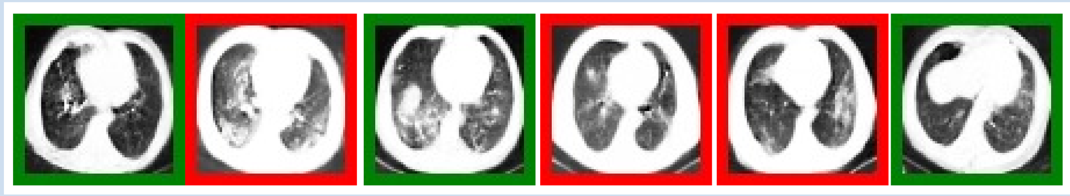
Conditional Synthetic Data Generation for Robust ML Applications with Limited Pandemic Data

Hari Prasanna Das†,Ryan Tran†,Japjot Singh†,Xiangyu Yue†,Geoffrey Tison*, Alberto Sangiovanni-Vincentelli†,Costas J. Spanos†

† *Department of Electrical Engineering and Computer Sciences, UC Berkeley* * *School of Medicine, UC San Francisco*

Introduction

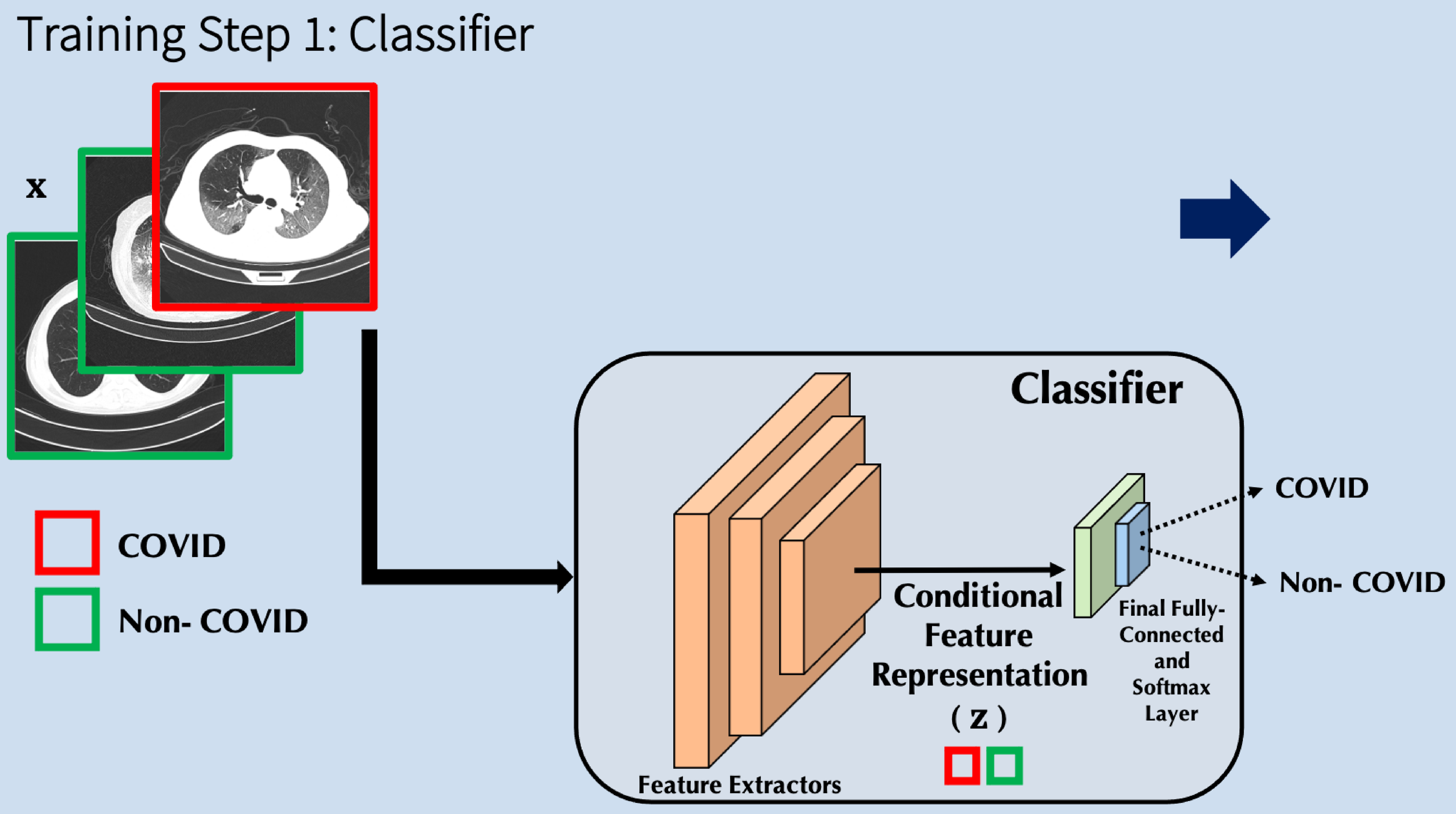
- At the onset of a pandemic, such as COVID-19, data with proper labeling/attributes corresponding to the new disease might be unavailable or sparse.
- At the same time, ML algorithms designed to fight pandemics must have good performance and be developed in a time-sensitive manner.

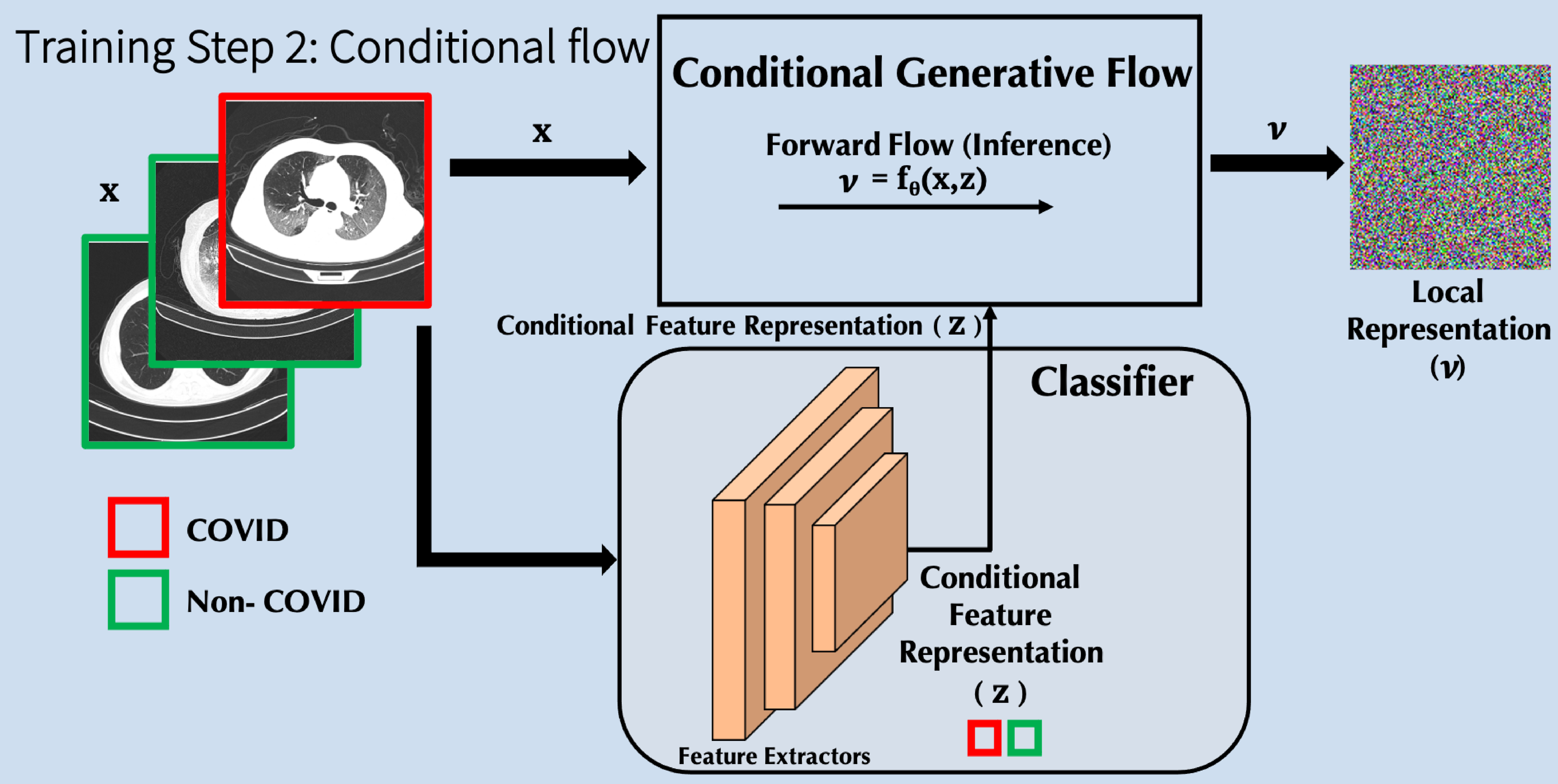


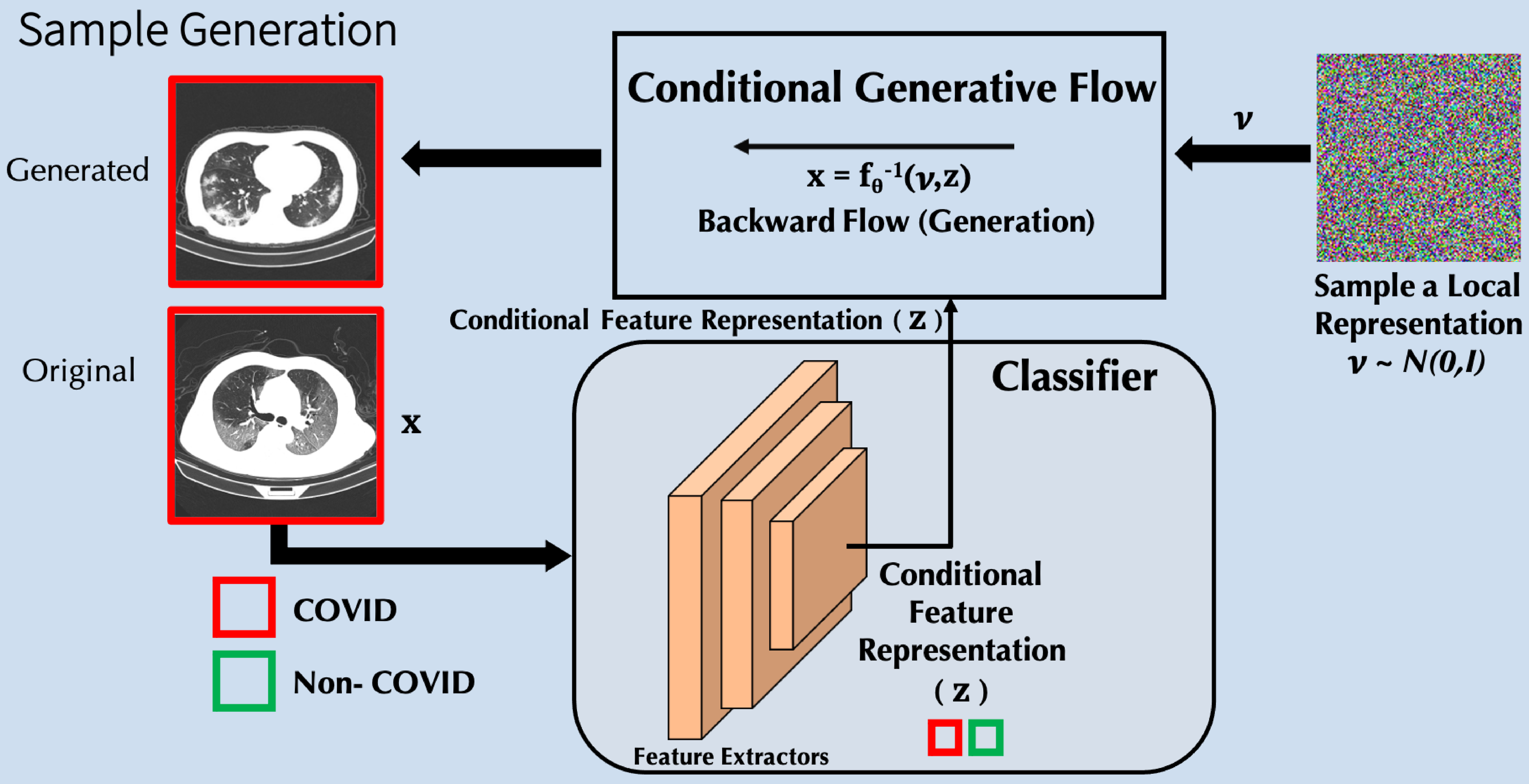
Synthetic CT scans generated by our model. Red border: COVID, Green: Non-COVID.

To tackle the challenges of limited pandemic data, we propose generating conditional synthetic data, to be used alongside real data for robust ML. We conduct experiments on chest CT scans corresponding to normal, COVID-19 and pneumonia afflicted patients.

Model Training and Synthetic Sample Generation







Training Step 1: Train a COVID/Non-COVID classifier to decouple the feature representations (z) for COVID, where, dim(z) $<<$ dim(input data).

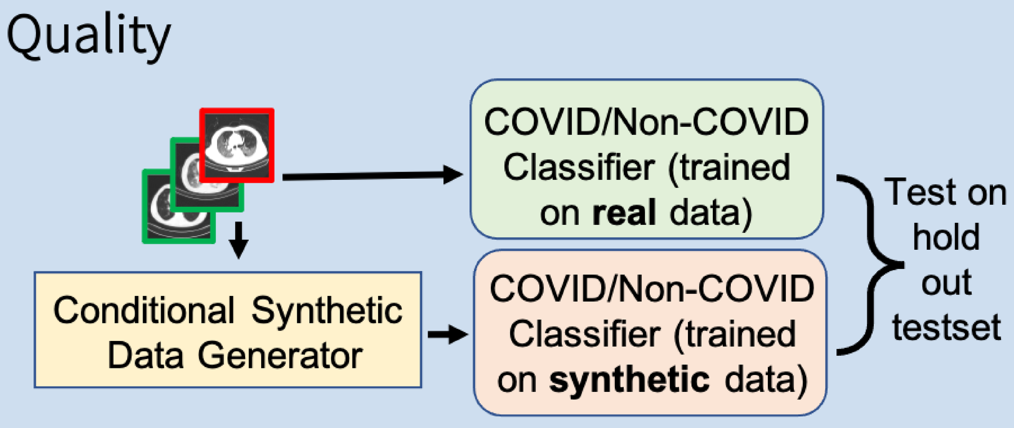
Training Step 2: Train a conditional generative flow [1] , which takes the input data x and conditional representation z, and gives local representation ν, $f_\theta(x, z) = \nu \sim \mathcal{N}(0, I)$.

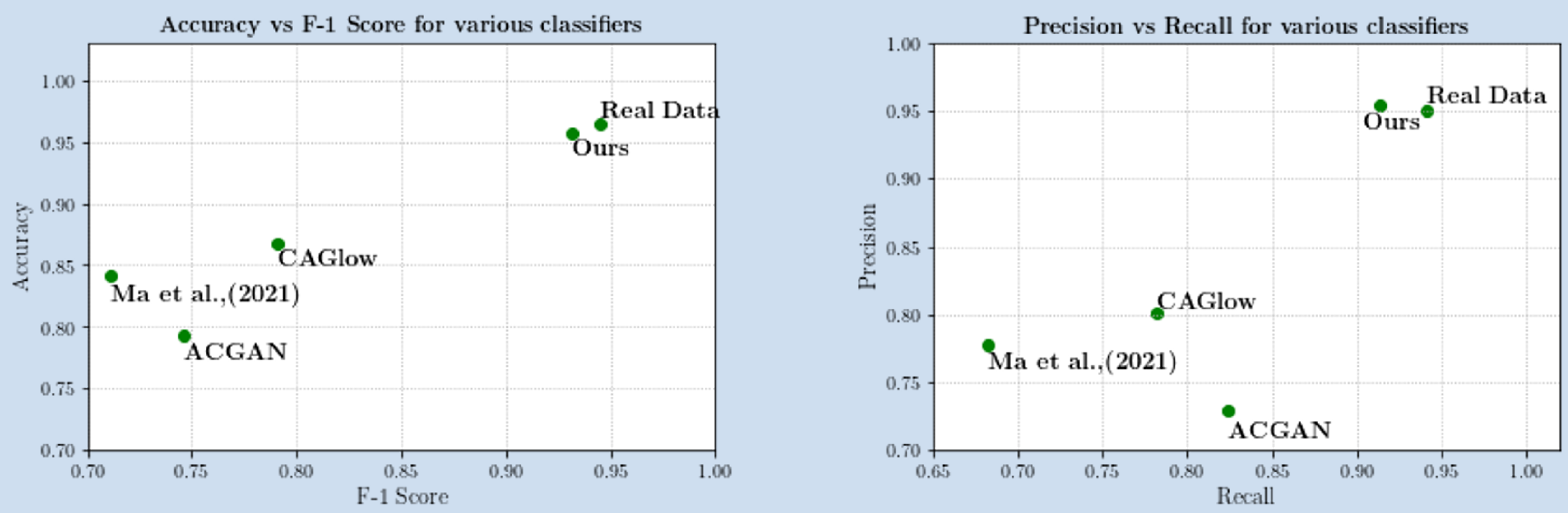
Sample Generation: Take an input sample x, keeping the conditional feature representation the same (z), sample a new local representation $\tilde{\nu}$, to generate a synthetic sample $\tilde{x} = f_\theta^{-1}(\tilde{\nu}, z)$.

Data Source and Pre-processing: We use chest CT scan dataset from [2], and combine Normal and Pneumonia classes into "Non-COVID"class ($\sim$ 80% samples) and kept the "COVID"class as it is ($\sim$ 20% samples).

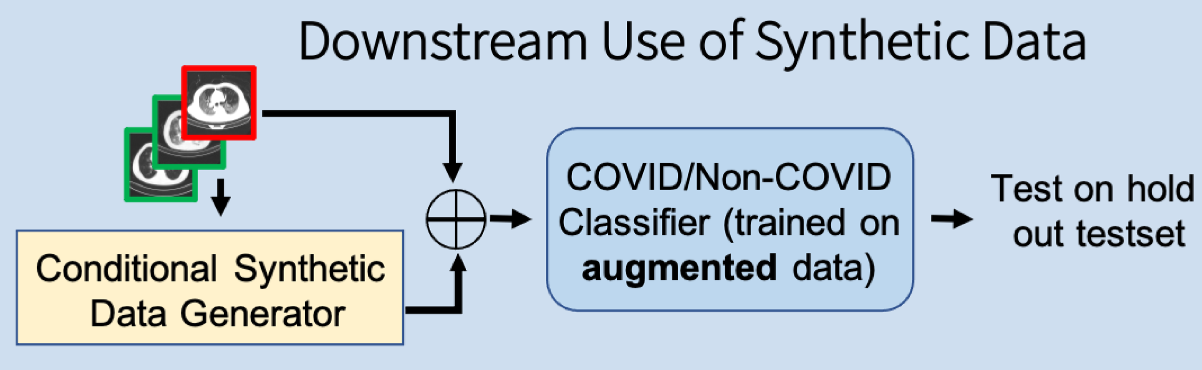
Results

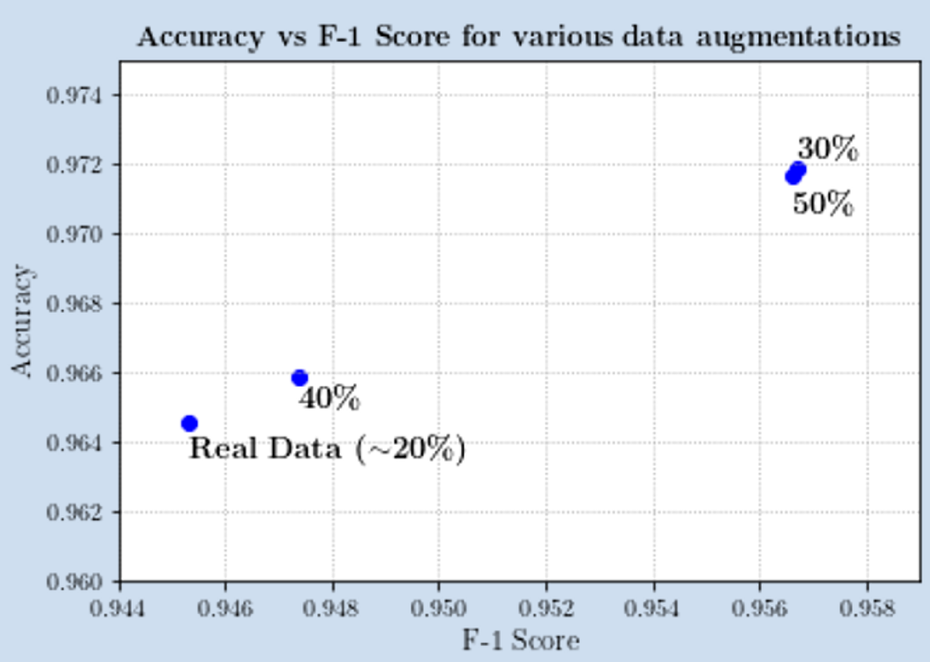
Conditional Synthetic Data Quality





Downstream Use of Synthetic Data





Acknowledgements

This research was funded in part by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore.

This work was also supported in part by the C3.ai Digital Transformation Institute and the Microsoft Corporation.

Key References

[1] Gunraj et al., "COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images." Frontiers in medicine (2020).

[2] Ma et al., "Decoupling global and local representations via invertible generative flows." In ICLR 2021.