Nama: Jalaluddin Al Mursyidy Fadhlurrahman
NIM: 23521059

Tugas Besar 2 – Artificial Neural Network

1. Deskripsi dataset:

Data yang digunakan merupakan kumpulan rekaman pengucapan 1 kata dalam bahasa Inggris yang diperoleh dari TensorFlow (*Speech Commands dataset*). Dataset ini memiliki 105.829 pengucapan dari 2.618 orang yang mengucapkan 35 kata. Setiap pengucapan selama 1 detik disimpan sebagai file berformat WAV dengan *sampling rate* 16 kHz. Data *training* berjumlah 85.511 data, dan data *test* berjumlah 4.890 data. Pada studi kasus ini, saya menggunakan dataset Mini Speech Commands, yang merupakan bagian dari dataset Speech Commands. Mini Speech Commands memiliki pengucapan 8 kata yaitu `'down'`, `'yes'`, `'go'`, `'no'`, `'up'`, `'right'`, `'stop'`, dan `'left'`. Data input berjumlah 8.000 data yang terbagi menjadi 6.400 data training, dan 800 data *test*.

Ekstraksi fitur dilakukan dengan mengubah audio ke domain frekuensi-waktu dengan menerapkan Short-Time Fourier Transform (tidak menggunakan Fast Fourier Transform) agar tidak kehilangan informasi waktunya. Informasi waktu pada *speech recognition* berguna untuk memberikan informasi urutan waktu tiap-tiap kata diucapkan. Data pada domain frekuensi-waktu kemudian divisualisasikan ke dalam *Spectrogram*.

Target dari studi kasus ini adalah 8 *class* kata.

2. Arsitektur ANN yang digunakan pada penelitian lain:
    a. Aaron Nichie dan Godfrey A. Mills (2013) dalam penelitiannya yang berjudul "Voice Recognition Using Artificial Neural Networks and Gaussian Mixture Models" menggunakan arsitektur *feed-forward* ANN dengan *supervised training* dari vektor fitur MFCC yang diekstrasi dari pengucapan. ANN ditraining dengan menggunakan algoritma *back propagation* yang, menurut penelitian ini, efektif dalam meminimalisir *recognition error rate*. Peneliti menggunakan *single input* dan *hidden layer neurons* dalam sebuah sistem-3-layer, serta memvariasikan jumlah neuron pada tiap-tiap layer dengan mentraining *network* sampai mencapai jumlah optimal yang memberikan *training* terbaik. Peneliti kemudian mendapatkan bahwa arsitektur ANN dengan 20 neuron pada *input layer* dan sebuah *hidden layer* dengan 30 neuron adalah yang paling cocok.
    b. Gouda, Sanjay Krisna, et. Al. (2020) dalam penelitiannya yang berjudul "Speech Recognition: Key Word Spotting through Image Recognition" menggunakan *spectrogram* sebagai hasil ekstraksi fitur suara. Peneliti kemudian menawarkan arsitektur berupa: *First Convolutional Layer*, *Second Convolutional Layer*, *Densely Connected Layer*, dan *Softmax Output Layer*. *First Convolutional Layer* akan menerima 32 neuron. Peneliti juga

menggunakan dataset yang sama dengan master dataset yang saya pakai, yakni dataset Speech Commands. Penelitian ini mendemonstrasikan pemecahan masalah *audio recognition* melalui pendekatan klasifikasi *image* yang sudah banyak dipelajari. Peneliti melihat betapa krusialnya *hyperparameter tuning* yang baik terhadap akurasi model.

3. Rancangan arsitektur yang akan digunakan:

Jumlah layer	9 layer yang terdiri dari: • 1 Input layer • 7 Hidden layer: (2 layer Conv2D, 1 layer Max Pooling, 2 layer Dropout, 1 layer Flatten, dan 2 layer Dense) • 1 Output layer
Input layer	Gambar *spectrogram* dengan dimensi 30 x 30 x 1
Hidden layer	Layer Convolusi 2D, Max Pooling Dropout, Flatten, dan Dense
Output layer	• Jenis layer: fully connected • Jumlah neuron: 8 • Fungsi aktivasi: ReLu

4. Jumlah parameter dari setiap layer:

Layer	**Output Shape**	**Parameter**
Conv 2D	(None, 28, 28, 32)	320
Conv 2D	(None, 26, 26, 64)	18496
MaxPooling	(None, 13, 13, 64)	0
Dropout	(None, 13, 13, 64)	0
Flatten	(None, 10816)	0
Dense	(None, 128)	1384576
Dropout	(None, 128)	0
Dense	(None, 8)	1032