

## BAGIAN A

→ The final accuracy comparison result will be presented in this table form:

Model	DNA Dataset	Fashion MNIST Dataset
PCA + CNN	57.14%	90.05%
CNN	57.14%	91.10%

## BAGIAN B

*Pertanyaan:* Apakah satu model secara konsisten mengungguli model lainnya di semua kumpulan data? Jika tidak, faktor apa yang mempengaruhi disparitas performa model pada kedua dataset tersebut?

→ Jawabannya ialah **tidak**. Dalam percobaan yang saya lakukan, hasil akurasi yang didapatkan dengan kedua model (CNN dan PCA + CNN) hanya memiliki selisih  $\pm 1\%$  (pada Fashion MNIST Dataset).

Dari hasil percobaan ini, saya melihat bahwa faktor yang memengaruhi performa model ialah jumlah data pada dataset, jumlah hidden layer pada model CNN, ukuran batch yang digunakan, dan jumlah komponen/fitur/dimensi pada dataset.

- **Jumlah data pada dataset:** Semakin banyak data pada dataset, semakin banyak pula data yang dapat digunakan oleh model untuk belajar pada proses training.

→ Hal ini dapat dilihat dari hasil akurasi model ketika menggunakan DNA dataset dengan Fashion MNIST dataset. Fashion MNIST dataset memiliki 60.000 data yang digunakan untuk training, sehingga model akan belajar dan menghasilkan bobot-bobot yang lebih akurat dibandingkan DNA dataset yang hanya memiliki 89 data secara keseluruhan, yang mana pada kasus ini saya membagi data dengan ratio 50:7 sehingga training set memiliki data sejumlah 76. Hal ini mengakibatkan nilai akurasi yang didapatkan oleh model pada DNA dataset bukan hanya rendah, tetapi juga stagnan (kemungkinan terjadi *overfitting* karena **jumlah data** yang **sedikit**).

- **Jumlah hidden layer pada model CNN:** Model menjadi lebih adaptif.

→ Dengan beberapa hidden layer, model akan lebih adaptif karena di setiap hidden layer yang tersedia akan terdapat hidden neurons yang membantu model untuk mempelajari data-data dengan lebih spesifik. Namun, jumlah hidden layer dan hidden neurons yang digunakan juga harus dipikirkan dengan baik agar performa model terjaga dan tidak ‘belajar’ terlalu keras sehingga terjadi *overfitting* yang mengakibatkan sifat generalisasi pada model berkurang (hasil klasifikasi menjadi lebih buruk).

- **Ukuran batch yang digunakan.**

→ Ukuran batch memengaruhi performa model dari segi *timing* dan *regularization*. Hal ini membantu mengurangi hasil estimasi error pada gradient ketika melakukan proses training, sehingga juga turut membantu mengontrol nilai akurasi. Ukuran batch yang besar dapat mempercepat komputasi training di tiap epochnya, namun jika terlalu besar dapat menurunkan performa model.

- **Jumlah komponen/*dimensional space* pada dataset.**

→ Jumlah komponen yang digunakan berpengaruh pada tingkat kecepatan komputasi, namun tidak menjamin keakuratan/tingginya hasil akurasi. Hal ini dibuktikan dengan metode PCA, yang mana waktu yang dibutuhkan selama 1 epoch jauh lebih cepat dibandingkan model dengan tanpa PCA. Hal ini dapat jauh lebih terlihat dengan jumlah data yang besar.

```
- 45s 22ms/step
- 0s 5ms/step -
- 0s 5ms/step -
- 0s 4ms/step -
- 0s 4ms/step -
```

(ETA tanpa PCA)

```
- 1s 14ms/step -
- 0s 5ms/step -
- 0s 5ms/step -
- 0s 5ms/step -
- 0s 5ms/step -
```

(ETA dengan PCA)

Namun, dengan penurunan jumlah komponen, akurasi akan cenderung lebih rendah karena memungkinkan adanya komponen yang tidak terlalu mewakili komponen lain yang telah direduksi/dihilangkan (hal ini dapat dilihat dari persenan variance ratio yang didapat setelah melakukan PCA). Oleh karena itu, nilai *n\_components*/jumlah komponen turut memengaruhi tingkat keakuratan prediksi data.