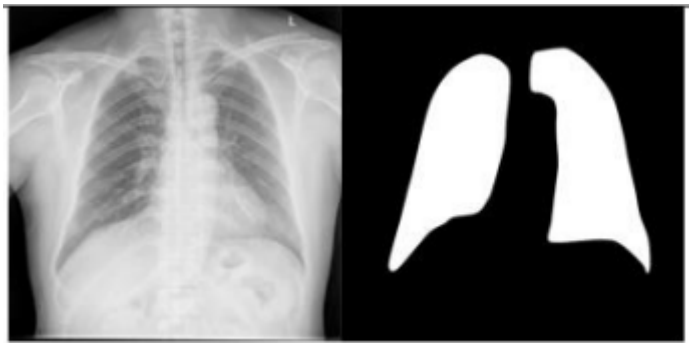
Dataset Segmentasi

Untuk dataset berasal dari repositori Kaggle Qatar University (QU) Database. Penggunaan dataset untuk proses pelatihan model segmentasi bersumber dari alamat berikut : https://www.kaggle.com/datasets/anasmohammedtahir/covidqu yang berisi citra *x-ray* dada beserta *mask* areanya, jumlah data asli dari sumber sebanyak 33,920 data citra *x-ray*. Berikut merupakan contoh dari data citra *x-ray* dan masking area pada Gambar 1.1.



Gambar 1.1. Dataset Citra Segmentasi (Kiri : *X-Ray* Dada, Kanan : *Masking* Areanya)

Dari sumber dataset awal, sebanyak 1.500 citra x-ray dada dipilih secara acak, terbagi dalam tiga kelas utama: COVID, Non-COVID (*Bacterial Pneumonia*), dan Normal. Proses pengambilan acak ini diimplementasikan melalui program Python dengan memanfaatkan beberapa *library* yang relevan. Setiap kelas memiliki data sebanyak 500 citra, dilengkapi dengan *mask* untuk memfasilitasi proses pelatihan data pada model yang akan dikembangkan. Citra dan mask masing-masing kelas disimpan dalam repositori terpisah, dirancang khusus untuk memastikan manajemen data yang efisien dan aksesibilitas selama tahap pengembangan serta evaluasi model. Informasi yang lebih rinci, termasuk distribusi data dan karakteristik kelas, dapat ditemukan pada Tabel 1.1. Tabel ini memberikan gambaran terperinci yang menjadi dasar penting untuk pengembangan model segmentasi citra x-ray dada.

Tabel 1.1. Dataset Segmentasi

Kelas	Jumlah Data Penelitian
Citra *X-Ray* dada COVID-19	500
Citra *X-Ray* dada Non-COVID-19 (Bacterial Pneumonia)	500
Citra *X-Ray* dada Normal	500
Total	1.500

Penelitian ini membagi 1500 data citra menjadi data *training* dan *testing* masing-masing perbandingannya adalah 9:1. Total data *training* yang diambil sejumlah 1350 data citra dan total data uji sejumlah 150 data citra. Data *testing* diambil per kelas secara acak dari dataset yang sudah diacak diawal (sejumlah 1350 data) yaitu 50 citra COVID-19, 50 citra Non-COVID-19 dan Normal 50 citra. Berikut rincian pembagian dataset pada Tabel 1.2.

Tabel 1.2 Pembagian dataset (Training dan Testing)

Kelas	***Training* (90%)**	***Testing* (10%)**
COVID-19	450	50
Non-COVID-19	450	50
Normal	450	50
Total	**1.350 (*random*)**	**150**

Data latih sejumlah 1350 diacak/*random* agar posisi atau indeks kelas tidak pada tempat yang sama selama proses *cross validation*. *K-Fold Cross Validation* digunakan dalam memvalidasi proses pelatihan data. Hal ini dilakukan untuk mengetahui terjadinya *overfitting* pada dataset yang digunakan saat pelatihan. Jumlah *k* yang digunakan pada penelitian ini adalah 5. Oleh sebab itu, data latih akan dibagi lagi menjadi 8:2 yaitu masing-masing data latih sejumlah 1080 data citra dan validasi sejumlah 270 data citra. Data latih, uji dan validasi (dengan *K-Fold Cross Validation*) akan digunakan pada proses pelatihan dan pengujian

model 2D V-Net. Berikut rincian pembagian data *Training, Validation, Testing* pada Tabel 3.3.

Tabel 1.3. Pembagian data *Training, Validation, Testing*

Proses *K-Fold Cross Validation*			
***K-Fold* (k=5)**	***Training***	***Validation***	***Testing***
Fold ke 1-5	1080	270	150
Total	**1500**		