and the progressive growth of particles.

This paper addresses a binary classification of atmospheric nucleation event images into two categories:

- **Negative events**: days on which no evidence of nucleation is observed. Particle size distributions remain stable over time and no new ultrafine populations appear.

- **Positive events:** days on which a nucleation event occurs, supported by the appearance and growth of ultrafine particles.

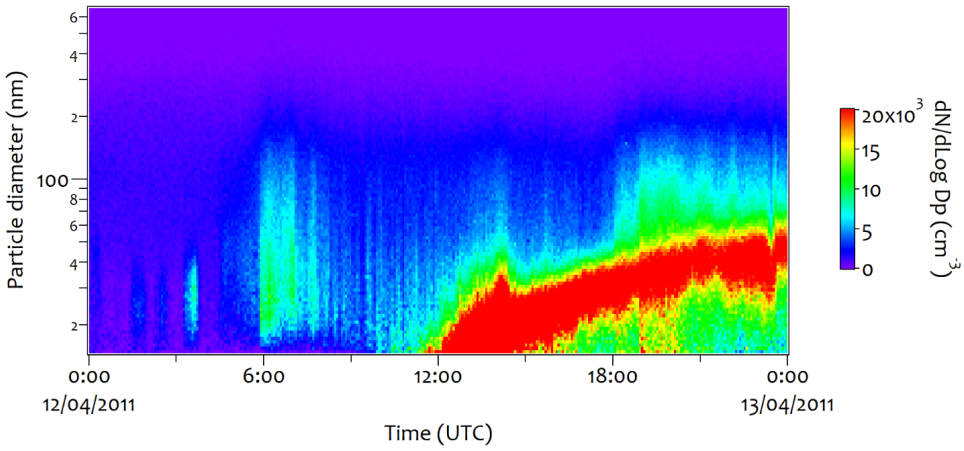

Figure 1: Example surface plots provided by CIEMAT.

In this type of graphical representation:

- The X-axis represents time, corresponding to the moments when the data was recorded.
- The Y-axis indicates the particle size, so that each row of the *surface plots* corresponds to a specific size range.
- The color of each cell reflects the particle concentration in that size range and at that point in time, with warm colors indicating higher concentrations and cool colors indicating lower concentrations.

In this way, *surface plots* allow the evolution of particle populations over time to be visualized and facilitate the identification of relevant patterns.

3.1 Preprocessing

In previous campaigns carried out by CIEMAT, measurements were recorded following a different methodology, so that the particle size ranges considered were different from those that are currently used. In fact, the ACTRIS network (Wiedensohler et al. (2012)) established a standard in order to ensure consistency in data collection and analysis. In order to make historical records compatible with the modern format, it was necessary to create and apply a specific conversion algorithm.

In addition, from time to time, instruments may cease to operate either suddenly and unexpectedly or due to scheduled maintenance shutdowns. Hence, such an issue prevents sufficient data from being obtained for those days and it is not possible to conduct any analysis on the occurrence of nucleation events. For this reason, days that do not reach a minimum threshold of valid information are discarded from the analysis, thus ensuring the quality of the samples and the reliability of the results.

Although we were initially provided with all the measurement particle sizes, as well as the entire time span, we have selected just the relevant ones to feed the AI-based models. To this end, a feature importance analysis was conducted, as well as some feedback was also provided by the experts at CIEMAT. Thus, we ended up with the time interval from 9–21 h and particle sizes from 10 to 110 nm.

Therefore, new surface plots were generated for both positive cases (e.g., Figure 2) and negative cases (e.g., Figure 3).

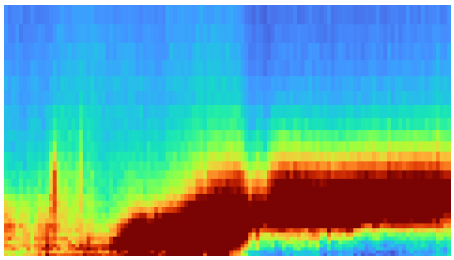

Figure 2: Preprocessed surface plots for a positive nucleation.

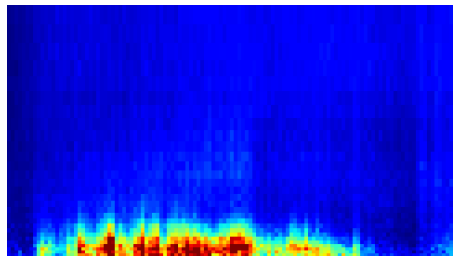

Figure 3: Preprocessed surface plots for a negative nucleation.

4 AI-based models

Given that detecting nucleation events in surface plots is a complex problem, different AI models were evaluated in order to compare their performance and determine the most appropriate approach.

Feature Extraction and Random Forest

In the first approach, simple statistical features were extracted from the raw image data (pixels), including: mean, standard deviation, maximum and minimum values. Then, a Random Forest (Breiman (2001)) classifier was fed with such an information.

Convolutional Neural Network (CNN)

The second approach addresses the use of CNN (Zhao et al. (2024)) from scratch, . Such a network consists of several layers that extract features and reduce dimensionality, followed by dense layers. To improve generalization ability, a regularization step was applied by using a dropout layer, "whereas the output layer follows a dense layer with sigmoid activation, so that independent probabilities are produced for each class."

Transfer Learning with MobileNet

Finally, a Transfer Learning approach was explored using the MobileNet (Wang et al. (2024)) architecture, using a pre-trained version on the ImageNet dataset. To adjust the model to the specific task, the last layers were unfrozen and a fine-tuning process was applied.

The upper layers of the network consisted of a global averaging of the extracted features, followed by a dense layer, then a dropout one for regularization purposes, and an output layer with sigmoid activation.

4.1 Methodology

The experimental procedure began with the definition of the training set, considering two main approaches:

- **Balanced set:** the same number of positive and negative samples was considered, ensuring an equal representation of both classes.

- **Unbalanced set:** it considers all available negative examples, thus increasing the weight of the predominant class in model training.

Given that CIEMAT researchers emphasized the importance of avoiding false negativo, we opted for a balanced training set, so that the training was not biased because of the available number of samples.

To assess the performance of these methods, we used some well-known standard metrics: precision, accuracy, F_1-score and recall.

In addition, to ensure that these techniques were able to appropriately generalize and to prevent overfitting, some mechanisms were put in place: a cross-validation procedure was employed, evaluating the aforementioned metrics on the validation set; and, the loss value was continuously monitored, stopping the training process when it did not improve for several consecutive iterations.

5 Results

Table 1 shows the overall performance achieved by each model:

On the one hand, models based on **statistical feature extraction and Random Forest** proved to be robust, showing good discrimination between positive and negative events. On the other hand, they may not capture complex spatial patterns as effectively as deep learning approaches.

On the one hand, **convolutional neural networks (CNN) trained from scratch** achieved the best overall performance, demonstrating the ability of deep learning to capture complex spatial patterns present in surface plots. On the other hand, they require more computational resources and training data.

Finally, the **Transfer Learning with MobileNet** approach allowed us to leverage prior knowledge from networks trained on large sets of natural images, obtaining competitive results. Although its accuracy was slightly lower than that of the CNN trained from scratch, this strategy proves useful when training resources are limited or when computing time needs to be reduced.

Table 1: Summarized (average) metrics for each model on the validation set.

Metrics	CNN	Random Forest	MobileNet
Accuracy	0.876	0.718	0.750
F1-score	0.831	0.756	0.753
Precision	0.865	0.829	0.744
Recall	0.801	0.694	0.763

6 Conclusions

The comparative study of different approaches to classifying atmospheric nucleation events using surface plots has led us to several relevant conclusions. First, the overall comparison indicates that the most appropriate model depends on the priorities of the study. When interpretability and simplicity are the main goals, Random Forest is a suitable choice. In contrast, if maximum accuracy and the ability to capture complex patterns are required, convolutional neural networks (CNN) trained from scratch are the most recommended option. Finally, when efficiency and the use of pre-trained models are considered important, Transfer Learning appears as a competitive strategy, and its performance could be further improved with larger datasets.

In addition, the results obtained validate the potential of automated surface plot analysis as a tool for detecting atmospheric nucleation events. This methodology offers a promising complement to traditional observation techniques, both for monitoring purposes and for scientific research.

7 Future Work

Looking ahead, several lines of research can be proposed. In the coming years, a larger volume of data will become available, including records from previous years that have not yet been processed or shared by CIEMAT CIEMAT (2025). This will enable more comprehensive and representative analyses.

Another direction will be to explore the direct use of time series data, avoiding the need for image-based techniques. This approach may provide new insights and complement the results obtained from surface plots.

Finally, the scope of the study can be broadened beyond the binary classification between positive and negative events, incorporating different subtypes of nucleation phenomena. These include days without evidence of events (no event), situations of progressive particle size reduction that persist long enough to be characterized (contraction), and Class I and II events, which differ in the degree of definition of their formation and growth rates. Expanding the classification in this way would allow for a more detailed and nuanced understanding of atmospheric nucleation processes.

Acknowledgments

This work was funded by the HYNU-CLIM project
(PID2024-161276OA-I00; by MCIN/AEI/10.13039/501100011033 and by 'ERDF A way of making Europe').

CITIC: CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01)

Bibliography

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL *https://doi.org/10.1023/A:1010933404324*.

CIEMAT. Centro de investigaciones energéticas, medioambientales y tecnológicas, 2025. Disponible en: *https://www.ciemat.es/*, accedido en septiembre de 2025.

M. Dal Maso, M. Kulmala, I. Riipinen, R. Wagner, T. Hussein, P. P. Aalto, and K. E. J. Lehtinen. Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from smear ii, hyytiälä, finland. *Boreal Environment Research*, 10:323–336, 2005.

G. S. Downward, E. J. van Nunen, J. Kerckhoffs, P. Vineis, B. Brunekreef, J. M. Boer, K. P. Messier, A. Roy, W. M. M. Verschuren, and Y. T. van der Schouw. Long-term exposure to ultrafine particles and incidence of cardiovascular and cerebrovascular disease in a prospective study of a dutch cohort. *Environmental Health Perspectives*, 126(12):127007, 2018.

M. Kulmala, H. Vehkamäki, T. Petäjä, M. Dal Maso, A. Lauri, V.-M. Kerminen, W. Birmili, and P. H. McMurry. Formation and growth rates of ultrafine atmospheric particles: a review of observations. *Journal of Aerosol Science*, 35(2):143–176, 2004.

S. Twomey. Influence of pollution on shortwave albedo of clouds. *Journal of the Atmospheric Sciences*, 34(7):1149–1152, 1977.

B. Wang, L. Yu, and B. Zhang. Al-mobilenet: a novel model for 2d gesture recognition in intelligent cockpit based on multi-modal data. *Artificial Intelligence Review*, 57:282, 2024. URL *https://doi.org/10.1007/s10462-024-10930-z*.

A. Wiedensohler, W. Birmili, A. Nowak, A. Sonntag, K. Weinhold, M. Merkel, B. Wehner, T. Tuch, S. Pfeifer, M. Fiebig, A. M. Fjäraa, E. Asmi, K. Sellegri, R. Depuy, and Venzac. Particle mobility size spectrometers: harmonization of technical standards and data structure to facilitate high quality long term observations of atmospheric particle number size distributions. *Atmospheric Measurement Techniques*, 5(3):657–685, 2012.

X. Zhao, L. Wang, Y. Zhang, et al. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57:99, 2024. URL *https://doi.org/10.1007/s10462-024-10721-6*.