On particle dispersion statistics using unsupervised learning and Gaussian mixture models

Nicholas Christakis (Νικόλαος Χρηστάκης)[1,2,a] and Dimitris Drikakis (Δημήτρης Δρικάκης)[1,b]
[1)]*Institute for Advanced Modeling and Simulation, University of Nicosia, Nicosia, CY-2417, Cyprus*
[2)]*Laboratory of Applied Mathematics, University of Crete, Heraklion GR-70013, Greece*

Understanding the dispersion of particles in enclosed spaces is crucial for controlling the spread of infectious diseases. This study introduces an innovative approach that combines an unsupervised learning algorithm with a Gaussian mixture model to analyze the behavior of saliva droplets emitted from a coughing individual. The algorithm effectively clusters data, while the Gaussian mixture model captures the distribution of these clusters, revealing underlying subpopulations and variations in particle dispersion. Using computational fluid dynamics simulation data, this integrated method offers a robust, data-driven perspective on particle dynamics, unveiling intricate patterns and probabilistic distributions previously unattainable. The combined approach significantly enhances the accuracy and interpretability of predictions, providing valuable insights for public health strategies to prevent virus transmission in indoor environments. The practical implications of this study are profound, as it demonstrates the potential of advanced unsupervised learning techniques in addressing complex biomedical and engineering challenges and underscores the importance of coupling sophisticated algorithms with statistical models for comprehensive data analysis. The potential impact of these findings on public health strategies is significant, highlighting the relevance of this research to real-world applications.

I. INTRODUCTION

Detecting early signs of virus carriers and examining virus transmission dynamics through the air during coughing and sneezing in indoor environments is essential for reducing the impact on public health. In pathogens research, limiting factors included the irreproducibility of in-vitro studies in in-vivo contexts due to the complexity of physiological processes and the behavior of genetically modified tissues.[1] Historically, mathematical models have been pivotal in addressing these issues, but they involve computationally expensive simulations.[2,3] Performing multi-phase computational fluid dynamics (CFD) simulations for virus transmission through the air has provided significant insights,[4,5]. Still, in a three-dimensional space, it can be computationally expensive. Furthermore, installing sensors indoors to monitor particles and environmental conditions on a large scale could be technically challenging. Developing methods to reduce the computational and experimental burden could provide a promising alternative.

Studying particle dispersion in enclosed spaces is essential for several reasons beyond the context of COVID-19. One significant aspect is indoor air quality. This involves understanding the dynamics of how particles disperse, which can help assess the health risks associated with indoor air pollutants, such as dust, allergens, and volatile organic compounds. Additionally, insights into particle movement can inform the design of ventilation systems to improve air quality and reduce exposure to harmful particles.

Regarding occupational safety, studying particle dispersion is crucial for managing exposure to workplace hazards, such as chemical fumes or fine particulates in industrial settings. It also ensures regulatory compliance by keeping particle levels within safe limits, often a legal requirement. From an environmental control perspective, controlling particle dispersion is essential in laboratories or clean rooms to prevent contamination of sensitive processes or products. Moreover, the potential for optimizing airflow and filtration systems based on particle dispersion studies to lead to more energy-efficient environmental control systems is a reason for optimism about the future.

In public health, understanding how particles, including pathogens, disperse is a critical piece of knowledge in managing the spread of airborne diseases beyond COVID-19. This knowledge is also invaluable for emergency preparedness, aiding in the development of effective response strategies in the event of a biological or chemical threat. The urgency of our research is underscored by the importance of understanding how particles disperse in managing the spread of airborne diseases.

Finally, in building design and architecture, properly managing particle dispersion is not just a technical consideration but a key factor in ensuring occupants' overall comfort and well-being. It also supports sustainable practices by reducing the need for excessive air filtration and conditioning. These considerations highlight the broad relevance of particle dispersion studies, which draw on principles from physics, engineering, and public health in promoting health, safety, and efficiency in enclosed spaces.

Machine Learning (ML) can reveal hidden patterns within noisy or incomplete data.[6–9] Unsupervised Learning (UL), in particular, has achieved significant success across various scientific domains, especially for complex problems involving data set grouping. These problems range from deciphering complex genetic structures[10] and understanding natural language semantics[11] to discovering astronomical phenomena[12], among other applications. UL is advantageous because it can handle unlabelled data, including experimental, compu-

[a)]Electronic mail: nchristakis@tem.uoc.gr
[b)]Electronic mail: Author to whom correspondence should be addressed: drikakis.d@unic.ac.cy.

tational, or field measurements, without user intervention. This capability is particularly relevant given the persistent gaps in our understanding of the physics underlying various processes, such as virus transmission. Several UL algorithms have been developed.[13–15] However, UL in engineering, physics, and biomedical applications encompasses higher uncertainty due to the elusive nature of underlying physical processes.[15] Unlike general data sets in economics, marketing, or media, pattern classification and prediction in fluid mechanics and engineering must be physics-relevant and explainable. Simplicity in ML algorithms is also essential to facilitate explainability and broader application with confidence.

Recent advancements in UL, such as the Reduce UNcertainty and Increase CONfidence (RUN-ICON) algorithm[14], have further expanded its applicability. This state-of-the-art algorithm has demonstrated effectiveness in reliably clustering artificially generated and real-world data. Applying the RUN-ICON algorithm to physics-based data produced by CFD can provide invaluable insights into the behavior of saliva droplets and aerosols emitted from coughing or sneezing. This analysis offers a data-driven perspective on virus transmission dynamics that was previously unattainable. It has only been very recently that combining UL and CFD in particle-driven dynamics has been attempted.[16] This combination offers a promising avenue for advancing our understanding in an era where precision and rapid response to emerging infectious diseases are paramount. This study is motivated by the potential of gaining better insight into particle dispersion through the air by incorporating a stochastic Gaussian Mixture Model (GMM) into the UL methodology. GMMs are used to model data distribution by representing it as a mixture of several Gaussian distributions, allowing for the identification of underlying sub-populations within the data.[17]

While UL effectively uncovers hidden patterns in unlabeled data, it often struggles to accurately model the complex, multimodal distributions commonly found in large datasets. On the other hand, GMMs, known for their ability to represent such distributions, require a robust method for initial pattern identification and clustering. By combining UL with GMMs, we can harness the exploratory power of UL to detect patterns without prior labels and then refine these patterns into a probabilistic framework using GMMs. This integrated approach is particularly crucial for analyzing complex data, such as particle trajectories, where traditional clustering methods, with their inherent limitations, fail to capture the full dynamics. Our study demonstrates that the synergy between UL and GMMs enables a more comprehensive and insightful analysis, offering a depth of understanding that can not be attained with either method alone.

By utilizing GMMs, we aim to enhance clustering and analysis capabilities, providing a deeper understanding of virus transmission dynamics based on the probabilistic nature of such events. This means that instead of assigning a data point to a single cluster, GMMs provide a more subtle view by indicating the likelihood that a point belongs to different clusters; in this way, when data of trajectories of varying size particles need to be analyzed, we may discover underlying motion patterns, segment particle trajectories into meaningful clusters, and analyze complex, overlapping trajectories in a probabilistic and flexible manner. Therefore, the study by Christakis et al.[16] was expanded to include, with the aid of GMMs, an analysis of phenomena in all three spatial directions individually, following the initial identification of dominant clusters by the RUN-ICON algorithm.

The paper is organized as follows. Section II discusses COVID-19 spreading and infection in enclosed spaces and refers to how the data were collected using CFD models. In Section III, the RUN-ICON algorithm is briefly described. Section III introduces the stochastic GMMs and presents the coupling methodology between RUN-ICON and GMM. The results are presented in Section V. The conclusions drawn from this study are summarized in Section VI.

II. MOTIVATION

The recent COVID-19 pandemic, with more than 770×10^6 confirmed cases and nearly 7×10^6 deaths[18], has underscored the critical need for rigorous and urgent studies on virus transmission. Understanding how droplets and aerosols expelled by infected individuals spread and infect others has become paramount in devising effective mitigation strategies.

This urgency has spurred extensive research into virus transmission in enclosed spaces.[4,5,19–24] These studies have explored air filtration, air purification, and the efficacy of high-efficiency particulate air (HEPA) filters designed to capture more than 99% of particles. They also provide high ventilation rates with a mix of 50% fresh air and 50% filtered air, achieving more than 20 air changes per hour (ACH) in a space.[25] ACH refers to the number of times the air in a room can be completely renewed in one hour based on the ventilation system's volumetric flow rate and the room's volume. However, this is an indicative value, as factors like flow recirculation, room geometry, and the placement of ventilation inlets and outlets can significantly affect air renewal time. These studies have demonstrated the inevitability of virus transmission when individuals are close.

Additionally, recent research has focused on aerosol transmission and ventilation configurations in car cabins[26,27] and buses[28,29]. Studies on virus transmission on cruise ships have mainly involved mechanistic modeling[30] and outbreak data analysis.[31] A comprehensive review of COVID-19 transmission on cruise ships[32] highlighted that high occupancy cabins pose an increased transmission risk. However, it did not extensively discuss the impact of cabin ventilation systems on transmission dynamics.

CFD studies with ventilation recommendations have so far been limited to smaller vessels.[33] The literature presents conflicting views on the recirculation of potentially contaminated air and the efficiency of ventilation rates on cruise ships. While Azimi et al.[30] recommended high ventilation rates in cruise ship cabins, opinions vary among researchers.[34–37]

Current standards and guidelines for room safety regarding airborne virus transmission emphasize high air exchange rates.[38–40] However, maintaining these high rates can be energy-intensive and may reduce comfort due to strong air

drafts. The American Society of Heating, Refrigerating and Air-conditioning Engineers (ASHRAE) Standard[40,41] provides a formula for calculating the minimum ventilation rate based on the number of occupants and the surface area of a room: $\dot{Q}outdoor-air = R_p \times N_p + R_a \times A$. Here, $\dot{Q}outdoor-air$ is the required outdoor air supply, R_p is the required outdoor flow rate per person, and R_a is the required outdoor flow rate per unit area, with $R_p = 2.5; \ell/(s\cdot$ person) and $R_a = 0.3; \ell/(s\cdot m^2)$. For a cabin with two occupants, this translates to 27 m^3/h. Designers of cruise ships have proposed stricter ventilation limits of 30 m^3/h per room occupant. Recommendations from the World Health Organization (WHO)[42], the European Federation of HVAC Associations[43], and other research studies[44] suggest 36 m^3/h per person. The Federal Public Service (FPS) Health, Food Chain Safety and Environment of Belgium has set a slightly higher standard of 40 m^3/h per person.[45]

Considering these guidelines, the reference flow rate value for a small cruise cabin with two occupants is 60 m^3/h. The latest Centers for Disease Control and Prevention (CDC) guidelines, based on the draft of ASHRAE Standard 241-2023[40], and other studies[46], propose a minimum of 5 ACH, translating to a flow rate of 200 m^3/h for the cabin under investigation. ASHRAE Standard 241-2023[40] recommends 15 l/s per occupant, equating to 108 m^3/h for the studied cabin. Based on CDC data, a typical home has less than 0.5 ACH, while ASHRAE Standard (62.2-2019)[47] recommends 0.35 ACH.

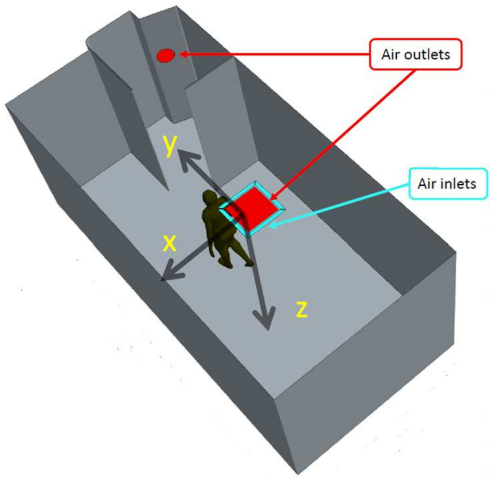


FIG. 1. The simulation domain used in the CFD simulations with the three directions (lateral x, streamwise y, and vertical z) depicted. The width of the room ($x-$direction) is 2.8 m, its length ($y-$direction) is 5.92 m and its height $z-$direction) is 2.4 m. The location of the coughing person is also shown, as well as the ventilation inlets and outlets.

A typical cruise ship cabin was considered, with details available in Ritos et al.[48] Figure 1 presents the simulation domain, where the three $x-$, $y-$ and $z-$ axes are depicted. The cabin height is 2.4 m, with the overall deck height of the ship being 2.8 m. A standard air conditioning unit is placed at the cabin's center with a square outlet (□48 cm) and four rectangular inlets (55×5 cm$\times$cm each) that expel air at an angle of 45^o. The coughing person is placed under the air conditioning unit. The bathroom area has an additional circular outlet (⌀25 cm). The CFD simulations used polyhedral non-uniform cells ($\approx 0.6 \times 10^6$), with significant refinement in all inlet and outlet regions. For example, at the mouth and up to a distance of 0.5 m in the streamwise direction, the cells have a maximum isotropic size of 4 mm compared to the overall targeted cell size of 4 cm. Enhanced quality triangles were used for the surface meshing method and 5 core mesh optimization cycles with a quality threshold of 0.6 for the entire mesh. The most significant volume change was 0.01 in less than 2% of the cells, while the mesh had 100% face validity. The maximum skewness angle was 86.5^o. The choice of this mesh has been taken after conducting a mesh convergence study on main local and global flow parameters, like the cabin's average temperature (T_{avg}), relative humidity (RH), momentum and mass conservation of the fluid.

No-slip boundary conditions have been applied to all walls, the ceiling, the floor, and the person's body. Outlet boundary conditions with a specified mass flow rate have been used in the air outlets, with $1/3$ of the flow directed to the bathroom outlet and the rest to the A/C outlet, leading to a pressure balance in the cabin. The air inlets provide the overall targeted mass flow rate with the air blowing at an angle of 45^o. Further details can be found in.[48]

Since the air outlets are along the $y-$axis, the direction along that axis is considered as **streamwise** and the one along the $x-$axis as **lateral**. The $z-$axis defines the **vertical** direction, along which gravity acts. Data from this configuration are used in the present work. The effect of droplet evaporation, highlighted in review papers[49], has also been included. Dhand and Li[50] emphasized the role of droplets in virus spread and the significance of evaporation and size alteration. Recently, it has been shown that droplets can spread up to five times more when high ventilation rates are used after a person coughs.[48] CFD simulations of coughing events at varying flow rates were performed, with particles of different sizes tracked in their trajectories. The initial conditions and parameters for the simulations were set to replicate realistic coughing scenarios. Data from this study for a typical cabin ventilation rate of 120 m^3/h were utilized for the presented work. Further details on the data are provided in[16]. colour black Note the CFD simulations always encompass assumptions regarding initial conditions, but in the present study, we aimed to minimize the uncertainty by introducing realistic data about the room's ventilation and physical modelling employed to perform the simulations.

III. DOMINANT CLUSTERS IDENTIFICATION WITH RUN-ICON

In this study, the RUN-ICON algorithm is initially utilized to determine the optimal number of clusters at different times of droplet dispersion.[14] RUN-ICON's main advantage over other UL techniques is its systematic approach to determining the optimal number of clusters, reducing reliance on subjective judgments. Instead of relying on the Sum of Squared Errors, RUN-ICON introduces novel metrics such as the Clustering Dominance Index (CDI) and Uncertainty to identify commonly dominant centers across multiple repetitions of the K-means++ algorithm. CDI represents the frequency of occurrence of specific clustering configurations, while Uncertainty measures the relative difference between upper and lower CDI bounds. By focusing on stability, RUN-ICON ensures that the chosen cluster centers accurately represent underlying patterns, mitigating the influence of outliers or noisy data points. This leads to more robust and interpretable clustering results, with the algorithm demonstrating an accuracy of 97% in identifying intended clustering configurations, significantly outperforming other algorithms[14,15]. This makes RUN-ICON particularly suitable for complex scenarios like particle-like dispersion and various realistic flow applications. It is also important to note that RUN-ICON makes no numerical assumptions. The steps of the algorithm are as follows:

1. Specify a particular number of clusters and perform the K-means++ clustering algorithm 100 times.

2. Compute the coordinates of the cluster centres after every run.

3. Compare the cluster centre coordinates from all 100 runs to identify the coordinates that appear most frequently.

4. Select the cluster centers that exhibit the highest frequency of appearance as the dominant centers corresponding to the chosen number of clusters and calculate CDI.

5. Repeat the preceding steps (steps 1-4) nine more times to obtain ten sets of dominant cluster centres.

6. Compute the average frequency of appearance for the dominant centers and their respective CDIs, derived from the ten repetitions, ensuring that they correspond to the same clustering centers. Calculate these metrics' upper and lower bounds and assess the variance within this range.

7. Repeat steps 1-6 for various cluster numbers, beginning with 3 clusters and extending up to 10 clusters.

8. Choose the cluster number with the highest average CDI and low variance (less than 30%) as the optimal number of clusters for the RUN-ICON algorithm.

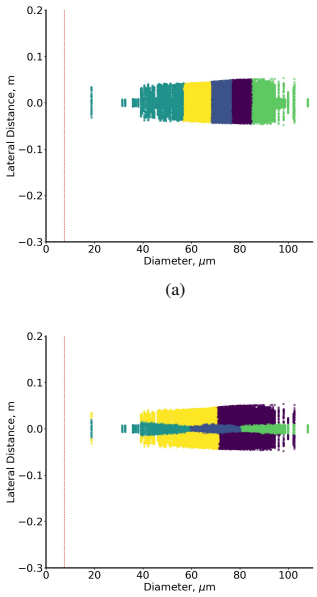


FIG. 2. Clustering of particles at time 0.12 s from the start of the coughing event in the lateral direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

IV. STOCHASTIC MODELING - MATHEMATICAL THEORY

GMM is a versatile probabilistic tool for clustering that can enhance the initial results obtained from any clustering algorithm.[51–53] Incorporating GMMs into UL algorithms significantly improves prediction accuracy and provides deeper insights into complex data sets. GMMs are particularly valuable in unsupervised learning due to several key benefits:

1. **Modeling data distributions**: GMMs model data as a mixture of several Gaussian distributions, ideal for capturing the inherent sub-populations within a data set.[54] This capability leads to a better understanding of the data structure, which is often necessary for accurate predictions in complex domains, such as virus transmission dynamics.

2. **Flexibility and adaptability**: Unlike more straightforward clustering methods, GMMs can handle clusters of different shapes and sizes.[55] This flexibility is crucial

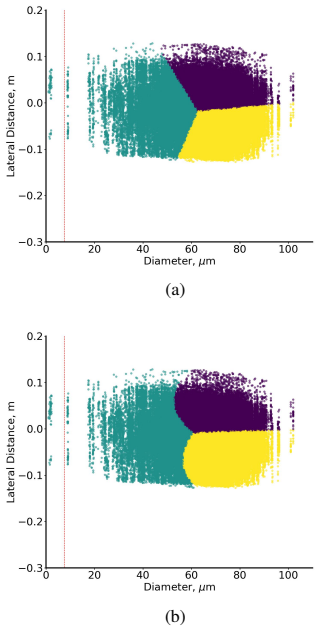


FIG. 3. Clustering of particles at time 2 s from the start of the coughing event in the lateral direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

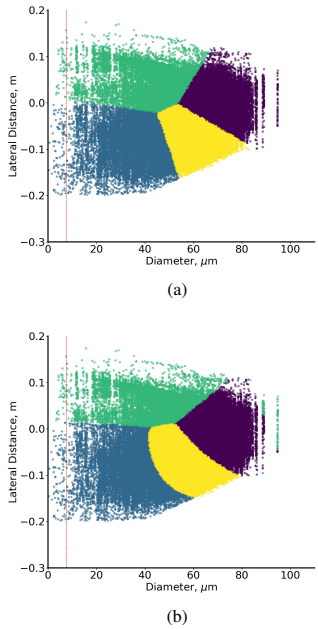


FIG. 4. Clustering of particles at time 4 s from the start of the coughing event in the lateral direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

when dealing with real-world data, which often exhibits varied distributions. By adapting to the underlying data distribution more effectively, GMMs may improve the robustness of UL algorithms.

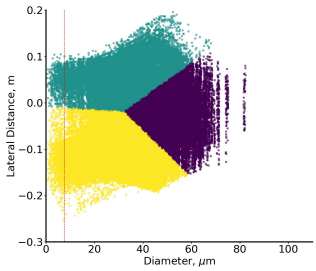
3. **Probabilistic framework**: GMMs provide a probabilistic approach to clustering, where each data point is assigned a probability of belonging to each cluster.[56] This probabilistic framework is beneficial for handling uncertainties and ambiguities in data classification, making GMMs well-suited for applications where precision is critical, such as predicting virus transmission pathways.

4. **Enhanced interpretability**: By representing data as a combination of Gaussian distributions, GMMs offer a more interpretable framework for understanding the underlying patterns and structures within the data.[57] This interpretability is essential for validating and explaining the results of UL algorithms, particularly in scientific applications where transparency and explainability are paramount.

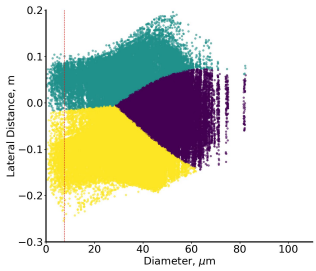
5. **Integration with other methods**: GMMs can be easily integrated with other ML techniques and domain-specific models.[58,59] This will allow for a more comprehensive data analysis by exploiting the strengths of all approaches involved, leading to more accurate and reliable predictions.

6. **Handling multimodal data**: Real-world data often exhibit multimodal distributions, with multiple subgroups or modes present within a single dataset. GMMs are particularly adept at handling such multimodal data, as they can identify and model each mode separately.[60] This ability is crucial for accurately capturing the complexity of virus transmission processes and other intricate phenomena.

All these benefits make GMMs a powerful tool for improving the predictions and insights derived from UL, particularly in complex and high-stakes applications such as virus

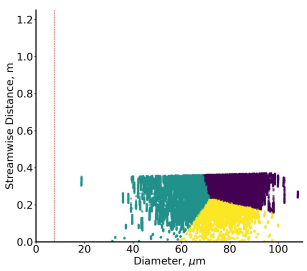


(a)

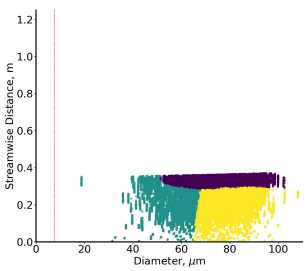


(b)

FIG. 5. Clustering of particles at time 8 s from the start of the coughing event in the lateral direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.



(a)



(b)

FIG. 6. Clustering of particles at time 0.12 s from the start of the coughing event in the streamwise direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

transmission studies. This section details how GMM is utilized to refine clusters based on the centers identified by a different clustering algorithm. It is well known, for example, that the K-means algorithm separates into clusters, mostly of spherical shape and almost equal size, which can be a limitation.[61] GMM addresses this by modeling the data as a combination of Gaussian distributions, each with its mean and covariance. In the present work, we show how to utilize GMM to improve clustering results initially generated by the RUN-ICON algorithm.[14] By using the dominant cluster centers from RUN-ICON as starting points for GMM, we achieve more precise and reliable clustering. All relevant equations are presented, and an implementation guide is provided. It has to be noted that the number of dominant clusters does not change every time, and it remains as predicted through RUN-ICON.

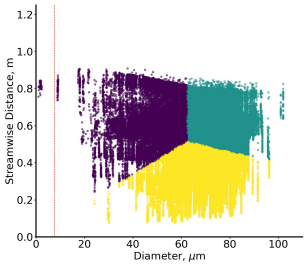
A. Setting up the framework

The GMM assumes that data points come from a mixture of multiple Gaussian distributions.[17] Suppose we have a dataset of N data points $\mathbf{x}$ that has already been separated into K different clusters. For each cluster, a mean μ_k (a $1 \times n$ matrix) is defined, each element of which is:

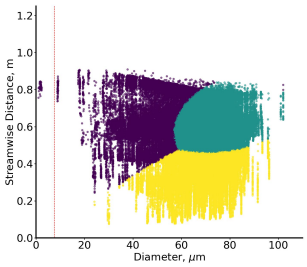
$$\mu_{kj} = \frac{\sum_{i=1}^{N_k} \mathbf{x}_{ij}^k}{N_k} \tag{1}$$

where:

- $k \in [1,...,K]$
- μ_{kj} is the j-th element of matrix μ_k ($j \in [1,...,n]$
- N_k is the number of points in cluster k.
- $\mathbf{x}_{ij}^k$ is the j-th coordinate of a point that belongs to cluster k

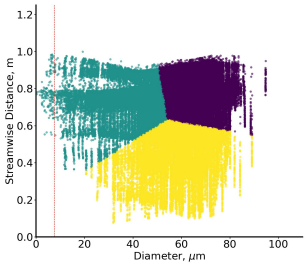


(a)

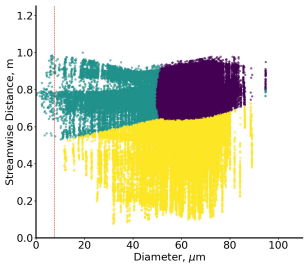


(b)

FIG. 7. Clustering of particles at time 2 s from the start of the coughing event in the streamwise direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.



(a)



(b)

FIG. 8. Clustering of particles at time 4 s from the start of the coughing event in the streamwise direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

We also define for every cluster k an $n \times n$ covariance matrix Σ_k as:

$$\Sigma_k = \frac{1}{N_k - 1}(\mathbf{X}_\mu^k)^T \mathbf{X}_\mu^k \tag{2}$$

where:

- $\mathbf{X}_\mu^k$ is the $N_k \times n$ matrix of all cluster-k points minus the corresponding to each row μ_k component.

Then, we may define the Gaussian distribution for points $\mathbf{x}$ belonging to cluster k as:

$$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T \Sigma_k^{-1}(\mathbf{x}-\mu_k)\right) \tag{3}$$

where

- $|\Sigma_k|$ and Σ_k^{-1} are the determinant and inverse of the covariance matrix, respectively.
- $(\mathbf{x}-\mu_k)$ is an $n \times 1$ vector, corresponding to the difference between point $\mathbf{x}$ and its respective component of μ_k

We may also define a weight π_k for each cluster, which indicates the proportion of points belonging to that cluster, i.e.:

$$\pi_k = N_k / N \tag{4}$$

As becomes obvious $\sum_{k=1}^{K} \pi_k = 1$. Then, based on GMM, the probability of a point $\mathbf{x}_i$ from the dataset belonging to cluster k is given by:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)} \tag{5}$$

For a complete derivation of this formula, the interested reader is referred to the work by Reynolds.[62]

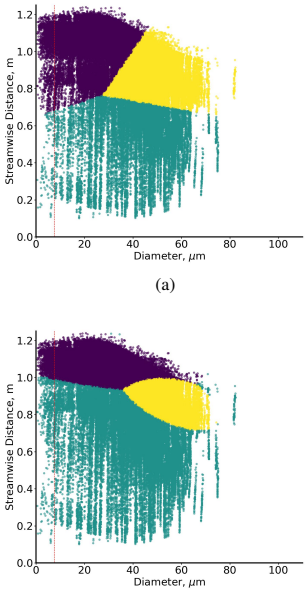


FIG. 9. Clustering of particles at time 8 s from the start of the coughing event in the streamwise direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

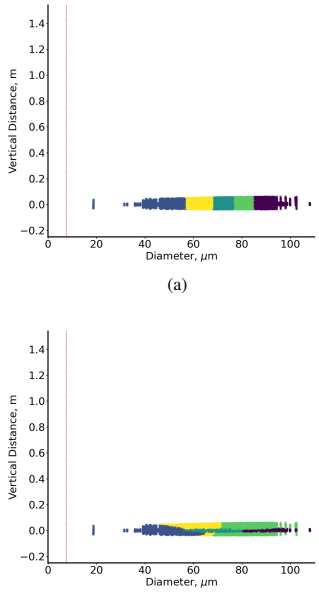


FIG. 10. Clustering of particles at time 0.12 s from the start of the coughing event in the vertical direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

B. Expectation-Maximization (EM) Algorithm

We may now introduce the EM algorithm, which is employed to find the maximum likelihood estimates of the Gaussian distributions for the clusters based on GMM. It iteratively performs two steps: the Expectation (E) step and the Maximization (M) step.

- **E-step** During this step, we calculate the probability γ_{ik} that data point $\mathbf{x}_i$ belongs to cluster k. For every point, K probabilities for belonging to any of the K clusters are calculated. A point is placed in a cluster based on the highest probability score. This cluster could now be different from the original cluster the point belonged to.
- **M-step** In this step, we redistribute points to clusters. We update the means, covariance matrices and weights for the new clusters of points, according to equations (1), (2) and (4), respectively. The number of clusters K remains the same, but the centers can change. The new centres are defined by the new calculated means.

C. Implementation Steps

In this section, we detail the implementation of our methodology for the integration of RUN-ICON with GMM in three distinct steps:

1. **Initialization with RUN-ICON**

   The RUN-ICON algorithm is utilized to find the dominant clustering, and the cluster centres, as determined by the algorithm, are set as the initial means of the GMM. The initial covariance matrices and weights are also calculated through equations (2) and (4), respectively.

2. **EM algorithm**

   Execute the EM algorithm to refine the GMM parameters:

   - E-step: Calculate the probabilities γ_{ik} for all points to belong to any of the k clusters.

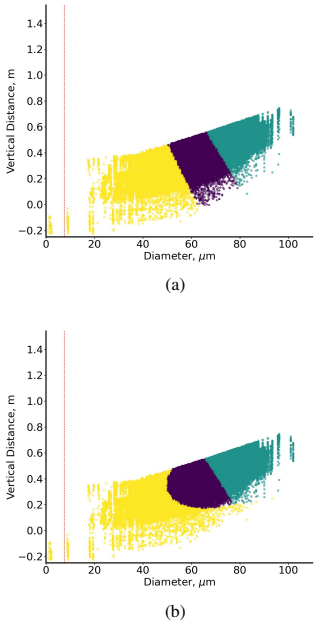


FIG. 11. Clustering of particles at time 2 s from the start of the coughing event in the vertical direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

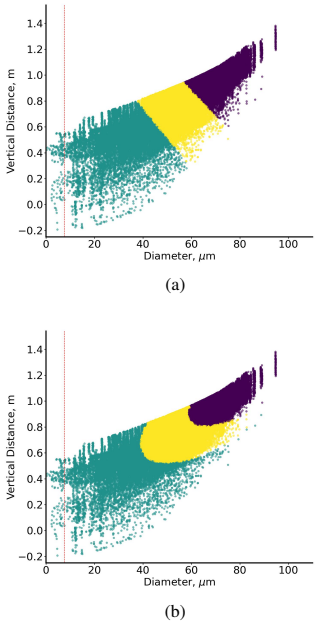


FIG. 12. Clustering of particles at time 4 s from the start of the coughing event in the vertical direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

- M-step: Redistribute points in K clusters based on highest probability scores and recalculate μ_k, Σ_k, and π_k.

3. **Check for convergence**

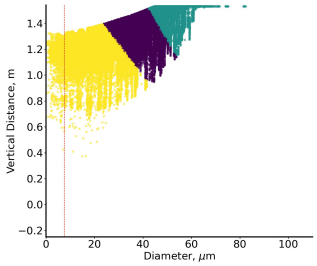
   To check for convergence, the previously and newly calculated means are compared. If they are below the tolerance set by the user, the algorithm terminates. Otherwise, we return to Step 2 and repeat.

Integrating RUN-ICON with GMM, we thus obtain a robust clustering method that leverages RUN-ICON's ability to form dominant clustering identification and GMM's flexibility for modeling complex data distributions. The EM algorithm iteratively refines the parameters, leading to more precise and dependable clustering outcomes. This dual approach helps understand droplets' nuanced behaviour over time and space.

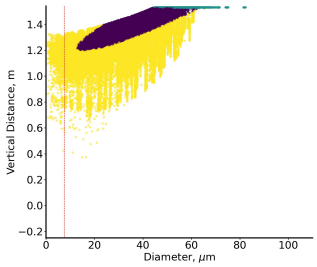
V. RESULTS AND ANALYSIS

In the work by Christakis et al.[16], where the RUN-ICON algorithm was applied to the study of droplet spreading in an enclosed space, no consideration was given to what was occurring in each of the three spatial dimensions (i.e., lateral, streamwise, and vertical) separately. Instead, the Euclidean norms of the three-dimensional position vectors from the mouth of a coughing person were considered. We utilize the RUN-ICON algorithm in the present work to predict the optimum cluster numbers. Then, we apply GMM to study what occurs in all three spatial dimensions separately and to examine whether the GMM model predicts different particle classifications in the clusters. A typical ventilation rate of 120 m^3/h was considered, with the person standing at the cabin's center underneath the air inlet, as shown in Figure 1. The streamwise direction is along the $y-$axis, the lateral direction is along the $x-$axis, and the vertical direction is along the $z-$axis.

Data from the CFD simulations mentioned in Section II for



(a)



(b)

FIG. 13. Clustering of particles at time 8 s from the start of the coughing event in the vertical direction. The $x-$axis is the particle diameter in μm, and the $y-$axis is the distance from the mouth in m. The red vertical line at 5 μm indicates the threshold between droplets and aerosols. (a) RUN-ICON predicts clustering, and (b) clusters are redistributed as predicted by GMM. Each colour indicates droplets belonging to the same cluster.

Time (s)	Number of particles	Maximum diameter (μm)	Minimum diameter (μm)
0.12	146,400	108	19
2	146,334	102	1
4	144,676	95	0.3
8	116,875	82	0.6

TABLE I. Times after the start of the coughing event, number of remaining particles and maximum-minimum diameters in the distribution.

this particular ventilation rate were utilized. Several particles were expelled and tracked in all three spatial dimensions four times (0.12 s, 2 s, 4 s, and 8 s). At the final time of 8 s, approximately 117,000 particles remained (since the effects of evaporation were considered), with diameters between 82 and 0.6 μm (see Table I). As can be seen, the evaporation rate accelerates with time; a considerable amount of particles evaporated in the last 4 s of the flow (27,801 particles), whereas in the first 4 s, only 1,724 particles evaporated. The diameters of particles, as expected, were reduced due to the effects of evaporation. The individual was considered to cough along the $y-$axis (streamwise), and the optimum numbers of clusters, as RUN-ICON predicted for each of the three directions, are given in Table II.

Time (s)	Optimum number of clusters		
	$x-$dir	$y-$dir	$z-$dir
0.12	5	3	5
2	3	3	3
4	4	3	3
8	3	3	3

TABLE II. Times after the start of the coughing event and optimum number of clusters in each of the three directions.

In the medical community, particles may be classified as droplets or aerosols based on their sizes (particles with diameters greater than 5 μm are classified as droplets. Otherwise, they are classified as aerosols).[63–65] This classification is crucial, as droplets are considered more significant for viral load transmission due to their larger sizes and, hence, the higher viral load that they carry. In the present study, since evaporation has been considered, we observed this distinction and analyzed how many particles fall off the droplet size category at all recorded times.

A. Lateral ($x-$) Direction

- Initial Time (0.12 s): RUN-ICON (Figure 2(a)) identified five clusters, symmetrically distributed on either side of the coughing individual. The clustering was predominantly size-based, with larger particles showing a higher initial momentum, resulting in similar travel distances.

  The application of GMM (Figure 2(b)) revealed a different clustering pattern, indicating that particles of the same size could travel varying distances. This suggests variability in particle velocities, possibly due to interactions among particles. The clusters formed by GMM included three thin clusters around the origin and two additional clusters, illustrating the velocity differences among similarly sized particles.

- 2 s: RUN-ICON (Figure 3(a)) predicted three clusters, with smaller particles forming a distinct cluster likely due to the evaporation of smaller droplets. Larger particles were separated based on their position relative to the coughing source, indicating similar travel distances due to higher momentum. GMM (Figure 3(b)) showed minimal changes, maintaining the clusters identified by RUN ICON.

- 4 s: RUN-ICON (Figure 4(a)) identified four clusters, with smaller particles gaining momentum and traveling slightly further than larger particles. GMM (Figure 4(b)) retained the same clustering pattern, confirming the initial analysis.

- 8 s: RUN-ICON (Figure 5(a)) predicted three clusters, with larger particles now traveling further than smaller ones. GMM (Figure 5(b)) confirmed these findings, showing similar particle distributions in the three clusters.

  In the lateral direction, particle travel distances did not exceed 0.2 meters, making this direction less significant than the other two dimensions.

B. Streamwise ($y-$) Direction

- Initial Time (0.12 s): RUN-ICON (Figure 6(a)) identified three clusters where both small and large particles traveled similar distances, with intermediate particles lagging. GMM (Figure 6(b)) maintained this clustering, indicating no significant alteration.
- 2 s: RUN-ICON (Figure 7(a)) predicted three clusters, with smaller particles gaining momentum while larger particles lagged slightly, likely due to interactions among larger particles. GMM (Figure 7(b)) showed similar cluster formations, supporting the initial results.
- 4 s: RUN-ICON (Figure 8(a)) identified three clusters with larger particles overtaking smaller ones. GMM (Figure 8(b)) confirmed these clusters with no significant changes.
- 8 s: RUN-ICON (Figure 9(a)) predicted three clusters, with smaller particles now overtaking larger ones, as larger particles settled closer to the floor and were less influenced by the airflow. GMM (Figure 9(b)) maintained the same clustering, showing consistent patterns.

  As observed, particles in the streamwise direction did not travel beyond 1.2 meters from the mouth, with the predicted distribution by GMM remaining consistent.

C. Vertical ($z-$) Direction

- Initial Time (0.12 s): RUN-ICON (Figure 10(a)) identified five clusters, with particles distributed above and below the line defining the streamline direction. This behavior mirrored the lateral direction due to initial particle dynamics. GMM (Figure 10(b)) exhibited similar clustering to the initial time in the lateral direction (Figure 2(b)), suggesting variability in particle velocities, possibly due to interactions among particles due to their close packing.
- 2 s: RUN-ICON (Figure 11(a)) predicted three clusters, with gravity significantly influencing larger particles, driving them towards the ground. GMM (Figure 11(b)) maintained this trend, showing similar clusters.
- 4 s: RUN-ICON (Figure 12(a)) identified three clusters, with an increased slope between large and small particles, indicating faster downward movement of larger particles. GMM (Figure 12(b)) confirmed this pattern, showing consistent clustering.
- 8 s: RUN-ICON (Figure 13(a)) predicted three clusters, with 97% of the particles having fallen at least 1 m below mouth level and larger particles having already settled on the floor. GMM (Figure 13(b)) slightly altered this clustering by gathering together in a cluster all particles that had already settled on the floor (approximately 27% of the distribution).

In the vertical direction, the separation seems to be primarily size-driven, with larger particles gaining momentum due to gravity and thus traveling faster than smaller ones.

D. Droplet-aerosol separation

The presented data shows that a significant amount of particles evaporates entirely within the time scale of 8 s (approximately 21%). However, this only occurs during the latter stages of the flow; within the first 4 s, less than 2% of particles from the initial distribution have evaporated.

As for the droplet-aerosol separation, it may be observed that at 2 s, much less than 1% of particles are below the 5 μm threshold (56 out of approximately 146,000). The number of particles below 5 μm remains relatively small even after 4 s (62 particles out of roughly 144,000). Only after 8 s does this number rise to approximately 0.6% of the particle size distribution, which is a tiny percentage of the distribution. Hence, we may conclude that by the end of the flow, the rate of smaller particles-aerosols in the particle size distribution is insignificant and has no critical effect on lowering virus transmission.

E. Comparisons with other works

In recent years, there have been limited studies where particle dispersion with airflow in indoor spaces has been modeled by utilizing CFD.[24,66,67] To the best of our knowledge, there are no studies in the literature that combine CFD with UL techniques, to extract patterns of particle dispersion in enclosed spaces, such as hospital rooms, ship cabins, etc. Hence, direct comparisons between the findings of the current study and past works are not feasible for several reasons:

- Traditional CFD studies focus on modeling airflow and particle trajectories under specific conditions, but they may not delve deeply into identifying underlying patterns in particle behavior. By incorporating clustering with the aid of UL and GMMs, the proposed method goes beyond merely tracking particle movement; it discerns intricate patterns in dispersion, clustering particles based on their properties and interactions. This level of analysis provides insights that standard CFD might not capture, particularly in understanding how different particles behave as a collective over time.

- The use of GMMs allows us to uncover hidden structures within the particle dispersion data. This approach can reveal complex, non-linear relationships in the data that CFD models, which rely on deterministic equations, might overlook. The resulting clusters can indicate different dispersion regimes or patterns that are not readily apparent in a conventional CFD analysis.

- Our approach integrates data from multiple spatial and temporal dimensions, providing a more holistic view of particle dispersion. Traditional CFD studies might analyze these dimensions separately or focus on a particular aspect, such as airflow patterns, without fully integrating them to discern broader patterns. By contrast, our method can simultaneously analyze these dimensions to identify clusters that reflect real-world particle behavior more accurately.

- By identifying clusters representing different dispersion behaviour, our method could improve the accuracy of predictions about where particles will travel under various conditions. Traditional CFD models might predict average behavior well. Still, our method's ability to identify specific clusters allows for more precise predictions, particularly in complex environments like a hospital room or a cruise ship cabin, where airflow and particle interactions can be highly variable.

Our method significantly advances traditional CFD approaches by integrating unsupervised and stochastic clustering techniques. This combination allows us to identify and analyze complex patterns in particle dispersion that traditional CFD studies might miss. Because of our approach's enhanced analytical depth, multidimensional integration and adaptive nature, direct comparisons with conventional CFD studies are not straightforward. Instead, our method should be seen as a more sophisticated tool that provides more profound and more precise insights into the dynamics of particle dispersion, especially in complex environments.

VI. CONCLUSIONS

In this study, CFD data were analyzed with the aid of Unsupervised Machine Learning to study airflow dynamics in a cruise ship cabin where a coughing individual was present. The trajectories of emitted droplets of different sizes were recorded in all three spatial dimensions. A typical ventilation rate was employed. Initially, the RUN-ICON algorithm was used to determine the optimum clustering configuration. Then, GMM was utilized to introduce stochasticity and account for the probabilistic nature of virus transmission dynamics.

The primary directions of particle dispersion are the streamwise ($y-$) and vertical ($z-$) directions, influenced by the airflow and gravity, respectively. The maximum lateral ($x-$) distance is approximately 12-13% of the distances covered in the other two directions. As particles disperse, evaporation reduces their number by approximately 21%, and by the end of the flow, only about 0.6% of particles fall below the 5 μm threshold.

In most cases, RUN-ICON results were not altered significantly by applying GMM, especially when the flow was primarily in one direction (i.e., streamwise and lateral). Significant differences were observed only at the initial stages when particles moved with similar velocities in multiple directions. The clustering seems to be size-based and reflects the dynamics of the flow. The similarity in clusters produced by both RUN-ICON and GMM indicated predictable patterns in particle dispersion, with smaller particles traveling further due to their ability to stay suspended in the air longer. This is particularly relevant for understanding the spread of airborne particles, such as respiratory emissions, where smaller particles could remain airborne and travel longer distances, potentially increasing the risk of transmission in enclosed spaces.

- Lateral direction: As mentioned, this is the less significant direction, with particles dispersing on either side of the mouth but without travelling substantial distances.

- Streamwise direction: Initially, smaller particles moved faster, with larger particles catching up later. However, since the larger particles reached the floor faster, they lost momentum, which enabled smaller particles to move further away in the streamwise direction. However, no particles reached a distance greater than 1.2 m from the mouth in that direction.

- Vertical direction: In this direction, the separation seemed to be size-based directly from the start due to gravity. By the final time of 8 s, almost all particles (97%)) were at a vertical distance below 1 m from the mouth.

The combination of RUN-ICON with GMMs in this study offers a robust framework for analyzing complex particle dynamics. RUN-ICON effectively identifies critical clusters within the data, providing a solid base that GMMs can use to model the underlying distributions more precisely. This integration enables a more detailed and probabilistic characterization of particle dispersion, capturing subtle variations and sub-populations within the trajectories. Consequently, this approach enhances the results' accuracy and interpretability, making it particularly useful for applications like modeling virus transmission in enclosed environments. As utilized in this work, the dual-method approach highlights the importance of considering initial conditions and dynamic interactions among particles to understand their dispersion accurately. Moreover, the present study provides insights into the behavior of respiratory droplets and aerosols, which are crucial for developing effective mitigation strategies in controlling airborne disease transmission.

The proposed combination of RUN-ICON for identifying potential cluster centre and their refinement with GMM can be an effective approach in various research areas, including beyond classical fluid dynamics, such as:

- Air quality analysis: Clustering data from various sensors to identify patterns and sources of air pollution.

- Wildlife monitoring: Analyzing sensor data to track animal movements and identify species in diverse habitats.
- Anomaly detection in transactions: Detecting unusual patterns in financial transactions that might indicate fraud or other anomalies.
- Market segmentation: Identifying different segments of customers based on their purchasing behavior or financial activities.
- Network intrusion detection: Modeling regular network traffic and identifying deviations that may indicate security breaches.
- Malware classification: Clustering software behaviors to identify new or evolving malware strains.

In these scenarios, the approach of utilizing RUN-ICON to find initial cluster centers before applying GMM can help to:

- Improve the accuracy of the GMM by providing it with a better initial estimate of cluster centers.
- Reduce the complexity of GMM fitting by starting with a more manageable number of well-defined clusters.
- Enhance interpretability by providing a clearer structure and distribution of the data.

This combination can leverage the strengths of both approaches, leading to more robust and insightful models in these diverse research areas.

VII. ACKNOWLEDGEMENTS

This paper is supported by the European Union's Horizon Europe Research and Innovation Actions programme under grant agreement No 101069937, project name: HS4U (HEALTHY SHIP 4U). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure, and Environment Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

VIII. AUTHOR DECLARATIONS

The authors have no conflicts to disclose.

IX. DATA AVAILABILITY

The authors' data supporting this study's findings are available upon reasonable request.

REFERENCES

[1]L. A. Caldera-Crespo, M. J. Paidas, S. Roy, C. I. Schulman, N. S. Kenyon, S. Daunert, and A. R. Jayakumar, "Experimental models of covid-19," Frontiers in Cellular and Infection Microbiology **11** (2022), 10.3389/fcimb.2021.792584.
[2]M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2008).
[3]J. Wang, "Mathematical models for covid-19: applications, limitations, and potentials," Journal of Public Health and Emergency **4** (2020).
[4]T. Dbouk and D. Drikakis, "On coughing and airborne droplet transmission to humans," Physics of Fluids **32**, 053310 (2020).
[5]T. Dbouk and D. Drikakis, "On respiratory droplets and face masks," Physics of Fluids **32**, 063303 (2020).
[6]N. Christakis, V. Barbaris, and A. Spentzos, "A new approach in financial modelling with the aid of artificial neural networks," Journal of Algorithms & Computational Technology **5**, 513–529 (2011), https://doi.org/10.1260/1748-3018.5.3.513.
[7]N. Christakis, P. Tirchas, M. Politis, M. Achladianakis, E. Avgenikou, and G. Kossioris, "Covid-liberty, a machine learning computational framework for the study of the covid-19 pandemic in europe. part 2: Setting up the framework with ensemble modeling," International Journal of Neural Networks and Advanced Applications , 27–42 (2021).
[8]K. Poulinakis, D. Drikakis, I. W. Kokkinakis, and S. M. Spottswood, "Machine-learning methods on noisy and sparse data," Mathematics **11** (2023), 10.3390/math11010236.
[9]D. Drikakis and F. Sofos, "Can artificial intelligence accelerate fluid mechanics research?" Fluids **8** (2023), 10.3390/fluids8070212.
[10]X. Shen, C. Jiang, Y. Wen, C. Li, and Q. Lu, "A brief review on deep learning applications in genomic studies," Frontiers in Systems Biology **2** (2022), 10.3389/fsysb.2022.877717.
[11]A. Vlachos, "Evaluating unsupervised learning for natural language processing tasks," in *Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011) pp. 35–42.
[12]K. T. Matchev, K. Matcheva, and A. Roman, "Unsupervised machine learning for exploratory data analysis of exoplanet transmission spectra," The Planetary Science Journal **3**, 205 (2022).
[13]M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," in *Supervised and Unsupervised Learning for Data Science*, edited by M. W. Berry, A. Mohamed, and B. W. Yap (Springer International Publishing, Cham, 2020) pp. 3–21.
[14]N. Christakis and D. Drikakis, "Reducing uncertainty and increasing confidence in unsupervised learning," Mathematics **11** (2023), 10.3390/math11143063.
[15]N. Christakis and D. Drikakis, "Unsupervised learning of particles dispersion," Mathematics **11** (2023), 10.3390/math11173637.
[16]N. Christakis, D. Drikakis, K. Ritos, and I. W. Kokkinakis, "Unsupervised machine learning of virus dispersion indoors," Physics of Fluids **36** (2024).
[17]T. Chen, J. Morris, and E. Martin, "Probability density estimation via an infinite gaussian mixture model: application to statistical process monitoring," Journal of the Royal Statistical Society Series C: Applied Statistics **55**, 699–715 (2006).
[18]WHO, "Who coronavirus disease (covid-19) dashboard data," (2021).
[19]Y. Yan, X. Li, X. Fang, P. Yan, and J. Tu, "Transmission of covid-19 virus by cough-induced particles in an airliner cabin section," Engineering Applications of Computational Fluid Mechanics **15**, 934–950 (2021).
[20]W. Wang, F. Wang, D. Lai, and Q. Chen, "Evaluation of sars-cov-2 transmission and infection in airliner cabins," Indoor Air **32**, e12979 (2022).
[21]F. Wang, T. Zhang, R. You, and Q. Chen, "Evaluation of infection probability of Covid-19 in different types of airliner cabins," Build Environ **234**, 110159 (2023).
[22]T. Dbouk and D. Drikakis, "On pollen and airborne virus transmission," Physics of Fluids **33(6)**, 063313 (2021).
[23]T. Dbouk and D. Drikakis, "Weather impact on airborne coronavirus survival," Physics of Fluids **32**, 093312 (2020).
[24]T. Dbouk and D. Drikakis, "On airborne virus transmission in elevators and confined spaces," Physics of Fluids **33**, 011905 (2021).

[25]S. Bushwick, T. Lewis, and A. Montañez, "Evaluating COVID Risk on Planes, Trains and Automobiles," (2020).
[26]V. Mathai, A. Das, J. Bailey, and K. Breuer, "Airflows inside passenger cars and implications for airborne disease transmission," Science Advances **7**, eabe0166 (2021).
[27]V. Mathai, A. Das, and K. Breuer, "Aerosol transmission in passenger car cabins: Effects of ventilation configuration and driving speed," Physics of Fluids **34**, 021904 (2022).
[28]K. Luo, Z. Lei, Z. Hai, S. Xiao, J. Rui, H. Yang, X. Jing, H. Wang, Z. Xie, P. Luo, W. Li, Q. Li, H. Tan, Z. Xu, Y. Yang, S. Hu, and T. Chen, "Transmission of SARS-CoV-2 in Public Transportation Vehicles: A Case Study in Hunan Province, China," Open Forum Infectious Diseases **7**, ofaa430 (2020).
[29]Y. Shen, C. Li, H. Dong, Z. Wang, L. Martinez, Z. Sun, A. Handel, Z. Chen, E. Chen, M. H. Ebell, F. Wang, B. Yi, H. Wang, X. Wang, A. Wang, B. Chen, Y. Qi, L. Liang, Y. Li, F. Ling, J. Chen, and G. Xu, "Community Outbreak Investigation of SARS-CoV-2 Transmission Among Bus Riders in Eastern China," JAMA Internal Medicine **180**, 1665–1671 (2020).
[30]P. Azimi, Z. Keshavarz, J. Laurent, B. R. Stephens, and J. G. Allen, "Mechanistic transmission modeling of covid-19 on the diamond princess cruise ship demonstrates the importance of aerosol transmission," medRxiv (2020), 10.1101/2020.07.13.20153049.
[31]L. Moriarty, M. Plucinski, and e. a. Marston, B.J., "Public health responses to covid-19 outbreaks on cruise ships — worldwide," MMWR Morb Mortal Wkly Rep **69**, 347–352 (2020).
[32]E. C. Rosca, C. Heneghan, E. A. Spencer, J. Brassey, A. Plüddemann, I. J. Onakpoya, D. Evans, J. M. Conly, and T. Jefferson, "Transmission of SARS-CoV-2 Associated with Cruise Ship Travel: A Systematic Review," Tropical medicine and infectious disease **7**, 290 (2022).
[33]L. Huang, I. Riyadi, S.and Utama, M. Li, P. Sun, and G. Thomas, "Covid-19 transmission inside a small passenger vessel: Risks and mitigation," Ocean Engineering **255**, 111486 (2022).
[34]A. Saunders, "Cruise lines change ship ventilation systems as part of overall covid strategy," (2020).
[35]K. Wiles, "Cruise ship ac systems could promote rapid coronavirus spread, prof says," (2020).
[36]O. Almilaji, "Air Recirculation Role in the Spread of COVID-19 Onboard the Diamond Princess Cruise Ship during a Quarantine Period," Aerosol and Air Quality Research **21**, 200495 (2021).
[37]J. Zhou, S. P. Chen, W. W. Shi, M. Kanrak, and J. Ge, "The impacts of COVID-19 on the cruise industry based on an empirical study in china," Marine policy **153**, 105631 (2023).
[38]M. Z. Bazant and J. W. M. Bush, "A guideline to limit indoor airborne transmission of COVID-19," Proc Natl Acad Sci USA **118**, 1–12 (2021).
[39]CDC, "COVID-19 Ventilation in Buildings 2023," The Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/community/ventilation.html (2023).
[40]ASHRAE, "ANSI/ASHRAE Standard 241-2023, Control of Infectious Aerosols," (2023).
[41]ASHRAE, "ANSI/ASHRAE Standard 62.1-2019, Ventilation and Acceptable Indoor Air Quality," (2019).
[42]WHO, "Roadmap to improve and ensure good indoor ventilation in the context of COVID-19," World Health Organization. https://www.who.int/publications/i/item/9789240021280 (2021).
[43]REHVA, "COVID-19 guidance 4.1, how to operate HVAC and other building service systems to prevent the spread of the coronavirus (SARS-CoV-2) disease (COVID-19) in workplaces," Federation of European Heating, Ventilation and Air Conditioning Associations. https://www.rehva.eu/activities/covid-19-guidance/rehva-covid-19-guidance (2021).
[44]Y. Li, P. Cheng, and W. Jia, "Poor ventilation worsens short-range airborne transmission of respiratory infection," Indoor Air **32**, e12946 (2022).
[45]FPS, "Legal framework regarding indoor air quality," Federal Public Service - Public Health. https://www.health.belgium.be/en/closer-legal-framework-indoor-air-quality (2022).
[46]J. G. Allen and J. D. Macomber, *Healthy Buildings: How Indoor Spaces Can Make You Sick–Or Keep You Well*, 2nd ed. (Harvard University Press, 2022).
[47]ASHRAE, "ANSI/ASHRAE Standard 62.2-2019, Ventilation and Acceptable Indoor Air Quality in Residential Buildings," (2019).
[48]K. Ritos, D. Drikakis, and I. Kokkinakis, "Virus spreading in cruiser cabin," Physics of Fluids (2023), https://doi.org/10.1063/5.0169992.
[49]X. Zhao, S. Liu, Y. Yin, T. T. Zhang, and Q. Chen, "Airborne transmission of covid-19 virus in enclosed spaces: An overview of research methods," Indoor Air **32**, e13056 (2022).
[50]R. Dhand and J. Li, "Coughs and sneezes: Their role in transmission of respiratory viral infections, including sars-cov-2," AJRCCM **202** (2020), 10.1164/rccm.202004-1263PP.
[51]B. Panić, J. Klemenc, and M. Nagode, "Gaussian mixture model based classification revisited: Application to the bearing fault classification." Journal of Mechanical Engineering/Strojniški Vestnik **66** (2020).
[52]P. E. Jebarani, N. Umadevi, H. Dang, and M. Pomplun, "A novel hybrid k-means and gmm machine learning model for breast cancer detection," IEEE Access **9**, 146153–146162 (2021).
[53]M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A comprehensive evaluation of optics, gmm and k-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins," Geochemistry , 126094 (2024).
[54]C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler, "Statistical mixture modeling for cell subtype identification in flow cytometry," Cytometry Part A: The Journal of the International Society for Analytical Cytology **73**, 693–701 (2008).
[55]S. M. Khansari-Zadeh and A. Billard, "Learning stable nonlinear dynamical systems with gaussian mixture models," IEEE Transactions on Robotics **27**, 943–957 (2011).
[56]J. Wiest, M. Höffken, U. Kreßel, and K. Dietmayer, "Probabilistic trajectory prediction with gaussian mixture models," in *2012 IEEE Intelligent vehicles symposium* (IEEE, 2012) pp. 141–146.
[57]Y. B. Varolgüneş, T. Bereau, and J. F. Rudzinski, "Interpretable embeddings from molecular simulations using gaussian mixture variational autoencoders," Machine Learning: Science and Technology **1**, 015012 (2020).
[58]M. Bitaab and S. Hashemi, "Hybrid intrusion detection: Combining decision tree and gaussian mixture model," in *2017 14th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC)* (IEEE, 2017) pp. 8–12.
[59]H. Wan, H. Wang, B. Scotney, and J. Liu, "A novel gaussian mixture model for classification," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (IEEE, 2019) pp. 3298–3303.
[60]L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: clustering, classification and density estimation using gaussian finite mixture models," The R journal **8**, 289 (2016).
[61]A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Information Sciences **622**, 178–210 (2023).
[62]D. A. Reynolds, *A Gaussian mixture modeling approach to text-independent speaker identification* (Georgia Institute of Technology, 1992).
[63]Y. Drossinos, T. P. Weber, and N. I. Stilianakis, "Droplets and aerosols: An artificial dichotomy in respiratory virus transmission," Health Science Reports **4**, e275 (2021).
[64]K. Randall, E. T. Ewing, L. C. Marr, J. L. Jimenez, and L. Bourouiba, "How did we get here: what are droplets and aerosols and how far do they go? a historical perspective on the transmission of respiratory infectious diseases," Interface Focus **11**, 20210049 (2021).
[65]M. L. Pöhlker, C. Pöhlker, O. O. Krüger, J.-D. Förster, T. Berkemeier, W. Elbert, J. Fröhlich-Nowoisky, U. Pöschl, G. Bagheri, E. Bodenschatz, *et al.*, "Respiratory aerosols and droplets in the transmission of infectious diseases," Reviews of Modern Physics **95**, 045001 (2023).
[66]T. N. Verma, A. K. Sahu, and S. L. Sinha, "Study of particle dispersion on one bed hospital using computational fluid dynamics," Materials Today: Proceedings **4**, 10074–10079 (2017).
[67]H. Tan, K. Y. Wong, M. H. D. Othman, H. Y. Kek, R. A. Wahab, G. K. P. Ern, W. T. Chong, and K. Q. Lee, "Current and potential approaches on assessing airflow and particle dispersion in healthcare facilities: a systematic review," Environmental Science and Pollution Research **29**, 80137–80160 (2022).