UNIVERSITY
OF ICELAND

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

UNIVERSITY OF ICELAND

MIPRO2022 – DS-BE

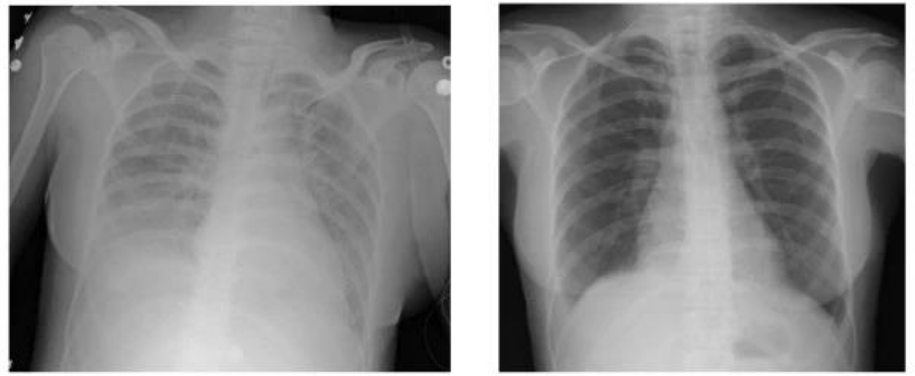
Lessons Learned on Using High-Performance Computing and Data Science Methods Towards Understanding the Acute Respiratory Distress Syndrome (ARDS)

C. Barakat, S. Fritsch, K. Sharafutdinov, G. Ingólfsson, A. Shuppert, S. Brynjólfsson, M. Riedel

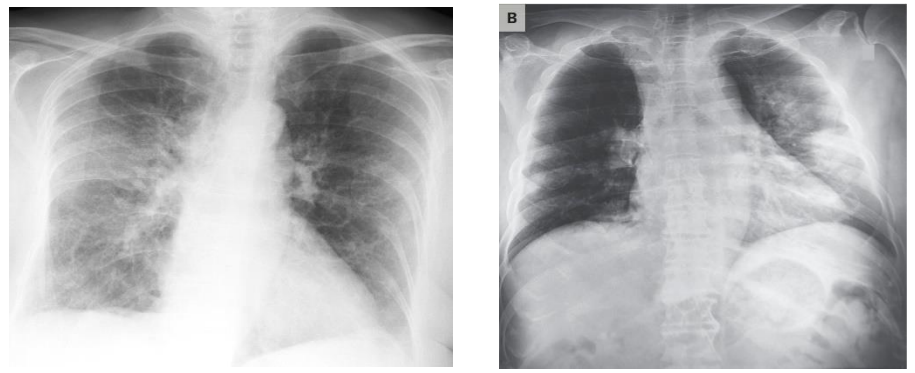
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Introduction

- Acute Respiratory Distress Syndrome (ARDS) affects many mechanically-ventilated ICU patients with a high mortality rate.
  - Causes infiltrates in the pulmonary compartments[1].
  - Gradual decrease in oxygenation[1].
  - Early diagnosis is associated with positive outcomes for the patients[2,3].
  - Generally diagnosed based on the Berlin Definition[4].
- Covid-19 caused by SARS-Cov-2 primarily affects the lungs[5,6].
  - Early diagnosis often associated with positive outcomes[5,6].
  - A lot of data available for analysis.



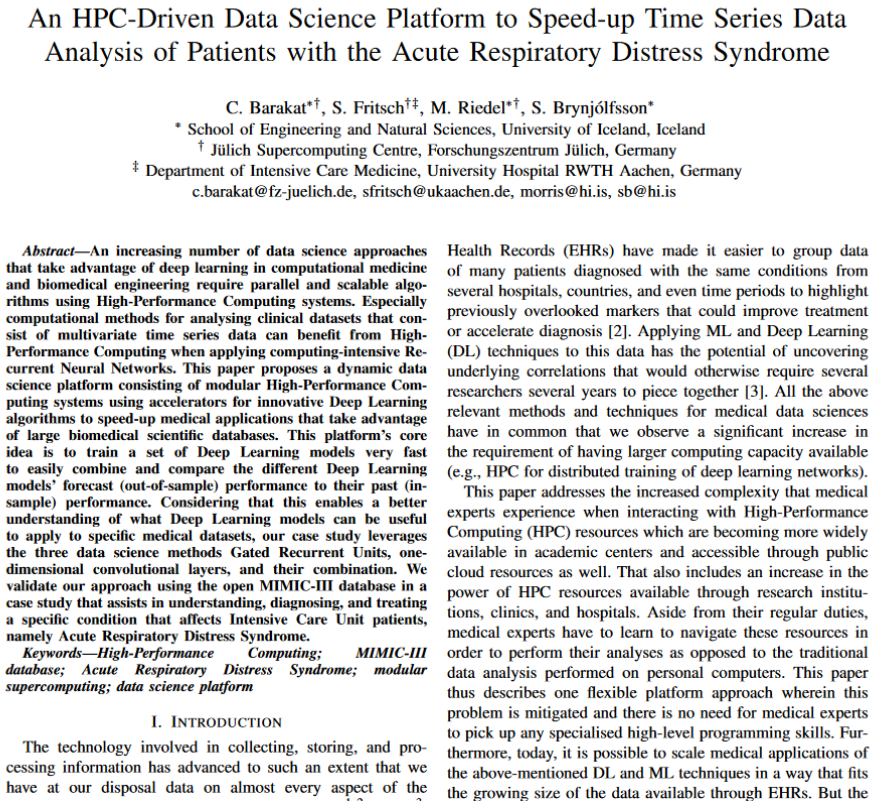
Comparison of ARDS (left) and healthy lungs (right)



Comparison of Covid-19 (left) and healthy lungs (right)

Aims of the Research

- Provide quick diagnosis methods for ARDS:
  - Generate patient-similar data to run simulations.
  - Design and build a neural network for ARDS diagnosis.
- Provide quick Covid-19 diagnosis:
  - Validate COVID-Net.
  - Test it on new data.
  - Retrain and deploy.
- Validate the previously developed HPC-enabled data science platform[7].



An HPC-Driven Data Science Platform to Speed-up Time Series Data Analysis of Patients with the Acute Respiratory Distress Syndrome

C. Barakat*†, S. Fritsch†‡, M. Riedel*†, S. Brynjólfsson*
* School of Engineering and Natural Sciences, University of Iceland, Iceland
† Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany
‡ Department of Intensive Care Medicine, University Hospital RWTH Aachen, Germany
c.barakat@fz-juelich.de, sfritsch@ukaachen.de, morris@hi.is, sb@hi.is

***Abstract*—An increasing number of data science approaches that take advantage of deep learning in computational medicine and biomedical engineering require parallel and scalable algorithms using High-Performance Computing systems. Especially computational methods for analysing clinical datasets that consist of multivariate time series data can benefit from High-Performance Computing when applying computing-intensive Recurrent Neural Networks. This paper proposes a dynamic data science platform consisting of modular High-Performance Computing systems using accelerators for innovative Deep Learning algorithms to speed-up medical applications that take advantage of large biomedical scientific databases. This platform's core idea is to train a set of Deep Learning models very fast to easily combine and compare the different Deep Learning models' forecast (out-of-sample) performance to their past (in-sample) performance. Considering that this enables a better understanding of what Deep Learning models can be useful to apply to specific medical datasets, our case study leverages the three data science methods Gated Recurrent Units, one-dimensional convolutional layers, and their combination. We validate our approach using the open MIMIC-III database in a case study that assists in understanding, diagnosing, and treating a specific condition that affects Intensive Care Unit patients, namely Acute Respiratory Distress Syndrome.**

***Keywords*—High-Performance Computing; MIMIC-III database; Acute Respiratory Distress Syndrome; modular supercomputing; data science platform**

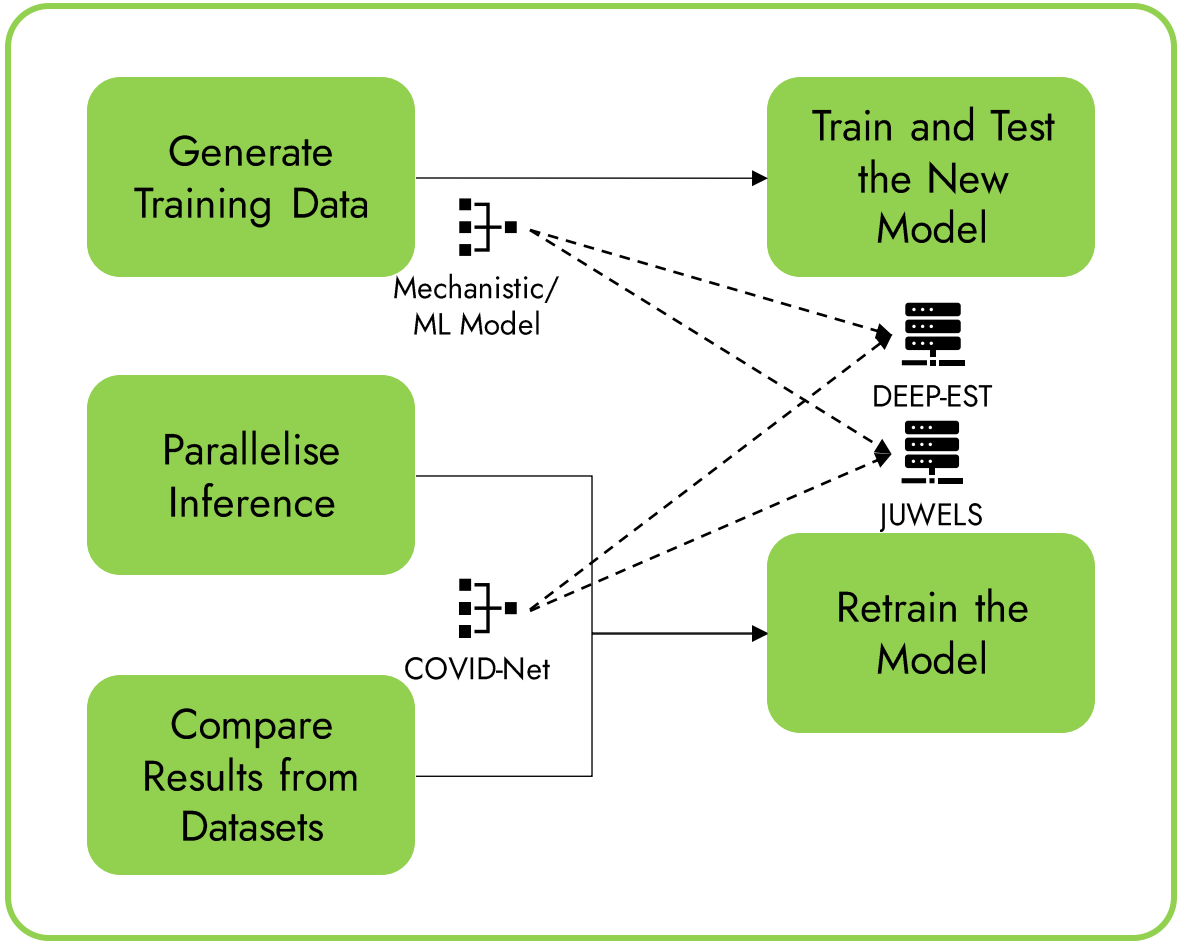
I. Introduction

The technology involved in collecting, storing, and processing information has advanced to such an extent that we have at our disposal data on almost every aspect of the Health Records (EHRs) have made it easier to group data of many patients diagnosed with the same conditions from several hospitals, countries, and even time periods to highlight previously overlooked markers that could improve treatment or accelerate diagnosis [2]. Applying ML and Deep Learning (DL) techniques to this data has the potential of uncovering underlying correlations that would otherwise require several researchers several years to piece together [3]. All the above relevant methods and techniques for medical data sciences have in common that we observe a significant increase in the requirement of having larger computing capacity available (e.g., HPC for distributed training of deep learning networks).

This paper addresses the increased complexity that medical experts experience when interacting with High-Performance Computing (HPC) resources which are becoming more widely available in academic centers and accessible through public cloud resources as well. That also includes an increase in the power of HPC resources available through research institutions, clinics, and hospitals. Aside from their regular duties, medical experts have to learn to navigate these resources in order to perform their analyses as opposed to the traditional data analysis performed on personal computers. This paper thus describes one flexible platform approach wherein this problem is mitigated and there is no need for medical experts to pick up any specialised high-level programming skills. Furthermore, today, it is possible to scale medical applications of the above-mentioned DL and ML techniques in a way that fits the growing size of the data available through EHRs. But the

Methodology

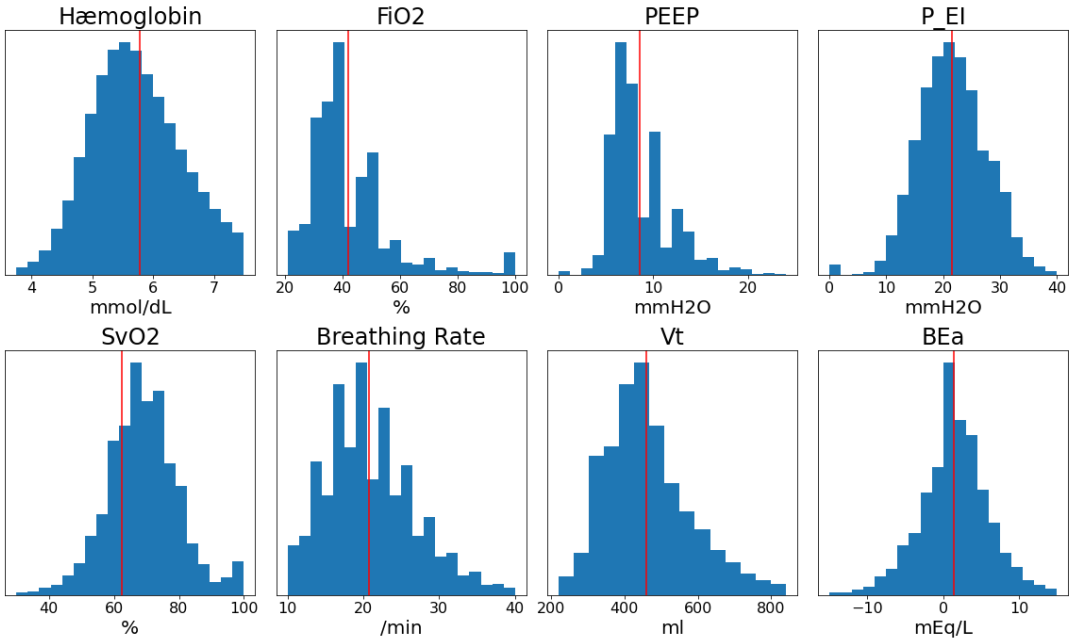
- Virtual Patient Simulator:
  - Matlab model ported to C.
  - Tuned outputs to most informative.
- Input Patient Data:
  - Understand distribution of available data.
  - Validate methods for random sampling.
  - Parallelise data generation.
- COVID-Net:
  - Validate original network.
  - Highlight speed up when using HPC
  - Test on data from hospital.
  - Retrain on the data.



Results – ARDS Diagnosis

- C-based model performance:
  - 100% comparable to the original simulation.
  - Faster in serial both locally and on HPC.
  - 100% faster in parallel on HPC, with potential for scale-up
- The patient parameters follow a Normal distribution:
  - Fits the boundaries defined by ICU personnel.
  - Can be modelled mathematically.
- Data generation can be parallelised.
- Python Generators are an efficient method for feeding into the simulation.

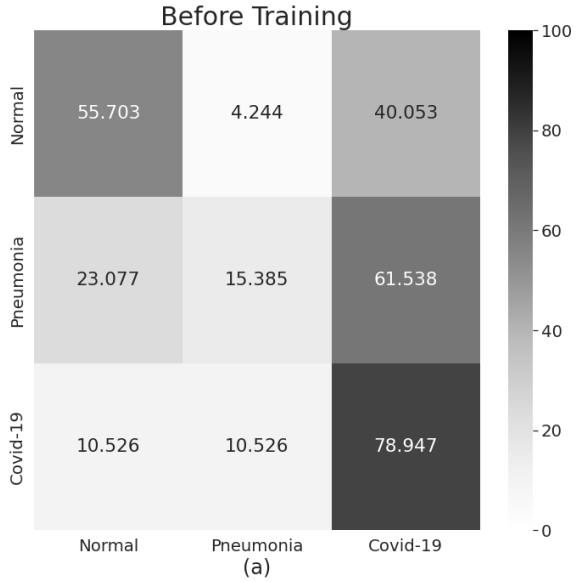
Platform	Execution Time
Original Simulation on Laptop	259.1 s
C in serial on DEEP with JupyterLab	108.8 s
C in parallel on DEEP on 48 CPUs	100.79 s

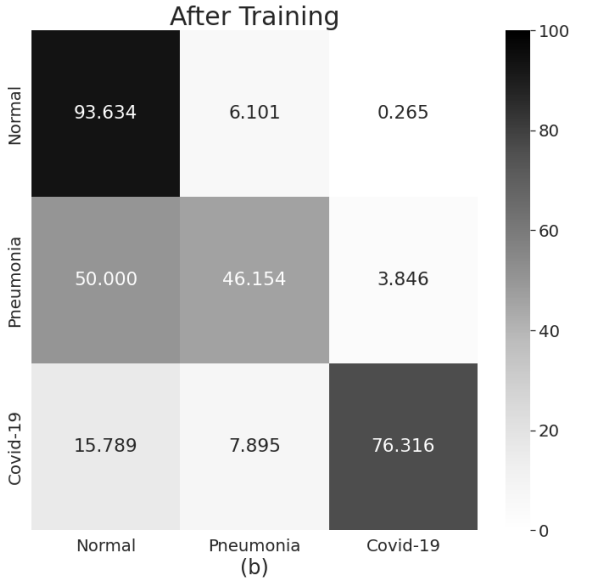


Results – Covid-19 Diagnosis

- Pre-trained COVID-Net:
  - Performs as expected on images similar the data it was trained on.
  - Underperforms on new images with different resolutions.
  - Inference was sped up and scaled up significantly.

- Retraining COVID-Net:
  - Almost impossible on local machine.
  - Using HPC made the process more efficient (speed-up, scale-up, multi-runs).
  - Retraining with new data showed improved inference.





Conclusions and Next Steps

UNIVERSITY OF ICELAND

- Validated the pre-established Data Science Platform:
  - Performed data analysis on available patient data.
  - Ported the virtual patient simulation to C.
  - Parallelised and scaled up virtual patient simulation.
  - Managed the transfer and storage of medical data for training COVID-Net.
  - Highlighted speed-ups possible using specialised hardware and software.
- Prepared the platform for large-scale data generation.
- Set up the groundwork for building the neural network-based simulation.
- Successfully retrained COVID-Net on new data.

Thank you!

Questions are Welcome

References

1. D. G. Ashbaugh, D. B. Bigelow, and B. E. Levine, 'Acute Respiratory Distress in Adults', *The Lancet*, vol. 290, no. 7511, pp. 319–323, Aug. 1967, doi: 10.1016/S0140-6736(67)90168-7.
2. M. Confalonieri, F. Salton, and F. Fabiano, 'Acute respiratory distress syndrome', *EUROPEAN RESPIRATORY REVIEW*, vol. 26, no. 144, p. 160116, Jun. 2017, doi: 10.1183/16000617.0116-2016.
3. S. Le *et al.*, 'Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS)', *Journal of Critical Care*, vol. 60, pp. 96–102, Dec. 2020, doi: 10.1016/j.jcrc.2020.07.019.
4. The ARDS Definition Task Force, 'Acute Respiratory Distress Syndrome: The Berlin Definition', *JAMA*, vol. 307, no. 23, Jun. 2012, doi: 10.1001/jama.2012.5669.
5. T. Acter, N. Uddin, J. Das, A. Akhter, T. R. Choudhury, and S. Kim, 'Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency', *Science of the Total Environment*, p. 138996, 2020.
6. G. French et al., 'Impact of Hospital Strain on Excess Deaths During the COVID-19 Pandemic — United States, July 2020–July 2021', *MMWR Morb Mortal Wkly Rep 2021*, vol. 70, no. 46, pp. 1613–1616, Nov. 2021, doi: 10.15585/mmwr.mm7046a5.
7. C. Barakat, S. Fritsch, M. Riedel, and S. Brynjolfsson, 'An HPC-Driven Data Science Platform to Speed-up Time Series Data Analysis of Patients with the Acute Respiratory Distress Syndrome', in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Croatia, Sep. 2021, pp. 311–316. doi: 10.23919/MIPRO52101.2021.9596840.
8. L. Wang, Z. Q. Lin, and A. Wong, 'COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images', *Sci Rep*, vol. 10, no. 1, p. 19549, Dec. 2020, doi: 10.1038/s41598-020-76550-z.