Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Project Plan

Data assimilation using Adversarial Neural Networks to help determine COVID infection risks in enclosed spaces

by

Shiqi Yin

Email: shiqi.yin21@imperial.ac.uk
GitHub username: acse-sy121
Repository: https://github.com/ese-msc-2021/irp-sy121

Supervisors:

Prof. Christopher Pain

Dr. Claire Heaney
Boyang Chen

June 2022

1 Introduction

COVID-19 is an outbreak that was first detected and broke out in Wuhan, China in late December 2019 [8], this epidemic is a pneumonia caused by the SARS-CoV-2 virus. Due to multiple mutations of the virus, the pathogenicity of the virus has decreased but its infectivity has increased significantly. The negative impact of the epidemic on the world is still severe, affecting hundreds of countries and retarding their economic development[5].

In a study of the trajectories of infected people [12], researchers found that SARS-CoV-2, the causative agent of the new infectious disease, has a strong capacity for person-to-person transmission and according to the report from WHO [1], COVID-19 can be transmitted by droplets, by touching the eyes, mouth or nose after contact with objects carrying the virus, or in poorly ventilated rooms. The virus's ability to transmit allows COVID-19 to be transmitted from an infected person to healthy people in an enclosed space. The study of the transmission of COVID-19 is therefore particularly important for outbreak prevention and control, and one such study is the study of COVID-19 transmission in enclosed spaces.

For the transmission of COVID-19 in enclosed spaces, aerosol transmission is a very common mode of virus transmission. In Björn Birnir's paper[3], he modelled the air flow in a restaurant and by analysing the changes in virus concentration levels in this enclosed space, concluded that the COVID-19 pandemic outbreak in the restaurant was due to an increase in the concentration level of virus-carrying droplets produced by aerosols carrying the SARS-CoV-2 virus in the air. And when people stay in such enclosed spaces for long periods of time, they become infected. Therefore, aerosol transmission is one of the main modes of transmission of COVID-19 in enclosed spaces and can be simulated by modelling aerosol transmission in enclosed spaces. As aerosol transmission is similar to the transmission of CO_2 from human respiration in enclosed spaces, this project uses a model of CO_2 transmission in enclosed spaces to simulate the transmission of COVID-19 and to predict the transmission of the virus in future times.

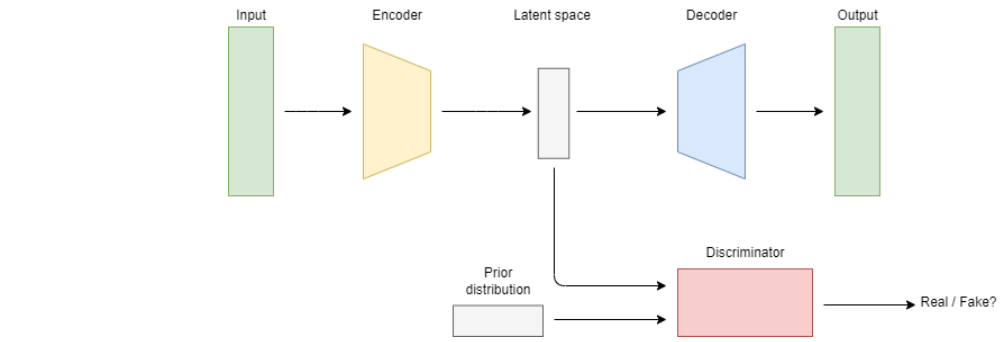
2 Literature Review

With the global outbreak of the COVID-19 epidemic, many scientific departments have begun to investigate the transmission of the epidemic, one method is Computational Fluid Dynamics (CFD), a model proposed by J.L. Hess and A.M.O. Smith[7] in 1967 and subsequently used in a variety of applications, including the modelling of the transmission of the virus. In Khalid M. Saqr's paper[13], he used CFD to model the transmission of COVID-19 in the air and proposed a reproducibility index for the COVID-19 model. Although CFD models can simulate the transmission of COVID-19 in air with high precision, the high computational cost make CFD models compute slowly [13].

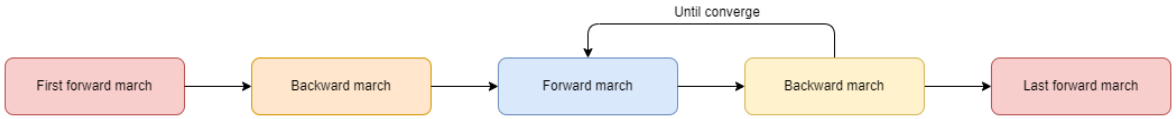
In terms of prediction, machine learning can also yield good prediction results and therefore neural networks can be used to predict the transmission of COVID-19 in air, one of the neural network models that can be used is the autoencoder (AE). The concept of autoencoder was proposed by Yann LeCun in 1987[9]. It consists of an Encoder and a Decoder. The input data is encoded by encoder, and then the data encoded by encoder is reconstructed by the decoder after decoding. Zhai et al. mentioned in their paper that traditionally, autoencoders have been used mainly for dimensionality reduction of data, but with the popularity of various machine learning models, autoencoders are also used as generative adversarial networks (GANs) in recent years[15]. However, GAN is difficult to train, resulting in trained models that may not work as well as expected. Therefore, this requires researchers to find a model to replace GAN. Makhzani et al. mentioned in their paper that adversarial autoencoder (AAE) is a combination of AE and GAN and it is easier to train. The main difference between AAE and GAN is the input to D. In AAE, the input to D is the latent space data generated by encoder in autoencoder part and the prior[2]. The core part to train AAE is similar to GAN, which is an adversarial training between encoder and discriminator. Then, the decoder can be well trained

by training the autoencoder part of the AAE model and generate reasonable data[11]. Figure 1a shows the model structure of AAE.

For the algorithms that will be used in this project which are prediction algorithm and data assimilation, in the paper by Vinicius L.S. Silva et al., the new model have the ability to predict the next m timestamps of data given the first n timestamps of data. This neural network model is called Pred-GAN. In addition, the paper mentions the application of data assimilation to Pred-GAN to improve the accuracy of the model predictions, which built the new model, DA-Pred-GAN[14]. Figure 1b shows the process of data assimilation.



(a) Architecture of the AAE model



(b) Progress of data assimilation

Figure 1: Description of AAE and data assimilation

A method used for pre-process the data is Reduced-order modelling (ROM). In Marco Fahl and Ekkehard W. Sachs's paper[6], they mentioned that Reduced-order modelling is required to decrease the computational difficulties, and one way of implementing reduced-order modelling is proper orthogonal decomposition. It was mentioned in the paper written by Liang et al.that there are three main methods of proper orthogonal decomposition, which are principal component analysis (PCA), arhunen–Loève decomposition (KLD) and singular value decomposition (SVD)[10].

3 Description of Problem and Objectives

Modelling the transmission of COVID-19 in air is complex due to the fact that air movement varies under different indoor conditions. Although Khalid M. Saqr has implemented CFD models to simulate the transmission of COVID-19[13], machine learning is also required to predict the transmission of COVID-19 due to the high computational cost of CFD models. Therefore, the objective of this project is to build an AAE model to predict the transmission of COVID-19 in enclosed spaces and use data assimilation to optimize the results of the prediction.

To achieve the objective, the dataset first needs to be pre-processed when input. This is due to the fact that the dimensionality of the input data is high that the model cannot accept data with such high dimensionality, one available method is to apply PCA for dimensionality reduction[4]. The second part of the pre-processing is to apply MinMaxScalar, which converts the range of the data to [0, 1], with the following code example.

```
from sklearn.decomposition import PCA
pca_compress = PCA(n_components=d)
X_train_compressed = pca_compress.fit_transform(input_data)
```

```
from sklearn.preprocessing import MinMaxScaler
scaler_minmax = MinMaxScaler((0,1))
X_data = scaler_minmax.fit_transform(CO2_data)
```

The data we will use in this project is a 3D dataset collected in a classroom, and generated by the CFD model. The dimension of the data is around 1 million and this requires us to use PCA to decrease the dimension. Figure 2 shows the schematic of the classroom where the data in dataset was collected and CO2 isosurface.

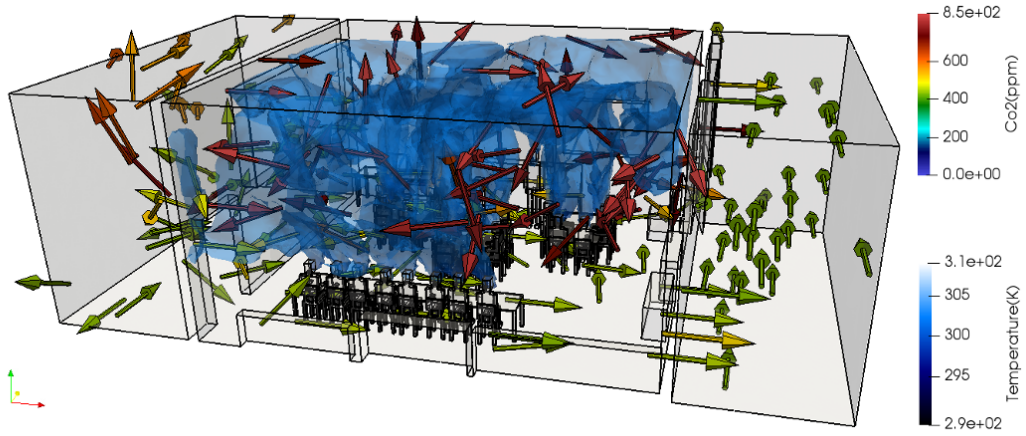
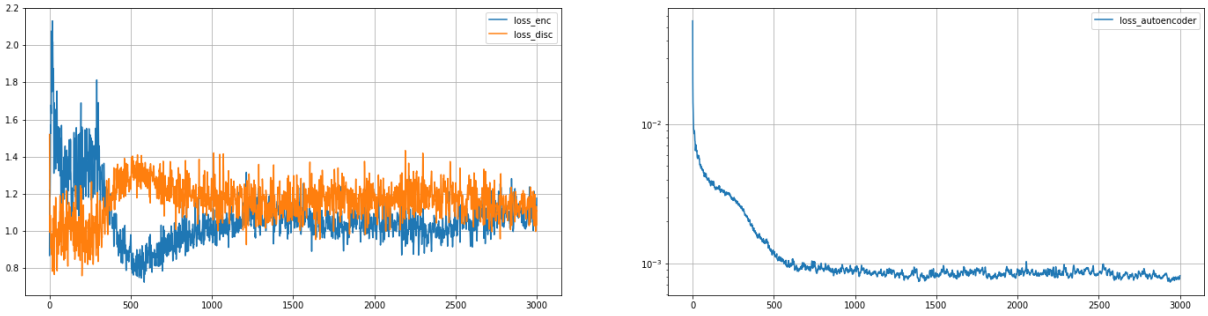


Figure 2: Visualization of data in the classroom

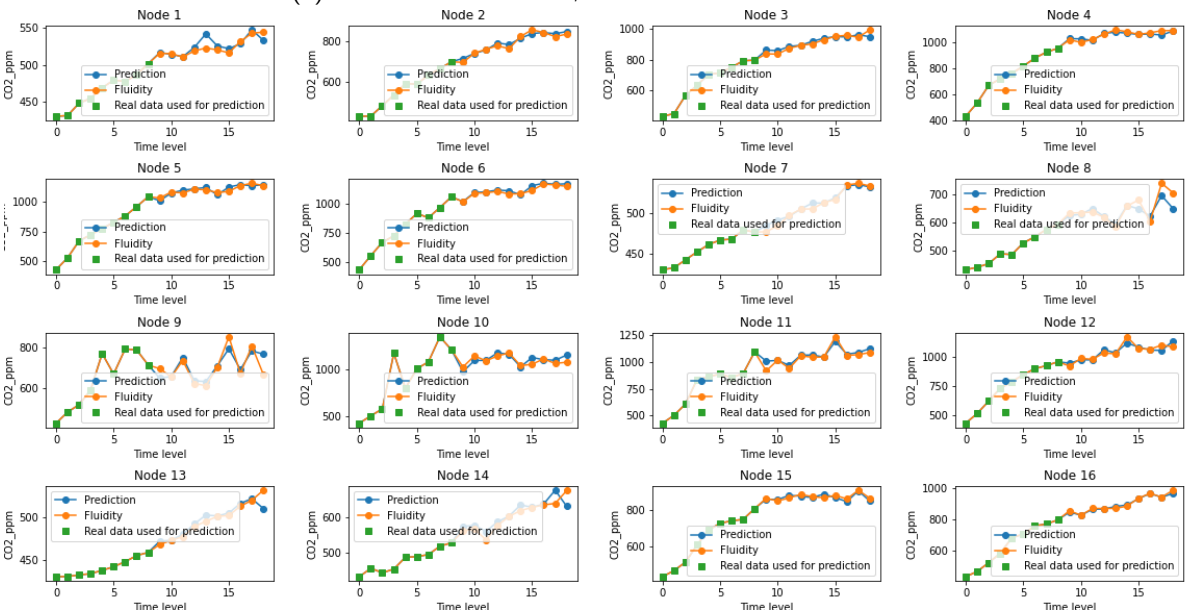
After pre-processing the data, we will build the AAE model, use the pre-processed data to generate the training set, train the built AAE model and output the loss of the model during training to determine whether the loss of the model converges. If we obtain the trained model, we will combine the prediction algorithm with the generated model by inputting data for the first m timestamps to obtain predicted CO_2 concentration levels for the next n timestamps and comparing them to the CO_2 concentration levels obtained from the CFD model. Subsequently, data assimilation will be applied to the model, and we will repeat the above steps and try to find the difference in precision between the predicted values in the two cases.

4 Progress to Date and Future Plan

For the current progress, I have pre-processed the existing data collected at several points and successfully built the AAE model and trained the model with the processed data, obtaining the trained model and outputting a loss. The prediction algorithm was then applied to the model and produced reasonable predictions and the results are shown in Figures 3a and 3b.



(a) Losses of the encoder, discriminator and autoencoder



(b) Prediction of the model

Figure 3: Results of the model

The following is a tentative version of future plans, which may be subject to change as the project progresses.

1. Implement data assimilation and apply it to existing models to output predicted values, and compare the results with the results without data assimilation to explore the impact of data assimilation on the predicted results.
2. Find the best hyperparameters for AAE by using grid search.
3. Obtain the real 3D data of detection in early July, apply PCA dimensionality reduction on the data and modify the existing model to meet the data, repeat the progress to date and step1.
4. Use code to show the difference between predicted data and data generated by CFD and visualize both data in a given enclosed space.
5. Work for the final report and presentation.

The Figure below shows the Gantt chart of the project plan.

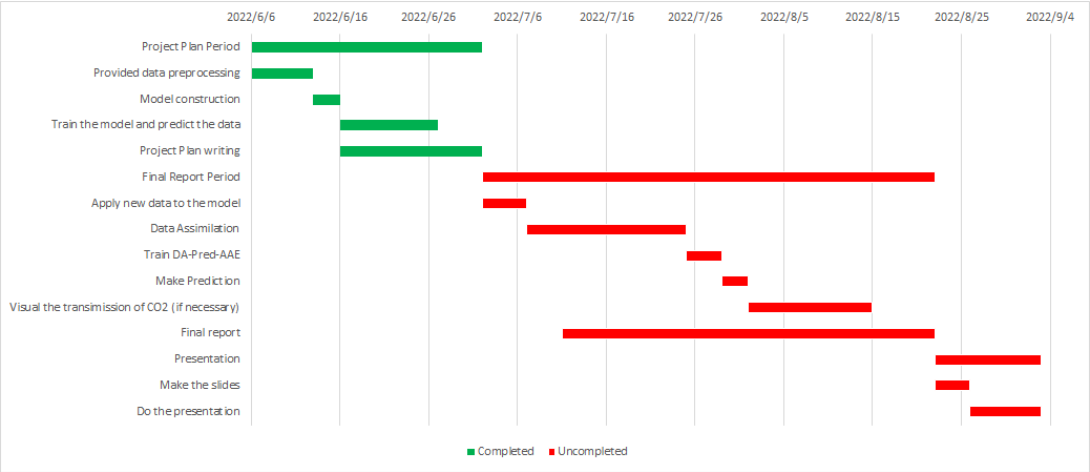


Figure 4: Gantt Chart of the project plan.

References

[1] Coronavirus disease (covid-19): How is it transmitted? `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted`. Accessed: 2022-06-21.

[2] Hamed Alqahtani, Manolya Kavakli-Thorne, and Gulshan Kumar. Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, (1), 2019.

[3] Björn Birnir. The build-up of aerosols carrying the sars-cov-2 coronavirus, in poorly ventilated, confined spaces. *medRxiv*, 2020.

[4] Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52(1):477–508, jan 2020.

[5] M. Chiah and A. Zhong. Trading from home: The impact of covid-19 on trading volume around the world. *Finance Research Letters*, 37, 2020.

[6] Marco Fahl and Ekkehard W. Sachs. Reduced order modelling approaches to pde-constrained optimization based on proper orthogonal decomposition. In Lorenz T. Biegler, Matthias Heinkenschloss, Omar Ghattas, and Bart van Bloemen Waanders, editors, *Large-Scale PDE-Constrained Optimization*, pages 268–280, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[7] J.L. Hess and A.M.O. Smith. Calculation of potential flow about arbitrary bodies. *Progress in Aerospace Sciences*, 8:1–138, 1967.

[8] D. S. Hui, E. I. Azhar, T. A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T. D. Mchugh, C Zamab, and C Cdab. The continuing 2019-ncov epidemic threat of novel coronaviruses to global health — the latest 2019 novel coronavirus outbreak in wuhan, china - sciencedirect. *International Journal of Infectious Diseases*, 91:264–266, 2020.

[9] Yann Lecun. *PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)*. Universite P. et M. Curie (Paris 6), June 1987.

[10] Y.C. LIANG, H.P. LEE, S.P. LIM, W.Z. LIN, K.H. LEE, and C.G. WU. Proper orthogonal decomposition and its applications—part i: Theory. *Journal of Sound and Vibration*, 252(3):527–544, 2002.

[11] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders, 2015.

[12] G. J. Milne and S. Xie. The effectiveness of social distancing in mitigating covid-19 spread: a modelling analysis. *Cold Spring Harbor Laboratory Press*, 2020.

[13] Khalid M. Saqr. Amicus plato, sed magis amica veritas: There is a reproducibility crisis in covid-19 computational fluid dynamics studies, 2021.

[14] Vinicius L. S. Silva, Claire E. Heaney, Yaqi Li, and Christopher C. Pain. Data assimilation predictive gan (da-predgan): applied to determine the spread of covid-19, 2021.

[15] Junhai Zhai, Sufang Zhang, Junfen Chen, and Qiang He. Autoencoder and its various variants. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 415–419, 2018.