**Prediction of Covid-19 cases & proposing faster diagnosis using lung ultrasound image recognition method.**

Our dataset contains 3 folders train, test and pred. Train folder will bwe used to train the model. Test folder will be used for validation and pred folder will be used to predict based on the models that we have trained.

**Data fitness:**

The quantity of image in the dataset, which is over 7500, is a good place to start. The values of the data show that there is sufficient information in the data to support our classification. Previously we used ultrasound lung images dataset which was not a better fit for our prediction on covid19. The current dataset is good enough for our project.

Ethical Assessment

## A. Data Collection

**A.1 Informed consent**: If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?

**Informed consent should be obtained in writing and should include a clear explanation of what the individual is agreeing to and what their rights are**.

**Our dataset includes xray images, and the dataset is extracted from Kaggle. While getting the dataset from Kaggle we have not got any personal information, hence privacy and confidentiality of the data is implemented.**

**A.2 Collection bias**: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?

**To minimize collection bias in X-ray image datasets, it's important to ensure that the dataset is diverse and representative of the population that the model will be used on. This can be done by collecting X-rays from a variety of individuals, including those with different ages, ethnicities, and medical histories. We don’t have much of these information.**

**A.3 Limit PII exposure**: Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn’t relevant for analysis?

**Due to the lack of any such information on the data sources, the dataset does not reveal any personally identifying information. During the extraction of images the privacy of data providers was respected.**

**A.4 Downstream bias mitigation**: Have we considered ways to enable testing downstream results for biased outcomes (e.g., collecting data on protected group status like race or gender)?

**Downstream bias refers to the biases that are introduced into machine learning models during the training, validation, or testing phase. This type of bias can result in models that perform poorly on certain populations or demographics, and can limit the usefulness of the models for real-world applications**.

## B. Data Storage

**B.1 Data security**: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?

**The group is aware of how crucial it is to handle the photos that were retrieved with care for the purposes of this research. The team members will ensure that only the five other members of the group have access to or utilise the data by keeping access restricted to themselves only.**

**B.2 Right to be forgotten**: Do we have a mechanism through which an individual can request their personal information be removed?

**We don’t have any personal information.**

**B.3 Data retention plan**: Is there a schedule or plan to delete the data after it is no longer needed?

**We are using the data only for project purpose. Hence, data will be removed from all our devices after our project is done.**

## C. Analysis

**C.1 Missing perspectives**: Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?

N/A

**C.2 Dataset bias**: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?

**The distribution of the classes in our picture collection, which contains 7943 photos, is as follows: COVID - 4269, Normal - 3666. The statistics are obviously out of proportion. These photographs will be used to train a machine learning model, and the more examples the model sees, the better. Our model's performance would also be affected by how much of a certain class of photos we used to train it with in comparison to other classes**.

**C.3 Honest representation**: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?

**The team has concentrated its efforts on accurately portraying the data in this data assessment in order to guide the project in the proper path rather than leading it wrong.**

**C.4 Privacy in analysis**: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?

**There is no PII information to be used throughout the course of the project.**

**C.5 Auditability**: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?

**The best possible documentation of the project's steps will be performed, including code with elaborative comments and adding notes to each method used to build the model.**

## D. Modeling

**We are currently working on modles for our dataset, to get an idea how the dataset can fit in the model. Trying to understand our dataset better.**

* **D.1 Proxy discrimination**: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
*  **D.2 Fairness across groups**: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?
*  **D.3 Metric selection**: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?
*  **D.4 Explainability**: Can we explain in understandable terms a decision the model made in cases where a justification is needed?
*  **D.5 Communicate bias**: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?