# **BAB IV HASIL PENELITIAN DAN PEMBAHASAN**

## **Pengambilan Data**

Dalam penelitian ini, dataset URL diambil dari Kaggle [] dan ditambah dengan data white list dari Phising Database milik mitchellkrogza []. Data utama adalah data dari Kaggle karena data tersebut sudah dikurasi dari beberapa sumber data kolektif dari Perusahaan (Canadian Institute for Cybersecurity, PishTank, PishStrom), sedangkan dataset dari mitchellkrogza divalidasi oleh tim mereka, sehingga data dari Kaggle dirasa lebih reliable untuk digunakan sebagai data inti. Berikut adalah data dari Kaggle dan whitelist dari mitchellkrogza:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## **Preprocessing**

## Balancing

A blue rectangular object with white text

Description automatically generated

Tahapan preprocessing pertama yang diberlakukan pada data adalah balancing data, dikarena data yang didapat tidak seimbang maka dilakukan penyeimbangan data mengunakan metode under-sampling dengan mengambil sampel sebanyak data paling sedikit yaitu sebanyak 23645 data tiap kelas, sebelum diberlakukan under-sampling dilakukan drop data duplikat terlebih dahulu.

## Normalisasi

Tahapan normalisasi dilakukan dengan menjadikan semua huruf ke huruf kecil dan menghilangkan symbol symbol.