

Learning Progress Review

Week-16

OPTIMISTIC

DigitalSkola



ALDIVA

LUTFIA

ASPRIZAL

MILLENTIA

GILANG





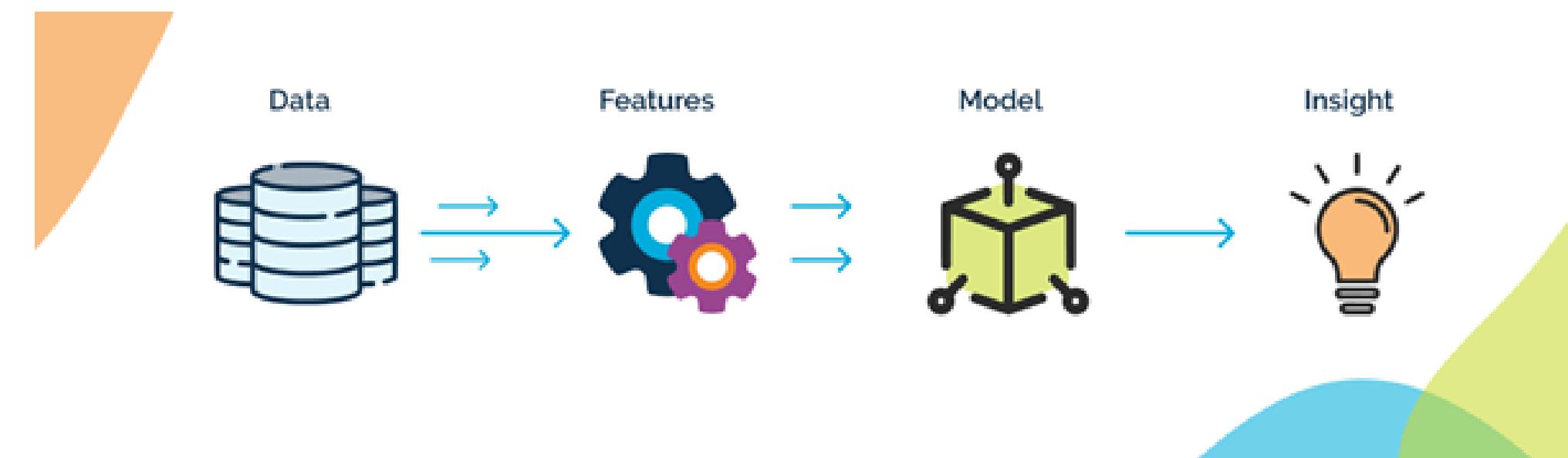
SESSION 46

ADVANCED DATA PREPROCESSING FOR TEXT



FEATURE ENGINEERING

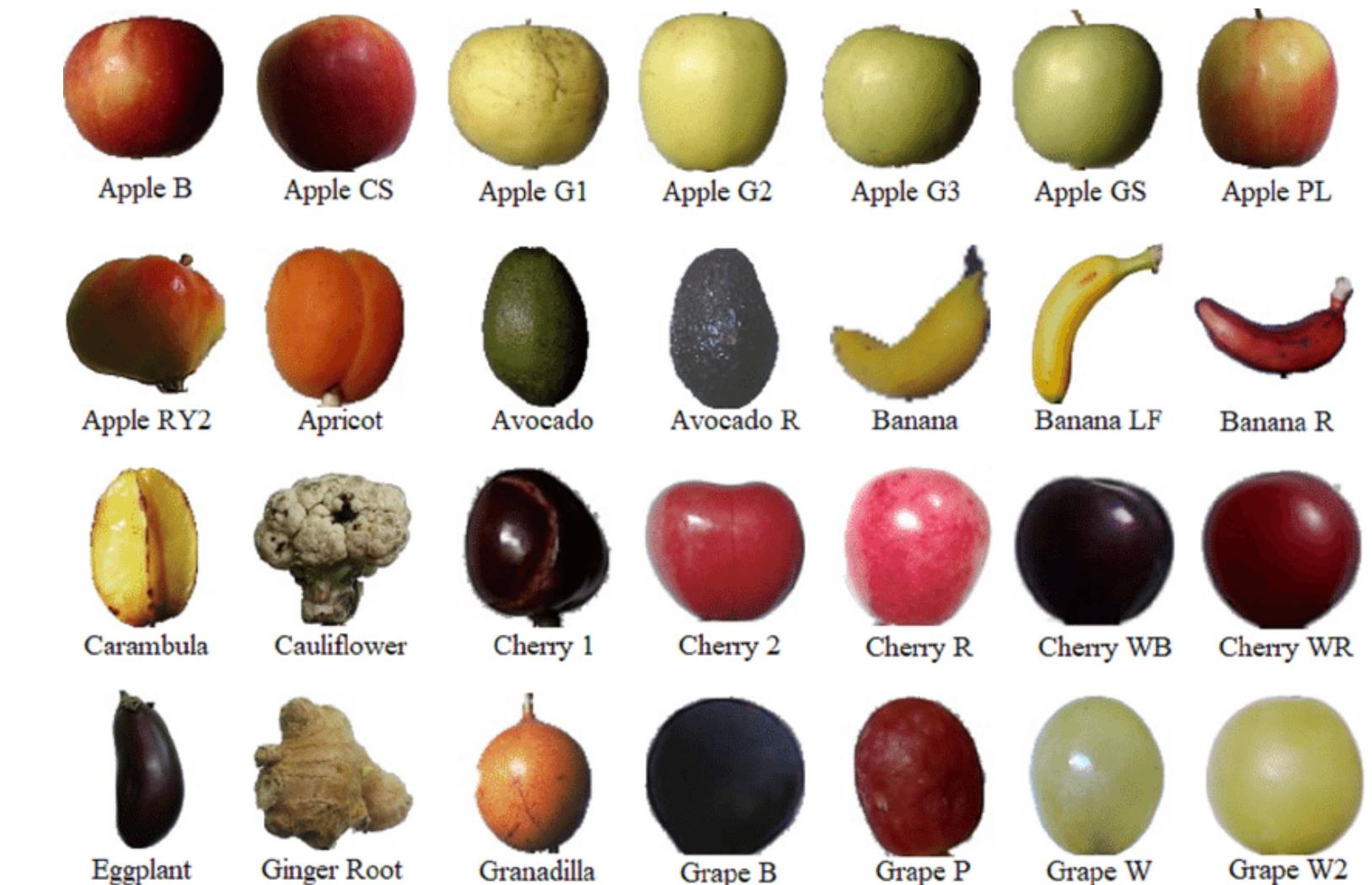
Feature engineering atau Rekayasa fitur adalah proses menggunakan pengetahuan domain untuk mengekstrak fitur (karakteristik, properti, atribut) dari data mentah atau mengacu pada proses penggunaan pengetahuan domain untuk memilih dan mengubah variabel yang paling relevan dari data mentah saat membuat model prediktif menggunakan machine learning. Rekayasa fitur telah banyak digunakan dalam kompetisi Kaggle dan machine learning projects.



DATASET TYPES



Wild Anim Dataset



Fruit dataset



TEXT CLASSIFICATION-DEFINITION

Klasifikasi Teks adalah penugasan dokumen teks ke satu atau lebih kategori yang telah ditentukan berdasarkan isinya. Dokumen teks pengklasifikasi Kelas A Dokumen teks Kelas B Dokumen teks Kelas C.

Pengklasifikasi:

Input : satu set m dokumen berlabel tangan $(x_1, y_1), \dots, (x_m, y_m)$

Output : classifier yang dipelajari $f: x \rightarrow y$



TEXT CLASSIFICATION-APPLICATIONS

- Klasifikasikan berita sebagai World, US, Business, Sci, Teknologi, Olahraga, Hiburan, Kesehatan, Lainnya.
- Mengklasifikasikan nama bisnis berdasarkan industri.
- Klasifikasikan esai siswa sebagai A, B, C, D, atau F.
- Klasifikasikan email sebagai Spam, Lainnya.
- Klasifikasikan email ke staf teknologi sebagai Mac, Windows, . . . , Lainnya.
- Klasifikasikan file pdf sebagai riset. kertas, Lainnya
- Klasifikasikan ulasan film sebagai Menguntungkan, Tidak Menguntungkan, Netral.



HANDLING TEXT DATASET

- TOKENIZATION

Tokenisasi adalah tugas umum dalam Natural Language Processing (NLP). Ini adalah langkah mendasar dalam metode NLP tradisional seperti Count Vectorizer dan arsitektur berbasis Advanced Deep Learning seperti Transformers. Tokenisasi adalah cara memisahkan sepotong teks menjadi unit yang lebih kecil yang disebut token. Di sini, token dapat berupa kata, karakter, atau subkata. Oleh karena itu, tokenisasi dapat secara luas diklasifikasikan menjadi 3 jenis – tokenisasi kata, karakter, dan subkata (karakter n-gram).



Pre-processing the Text

1. Menghapus stop words

- Punctuations
- Prepositions

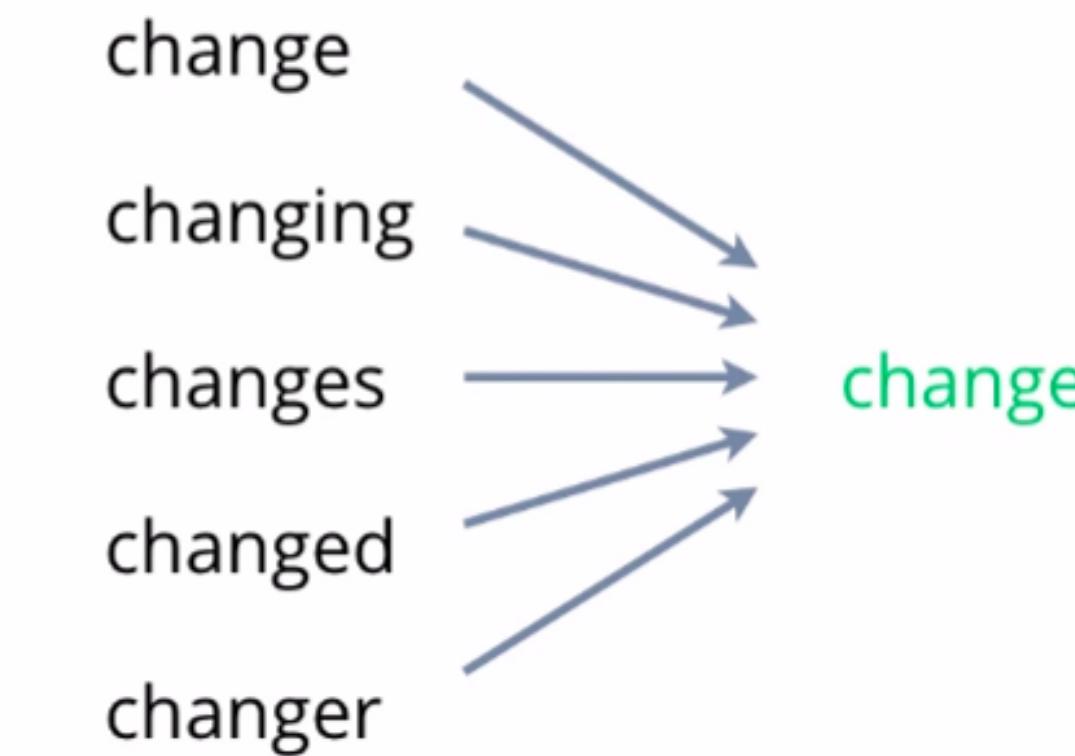
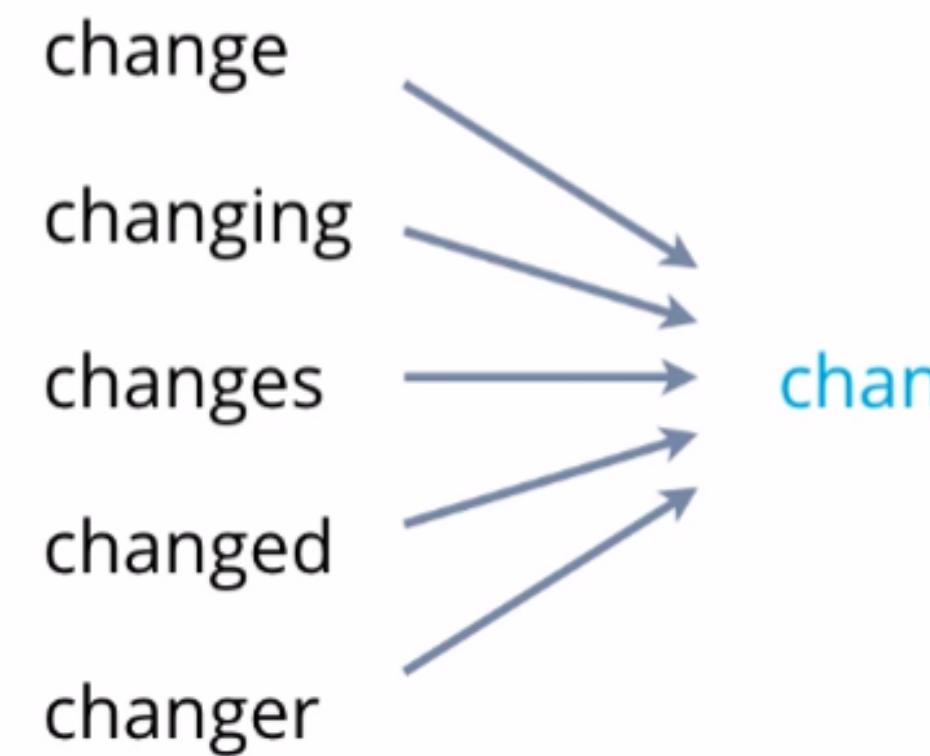
2. Stemming

- Stemming adalah proses mereduksi kata-kata yang diinfleksikan yang diturunkan menjadi kata dasar.

3. Lemmatization

- Lemmatization adalah proses merubah kata menjadi kata dasar. Bedanya dengan stemming, kata dasar hasil lemmatization lebih mudah dipahami.

Stemming vs Lemmatization



Feature Extraction

Bag Of Words

- Bag of Words atau yang biasa disebut sebagai BOW merupakan representasi dari kata-kata menjadi vektor angka 0 (nol) dan 1 (satu). Kata-kata tersebut dapat berasal dari berbagai kalimat dan paragraf.

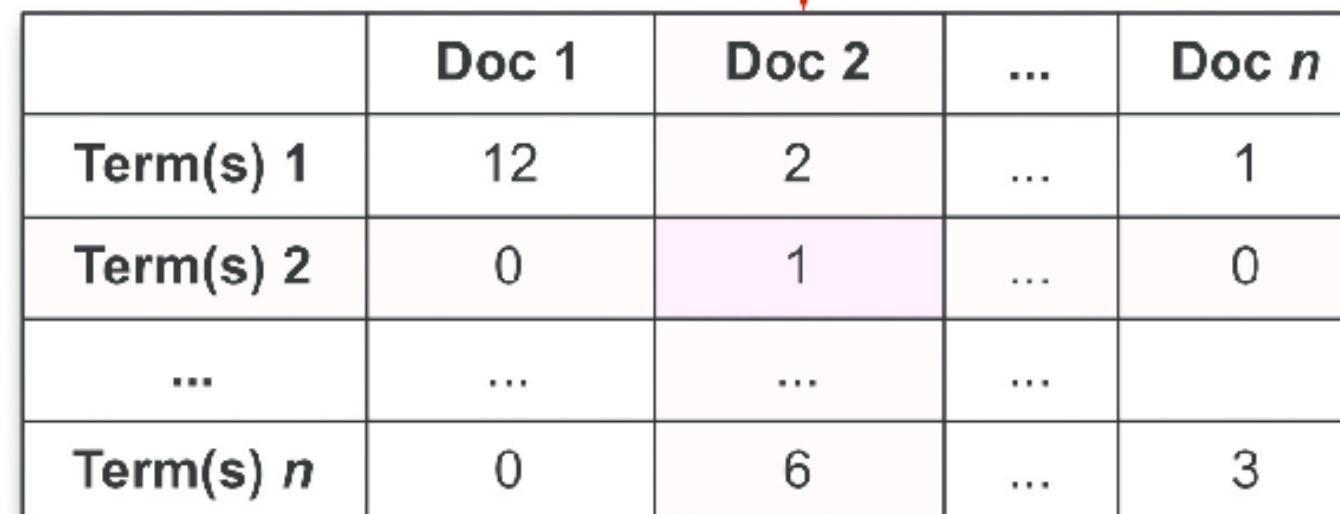
	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Feature Extraction

Term Frequency-Inverse Document Frequency

- TF-IDF adalah suatu metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$



A red arrow points from the formula $\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$ down to the matrix. A red curly brace groups the columns "Doc 1", "Doc 2", ..., "Doc n".

	Doc 1	Doc 2	...	Doc n
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...
Term(s) n	0	6	...	3



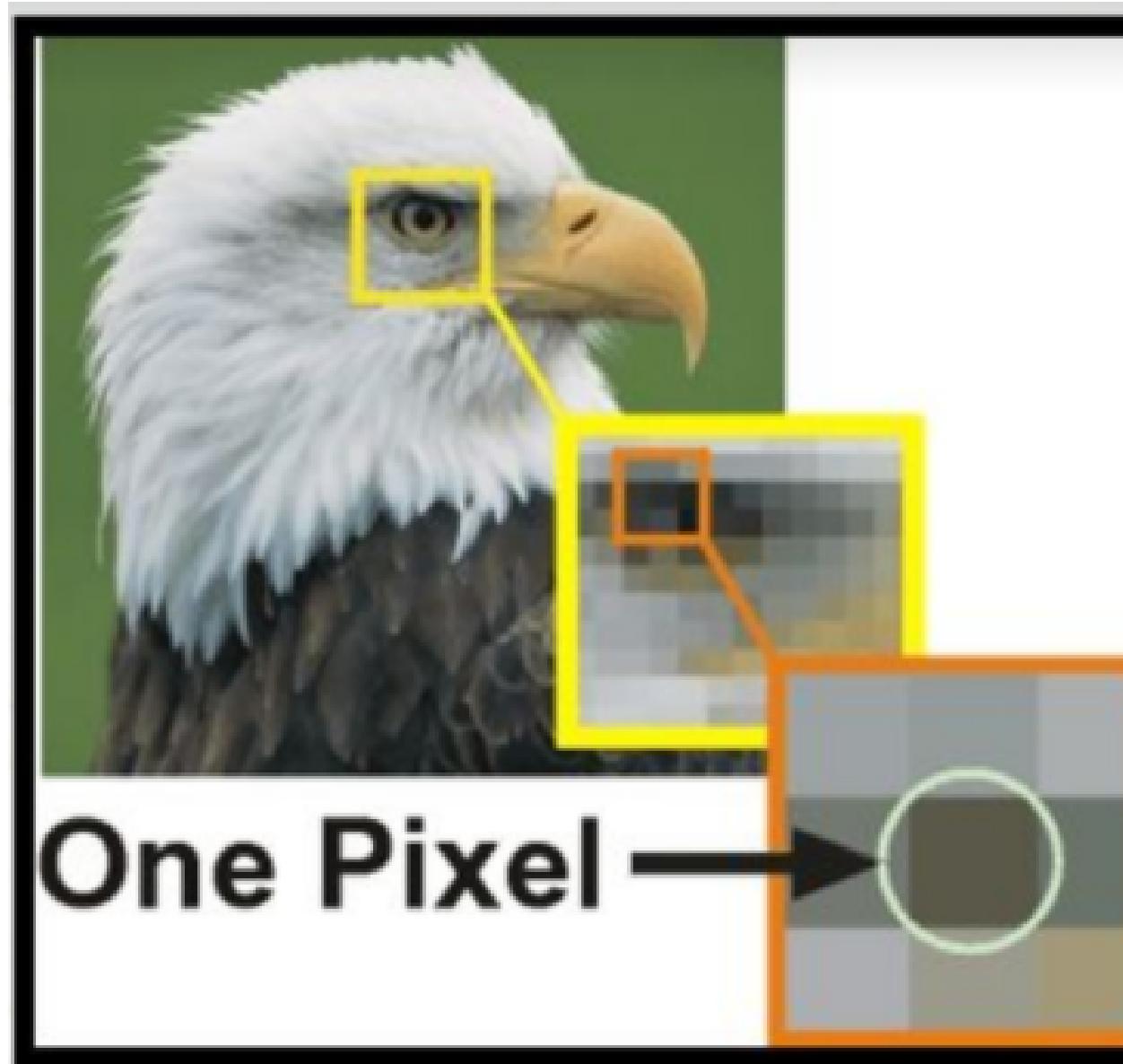
SESSION 47

ADVANCED DATA

PREPROCESSING

FOR IMAGES

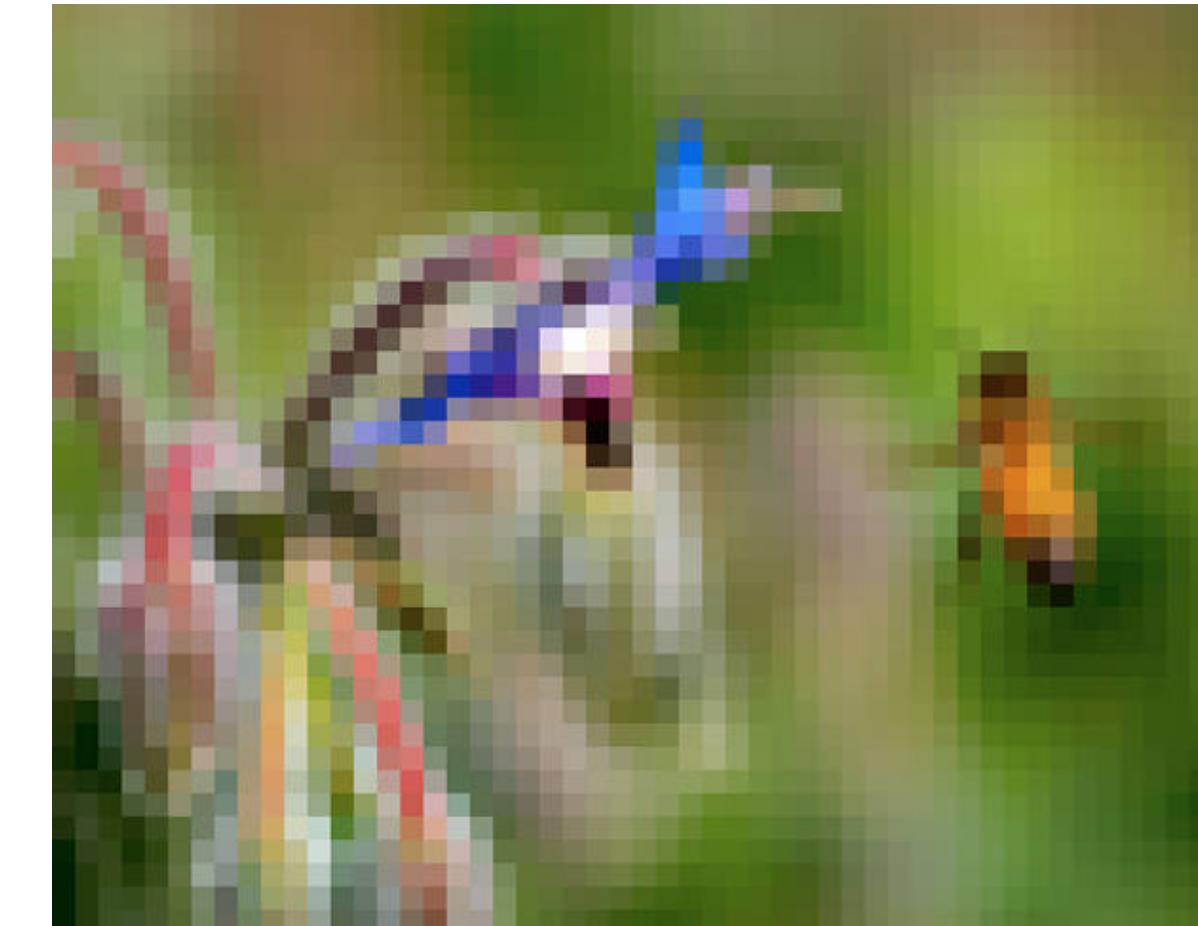
IMAGE DATA



Setiap foto, dalam bentuk digital, terdiri dari pixel. Mereka adalah unit informasi terkecil yang membentuk sebuah gambar. Biasanya bulat atau persegi, mereka biasanya diatur dalam kotak 2 dimensi.

PIXEL

Kata pixel berarti elemen gambar. Cara sederhana untuk menggambarkan setiap piksel adalah dengan menggunakan kombinasi tiga warna, yaitu Merah, Hijau, Biru. Inilah yang kami sebut gambar RGB.



PIXEL IN PYTHON

Di dalam python untuk menampilkan sebuah warna diperlukan sebuah matrix yang terdiri dari kumpulan angka dari sebuah warna.

Oleh karena itu pada satu gambar RGB kita memerlukan 3 matrix yang mampu mewakili setiap warna.

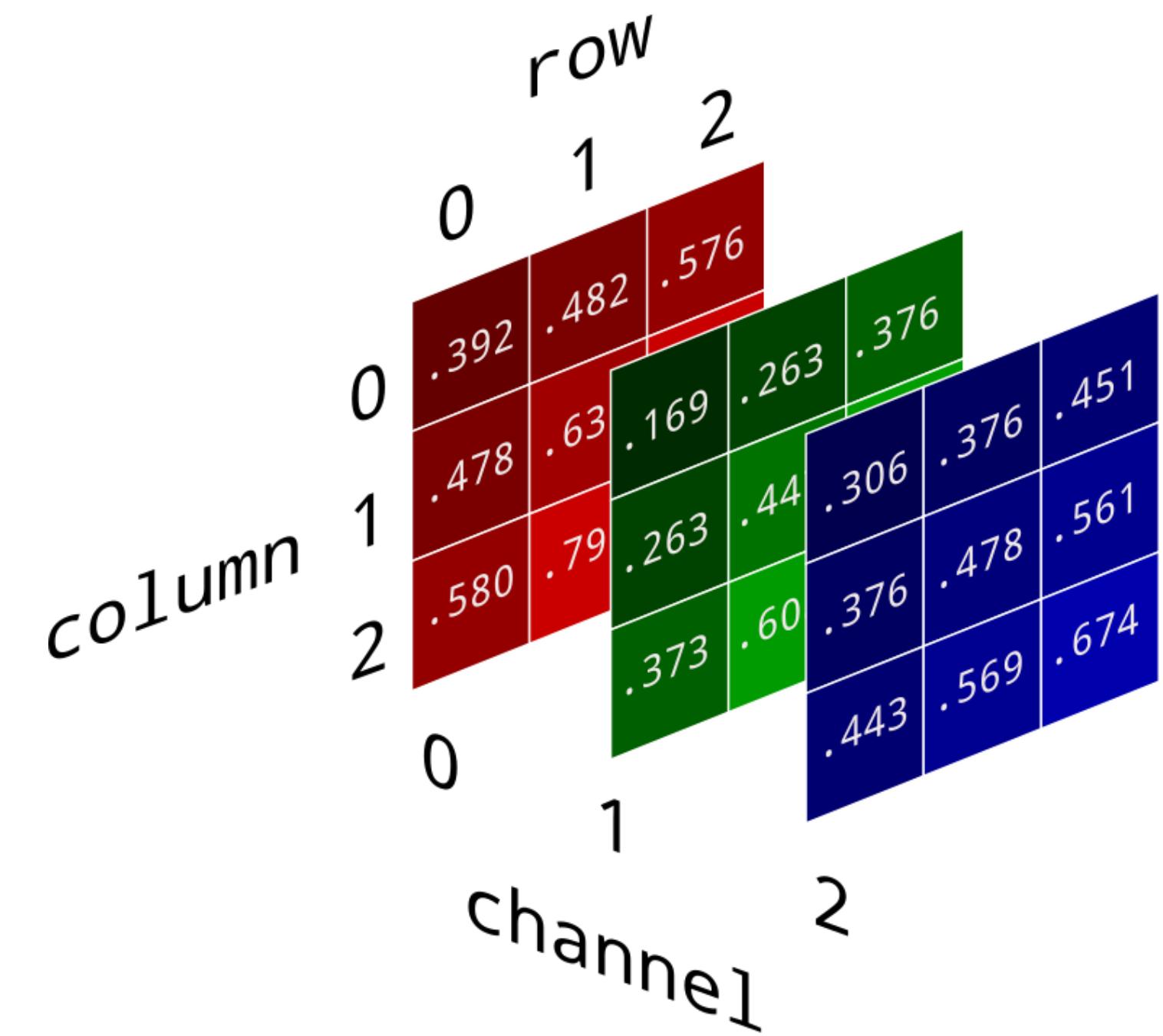
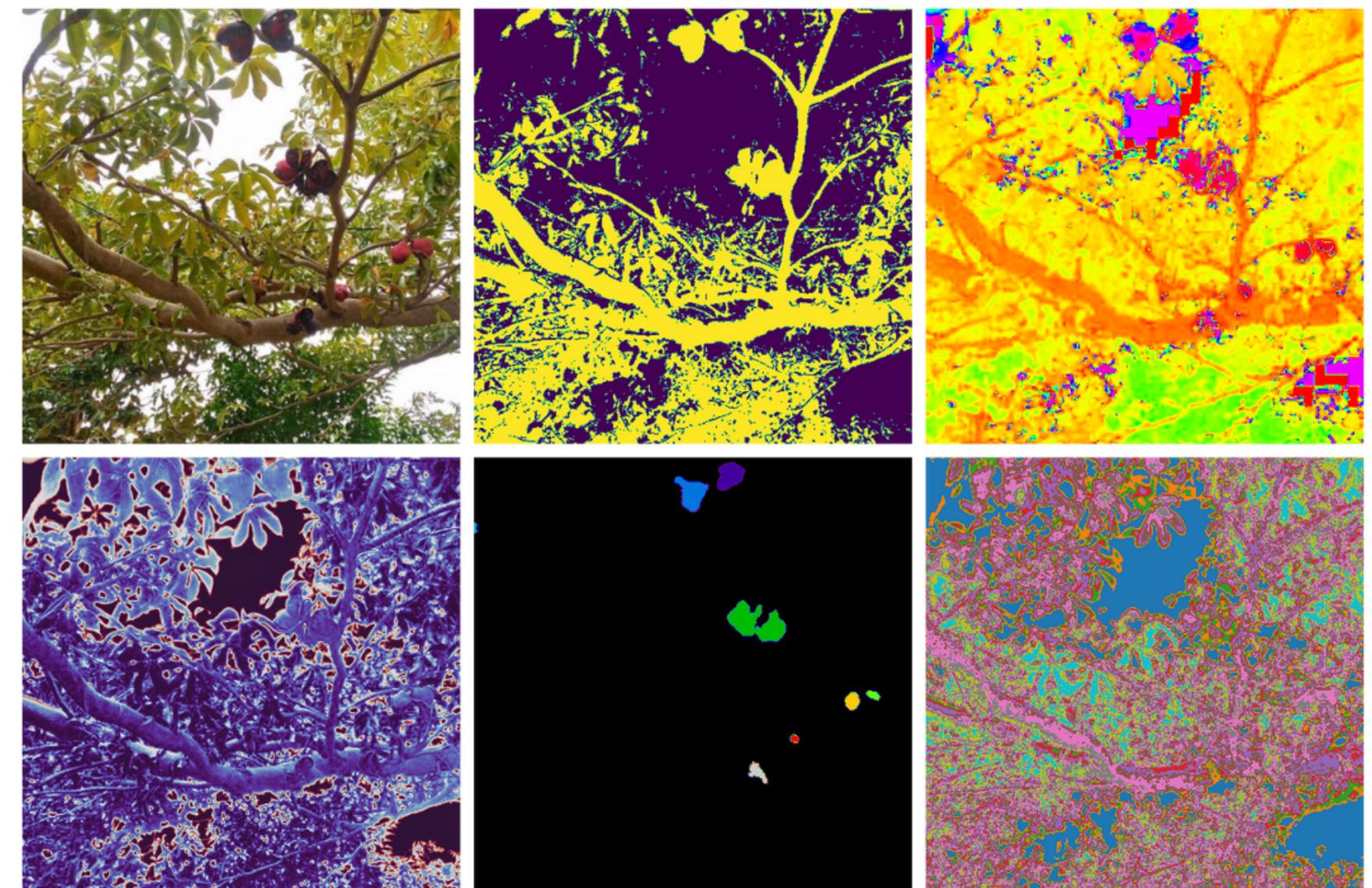


IMAGE IN PYTHON

Gambar didalam python dapat dipresentasikan di berbagai warna yaitu:

1. RGB (Red, Green and Blue)
2. Grey
3. HSV (Hue, Saturation and Value)



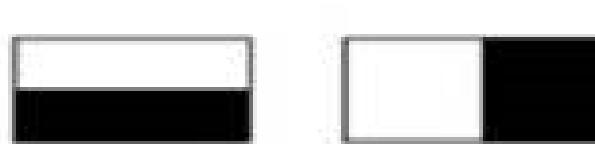
Feature Extraction

Haar Like Feature

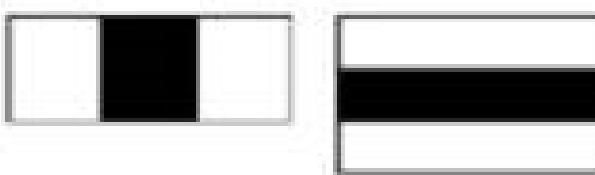
- Adanya fitur Haar ditentukan dengan cara mengurangi rata-rata piksel pada daerah gelap dari rata-rata piksel pada daerah terang. Jika nilai perbedaannya itu diatas nilai ambang atau threshold, maka dapat dikatakan bahwa fitur tersebut ada. Nilai dari Haar-like feature adalah perbedaan antara jumlah nilai-nilai piksel gray level dalam daerah kotak hitam dan daerah kotak putih.
- Haar-like feature memproses gambar dalam kotak-kotak, dimana dalam satu kotak terdapat beberapa pixel. Per kotak itu pun kemudian di-proses dan didapatkan perbedaan nilai (threshold) yang menandakan daerah gelap dan terang. Nilai – nilai inilah yang nantinya dijadikan dasar dalam image processing.

Feature Extraction

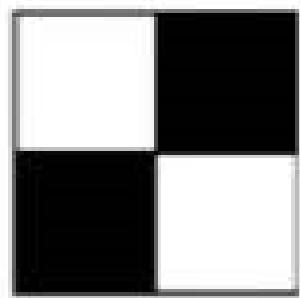
Haar Like Feature



Edge Features

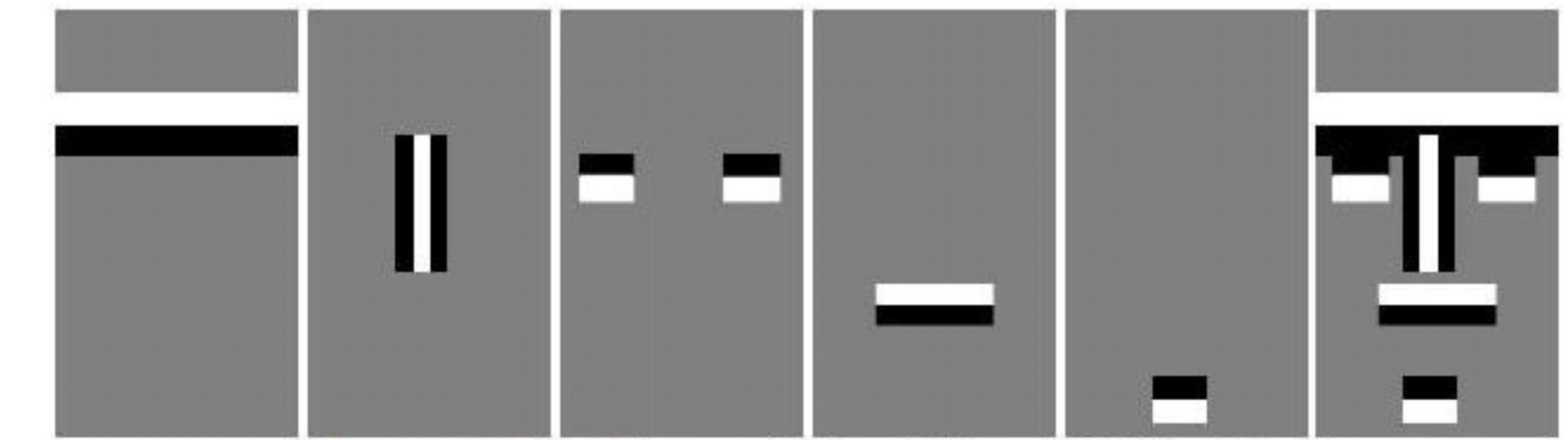


Line Features



Four-rectangle Features

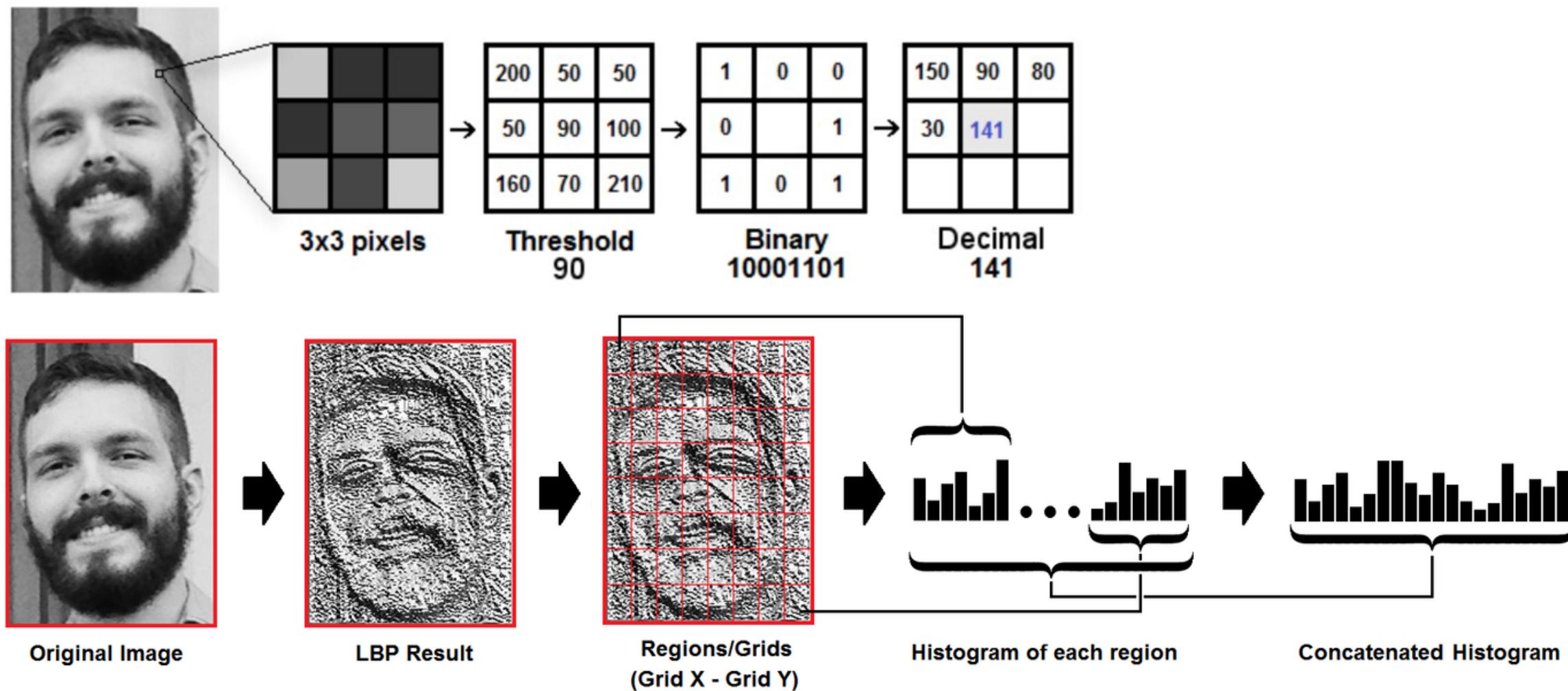
Important Features for Face Detection

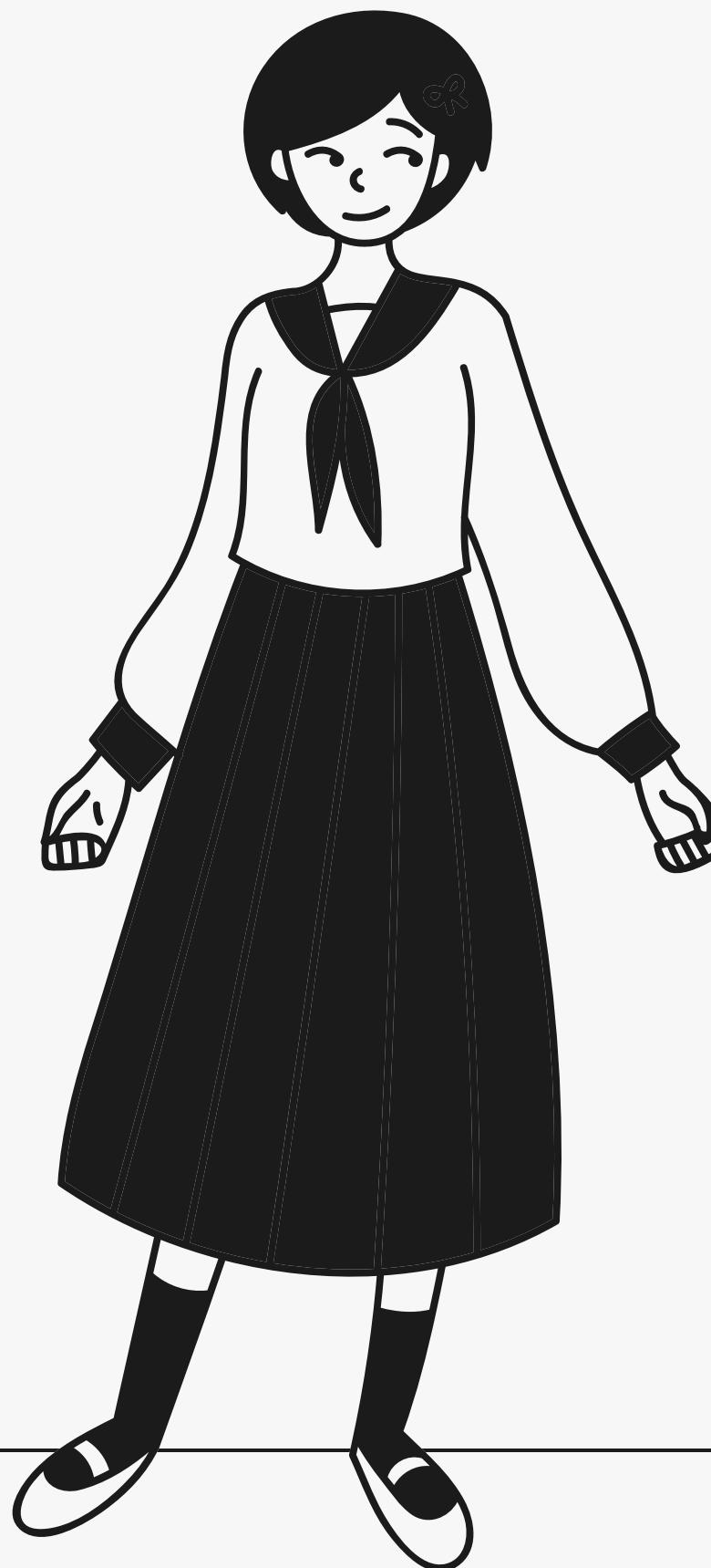


Feature Extraction

Local Binary Pattern(LBP)

Algoritma LBP (Local Binary Pattern) adalah salah satu algoritma yang dapat digunakan untuk melakukan klasifikasi berdasarkan tekstur gambar. LBP Dapat digunakan apabila gambar tersebut berwarna abu-abu. Prinsipnya adalah dengan mencari pola dari sekitar pixel yang ada.





SESSION 48

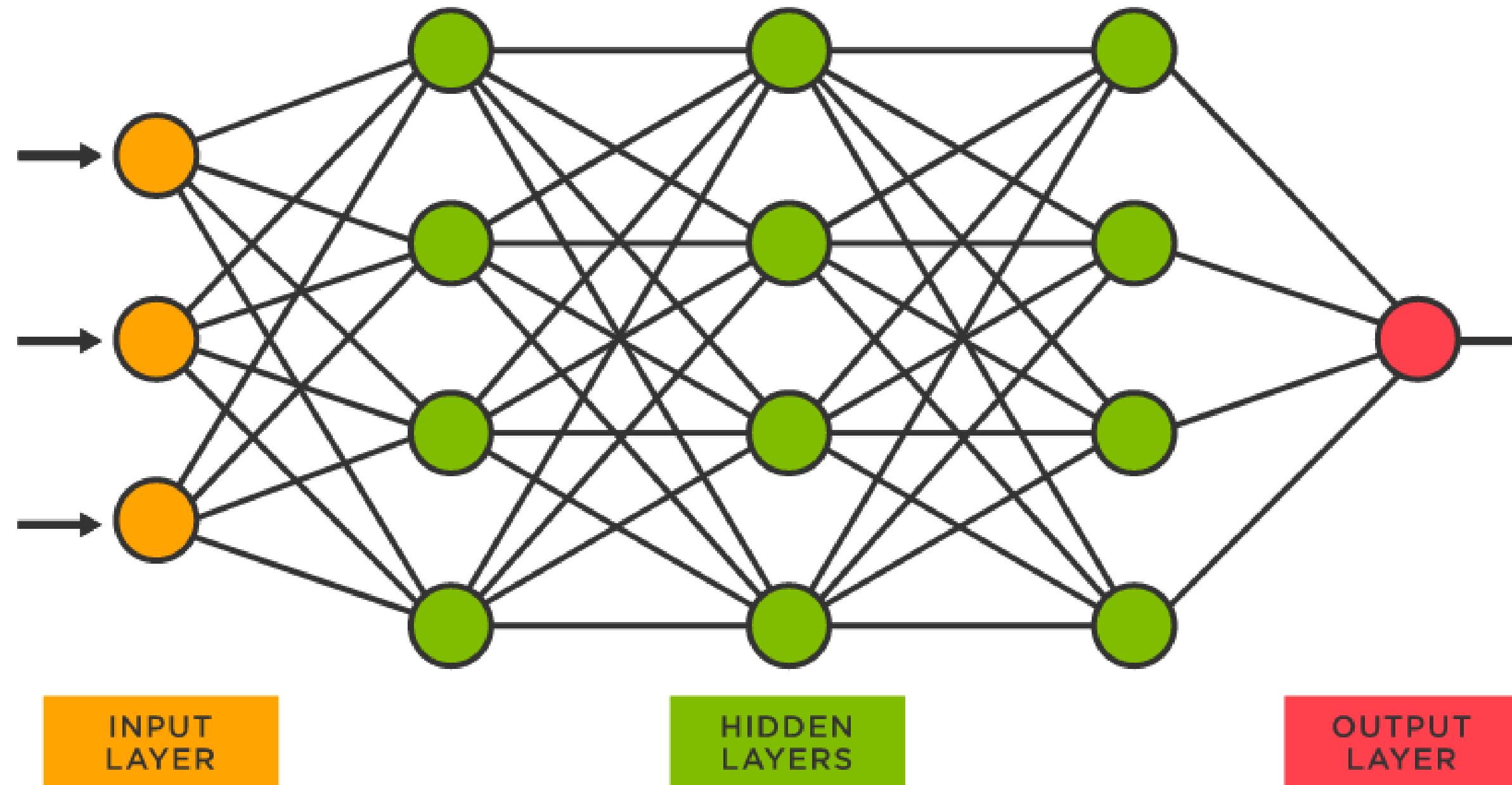
Neural Network



Definisi

- Neural Network (NN) adalah suatu metode pembelajaran yang diinspirasi dari jaringan sistem pembelajaran biologis yang terjadi dari jaringan sel syaraf (neuron) yang terhubung satu dengan yang lainnya.
- Neural network juga disebut dengan Artificial Neural Network.
- Artificial Neural Network (ANN) atau Jaringan Syaraf Tiruan merupakan sebuah teknik atau pendekatan pengolahan informasi yang terinspirasi oleh cara kerja sistem saraf biologis, khususnya pada sel otak manusia dalam memproses informasi.
- Elemen kunci dari teknik ini adalah struktur sistem pengolahan informasi yang bersifat unik dan beragam untuk tiap aplikasi. Neural Network terdiri dari sejumlah besar elemen pemrosesan informasi (neuron) yang saling terhubung dan bekerja bersama-sama untuk menyelesaikan sebuah masalah tertentu, yang pada umumnya adalah masalah klasifikasi ataupun prediksi.

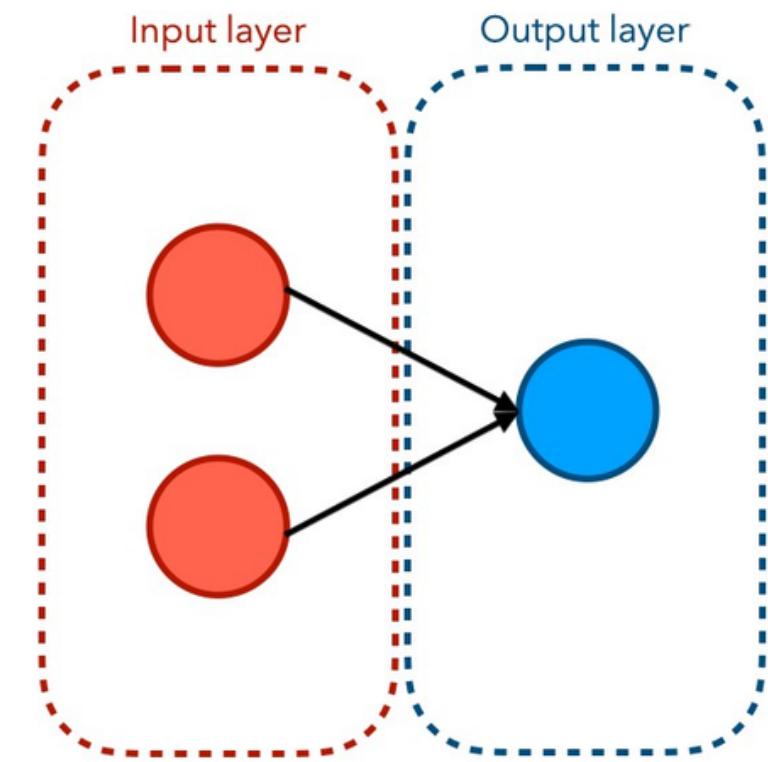




PEMODELAN JARINGAN PADA ANN

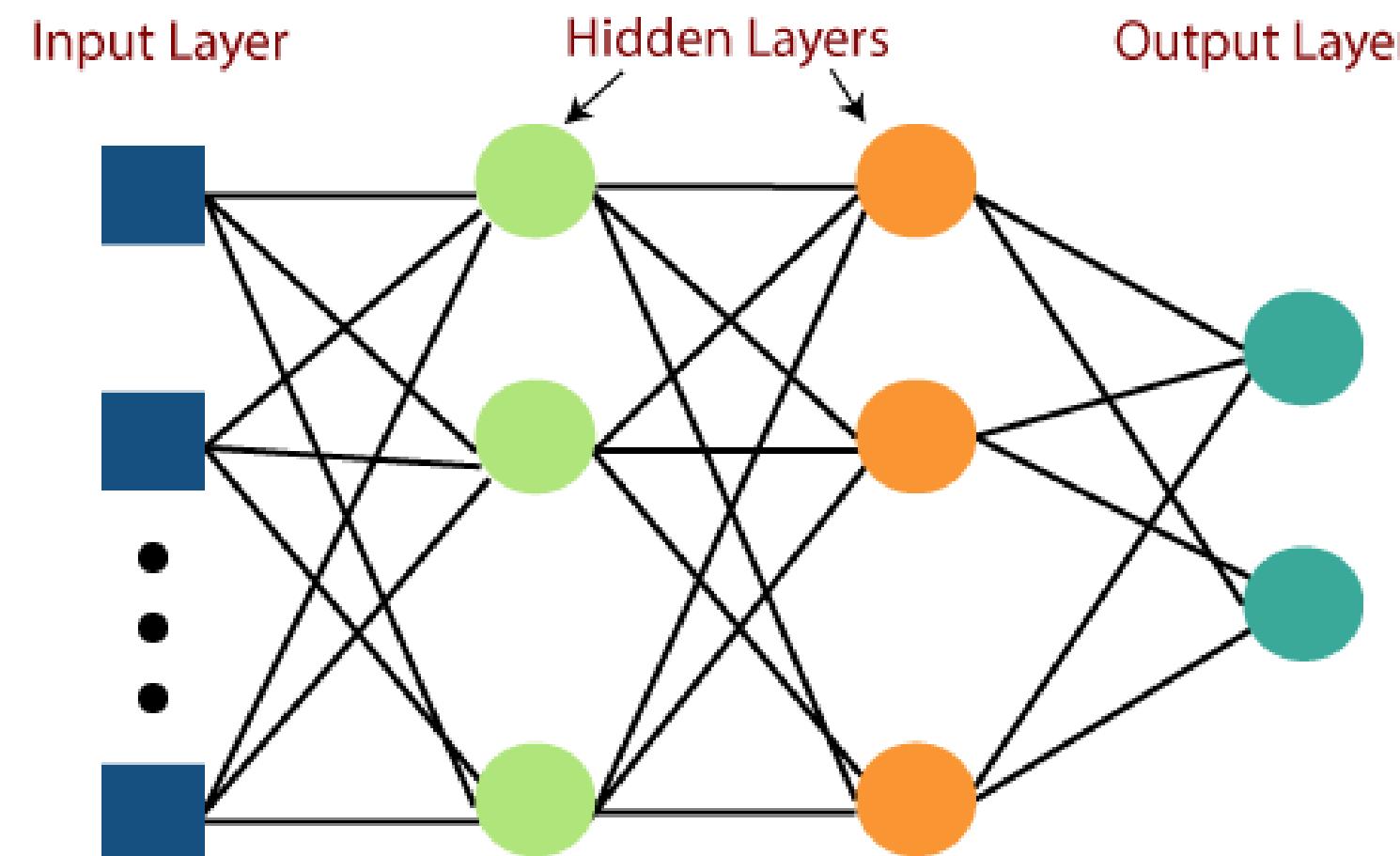
1. Single layer

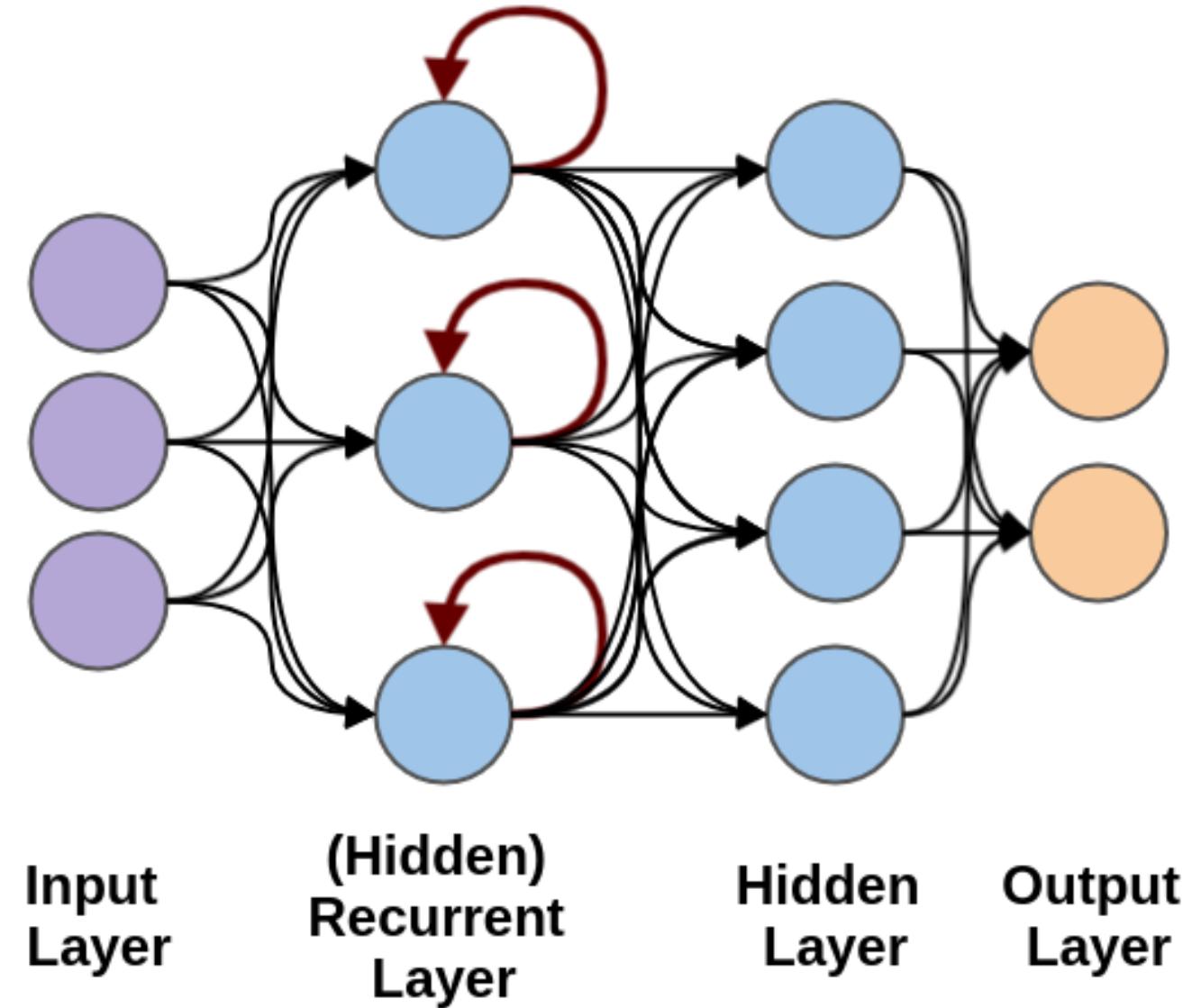
- Dalam ANN, neuron disusun dalam bentuk lapisan (layer). Pembentukan ANN yang paling sederhana yaitu single layer.
- Cara kerja dari single layer, input layer yang berasal dari sumber node diproyeksikan langsung ke output layer dari neuron (node komputasi), tetapi tidak berlaku sebaliknya.
- Pemodelan ini merupakan jenis jaringan feedforward yang dapat dilihat pada gambar disamping.
- Pada gambar tersebut input dan output memiliki 4 node, namun yang dimaksud dengan single layer yaitu output dari jaringan, sedangkan inputnya tidak memiliki pengaruh karena pada saat melakukan input tidak terjadi proses komputasi.



2. Multi layer

Cara kerja multi layer adalah input layer menyuplai input vektor pada jaringan, kemudian input yang dimasukkan melakukan komputasi pada layer yang kedua, lalu output dari layer yang kedua digunakan sebagai input dari layer yang ketiga dan seterusnya. Ilustrasi jaringan multi layer dapat di lihat pada gambar disamping.





3. Recurrent network

- Reccurent network terbentuk karena pada jaringan single layer dan multi layer harus memiliki feedback untuk dirinya sendiri pada setiap loop jaringan nya, pada reccurent network jaringan tidak memerlukan feedback untuk dirinya sendiri melainkan feedback dari input yang digunakan. Ilustrasi jaringan reccurent network dapat dilihat pada gambar disamping.

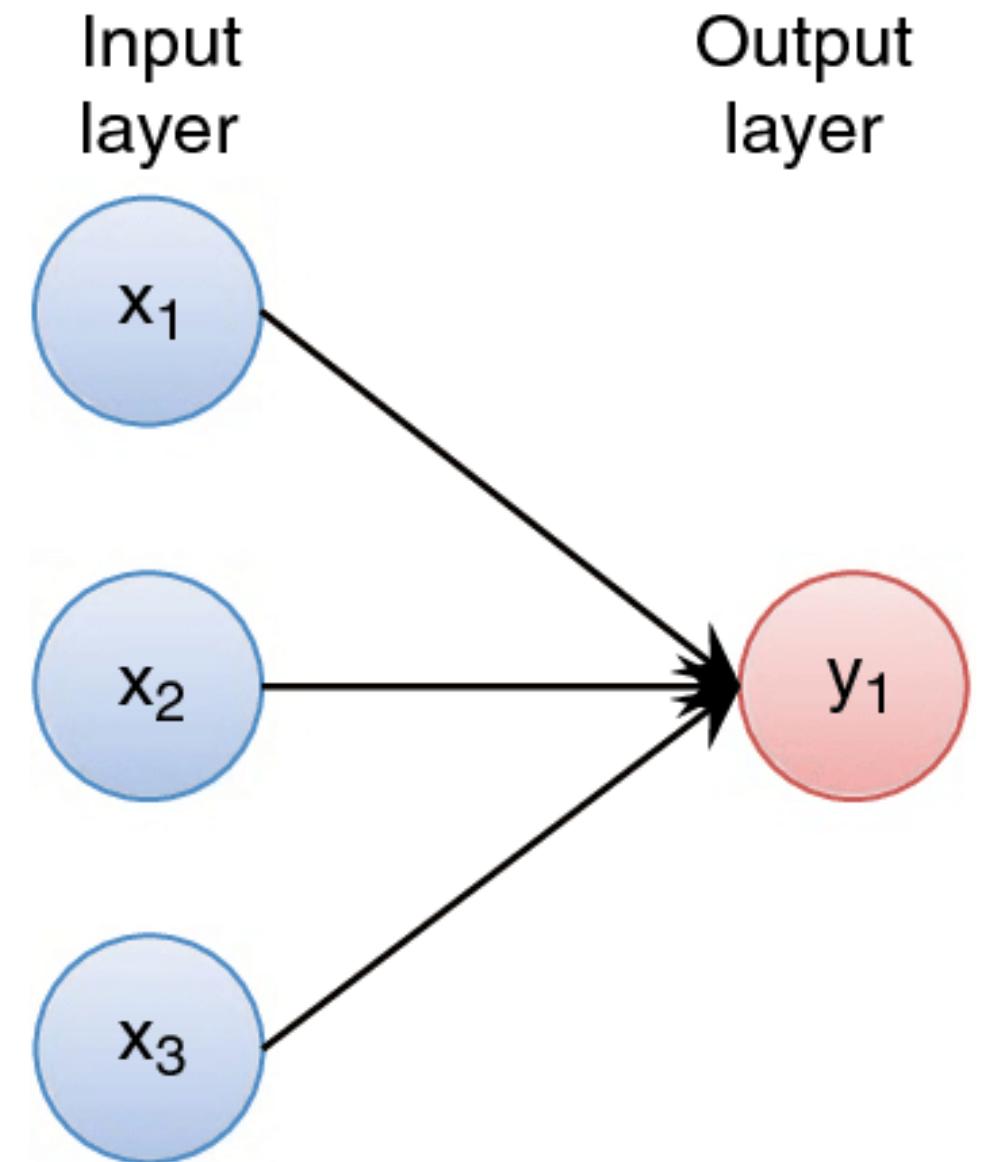
METODE PADA ANN

Sebelum membahas metode pada ANN, perlu diketahui definisi dari perceptron. Perceptron merupakan salah satu jaringan feedforward yang terdiri dari sebuah retina yang digunakan untuk akuisisi data yang mempunyai fixed-weighted connection dengan neuron layer yang pertama.

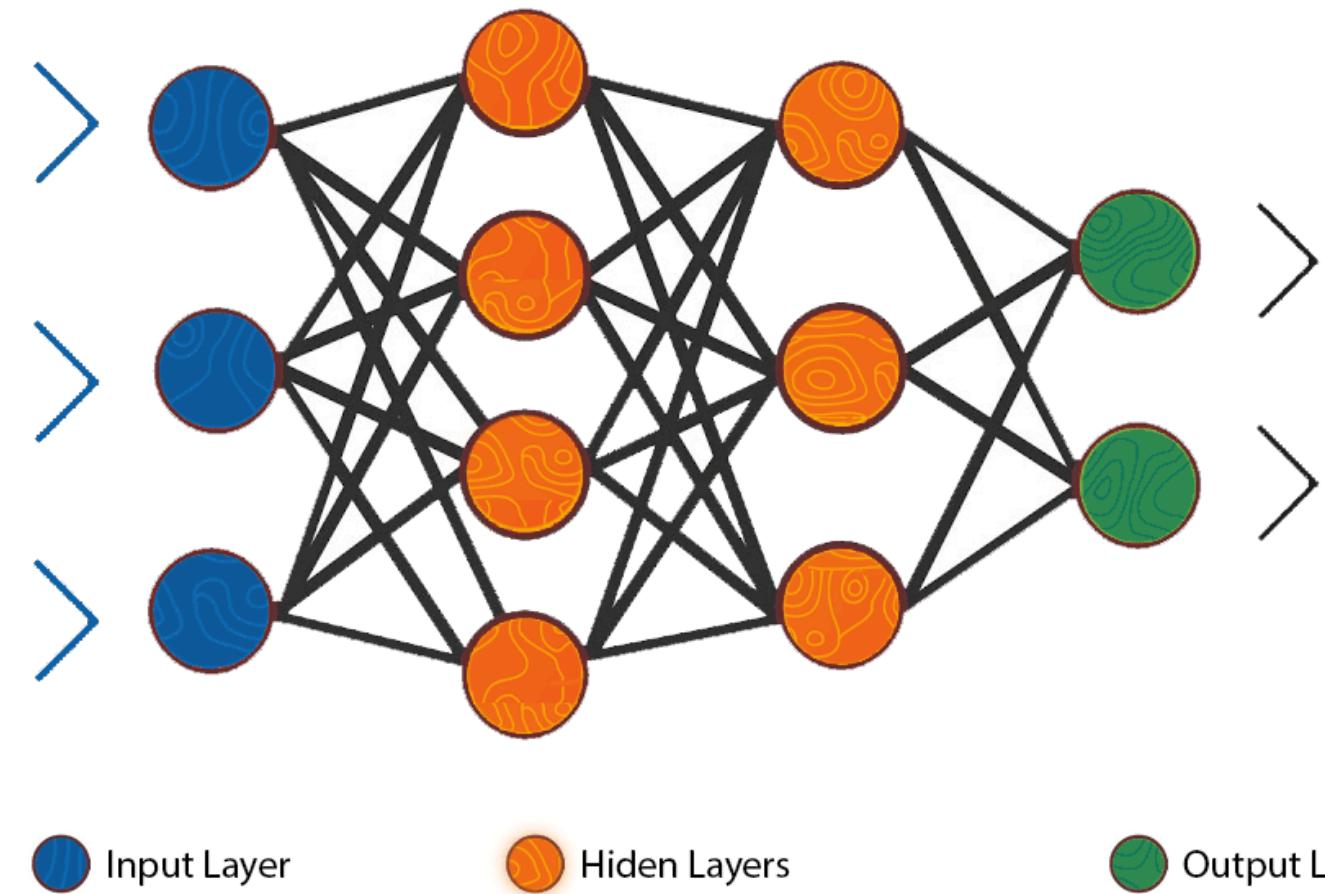
Metode - metode pada ANN:

1. Single Layer Perceptron (SLP)

- SLP merupakan sebuah perceptron yang memiliki satu variable weight dan satu variable layer dari output neuron Ω .
- Sebuah perceptron dengan beberapa output neuron dapat juga dianggap seperti beberapa perceptron berbeda dengan input yang sama.



2. Multi Layer Perceptron (MLP)



- Multi layer perceptron adalah sebuah perceptron dengan dua atau lebih trainable weight layer.
- Pada SLP dapat membagi input space dengan sebuah hyperlane sedangkan MLP dapat mengklasifikasi convex polygon dari proses hyperlane dengan mengenali pattern yang terletak di atas hyperlane.
- Sebuah n-layer perceptron adalah n-variable weight layer dan n+1 neuron layer dengan neuron layer 1 sebagai input layer. Ilustrasi dari MLP dapat dilihat pada gambar disamping.



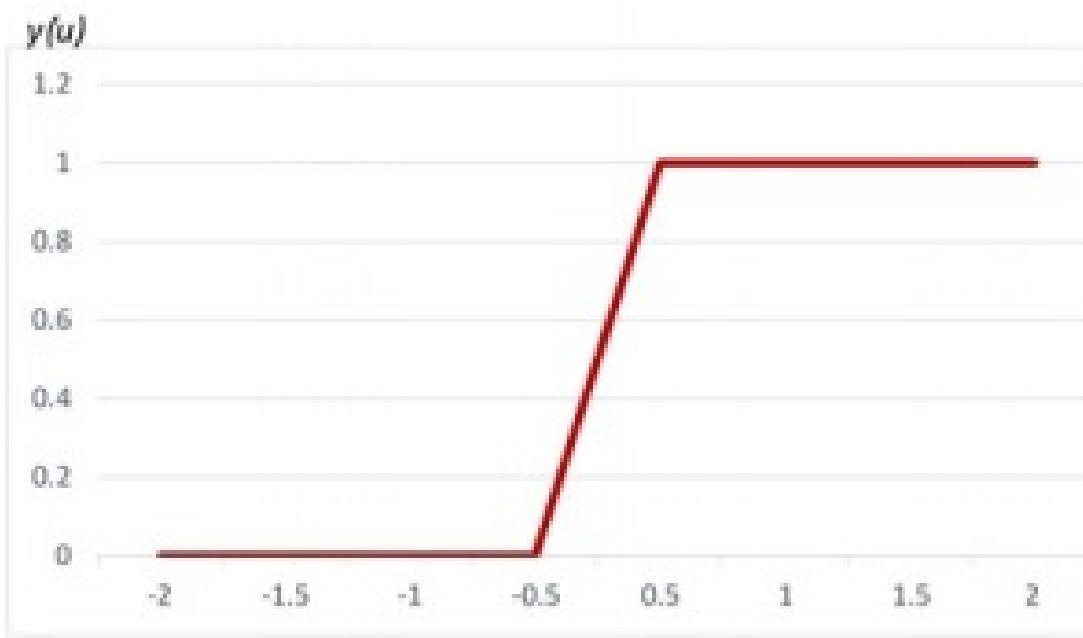
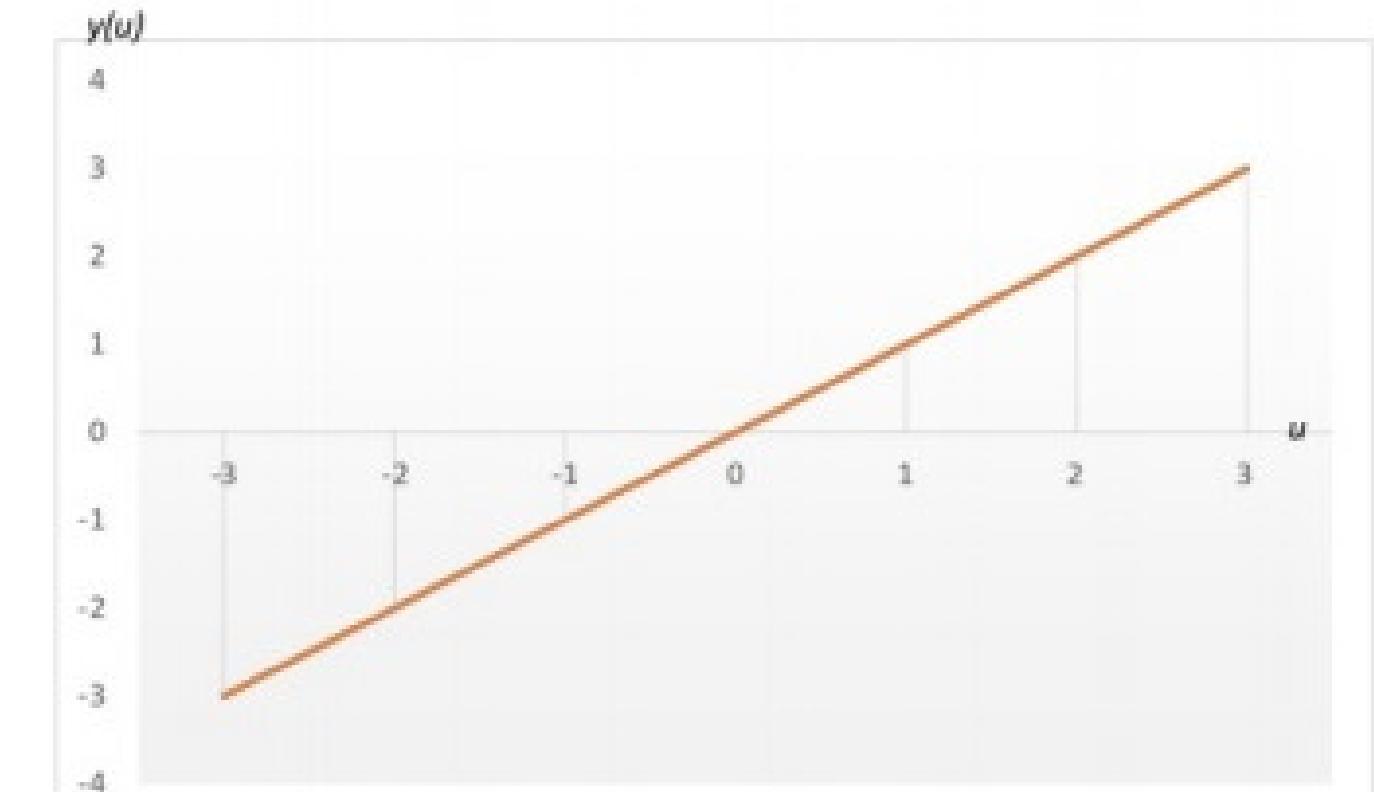
FUNGSI AKTIVASI PADA ANN

1. Fungsi Linear

- Fungsi linear menghasilkan output yang sama dengan hasil kombinasi linear.

2. Fungsi Step

- Fungsi step terbagi atas 2, yaitu fungsi step biner atau dikenal juga sebagai fungsi hard limit dan fungsi step bipolar.



Fungsi Aktivasi Step Biner (a)



Fungsi Aktivasi Step Bipolar (b)



3. Fungsi Sigmoid

- Fungsi sigmoid digunakan untuk memperoleh output yang bersifat nonlinear.

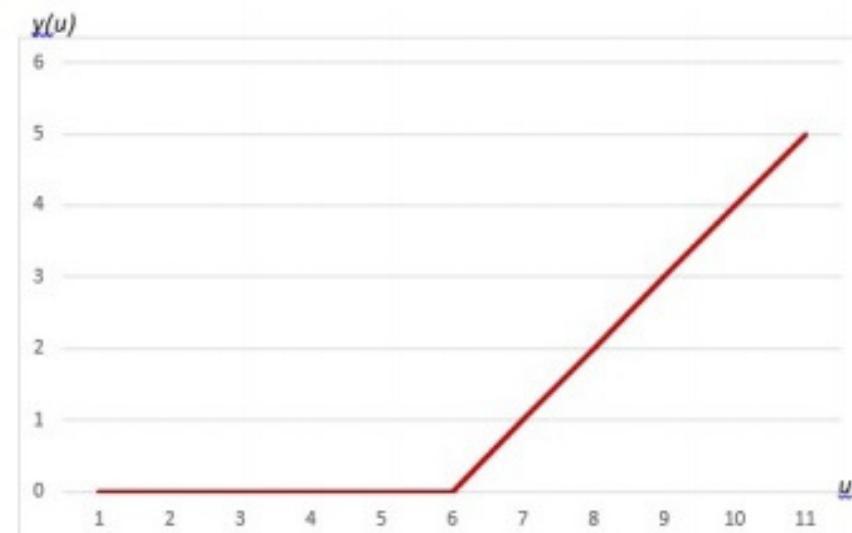
4. Fungsi Hyperbolic Tangent Sigmoid

- Fungsi ini dikenal juga sebagai fungsi tanh.

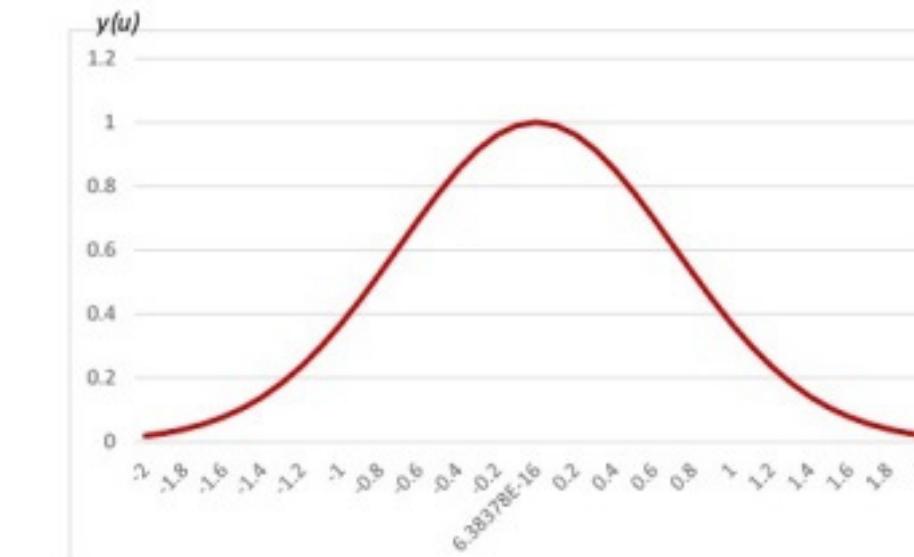
5. Fungsi Rectified Linear Unit (ReLU)

- Fungsi ini digunakan pada konteks convolutional neural networks.

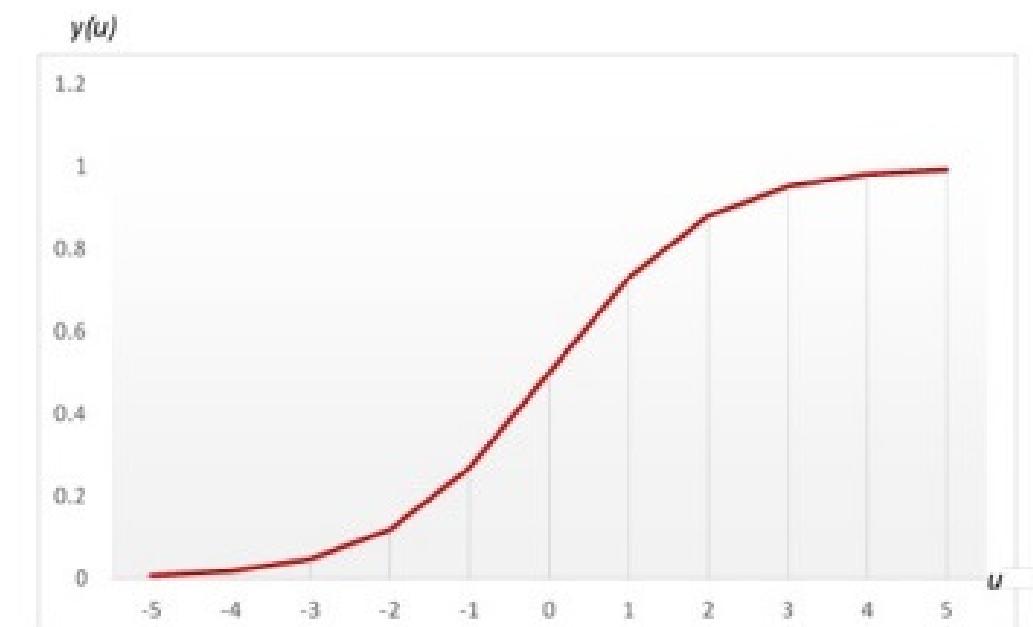
6. Fungsi Gaussian



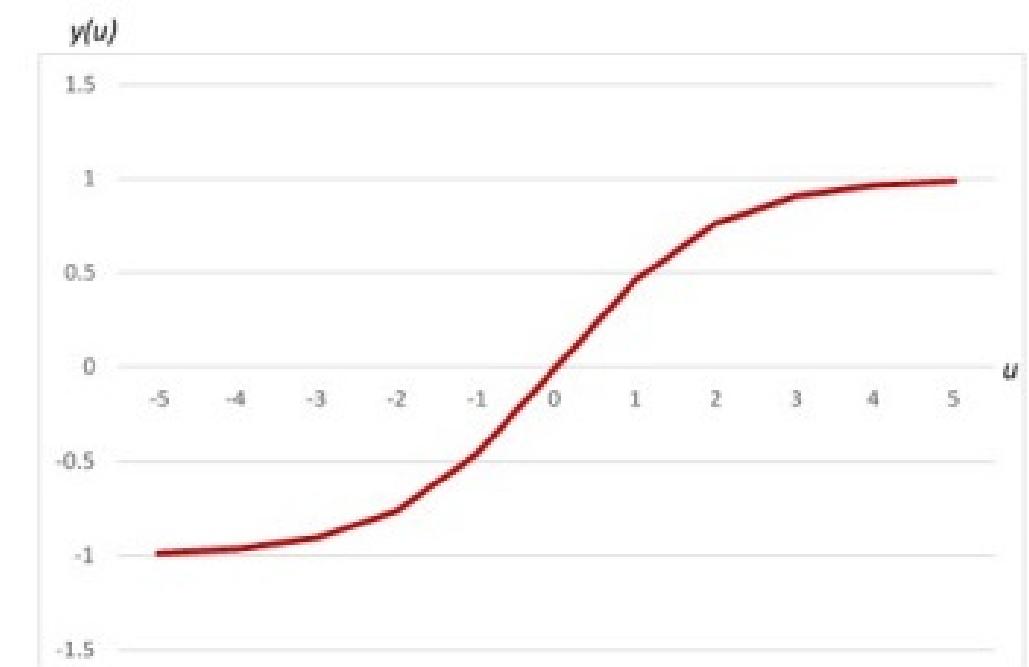
Gambar 6: Fungsi Aktivasi Gaussian



Gambar 7: Fungsi Aktivasi Gaussian



Gambar 4: Fungsi Aktivasi Sigmoid



Gambar 5: Fungsi Aktivasi *Hyperbolic Tangent* Sigmoid



KELEBIHAN DARI ANN

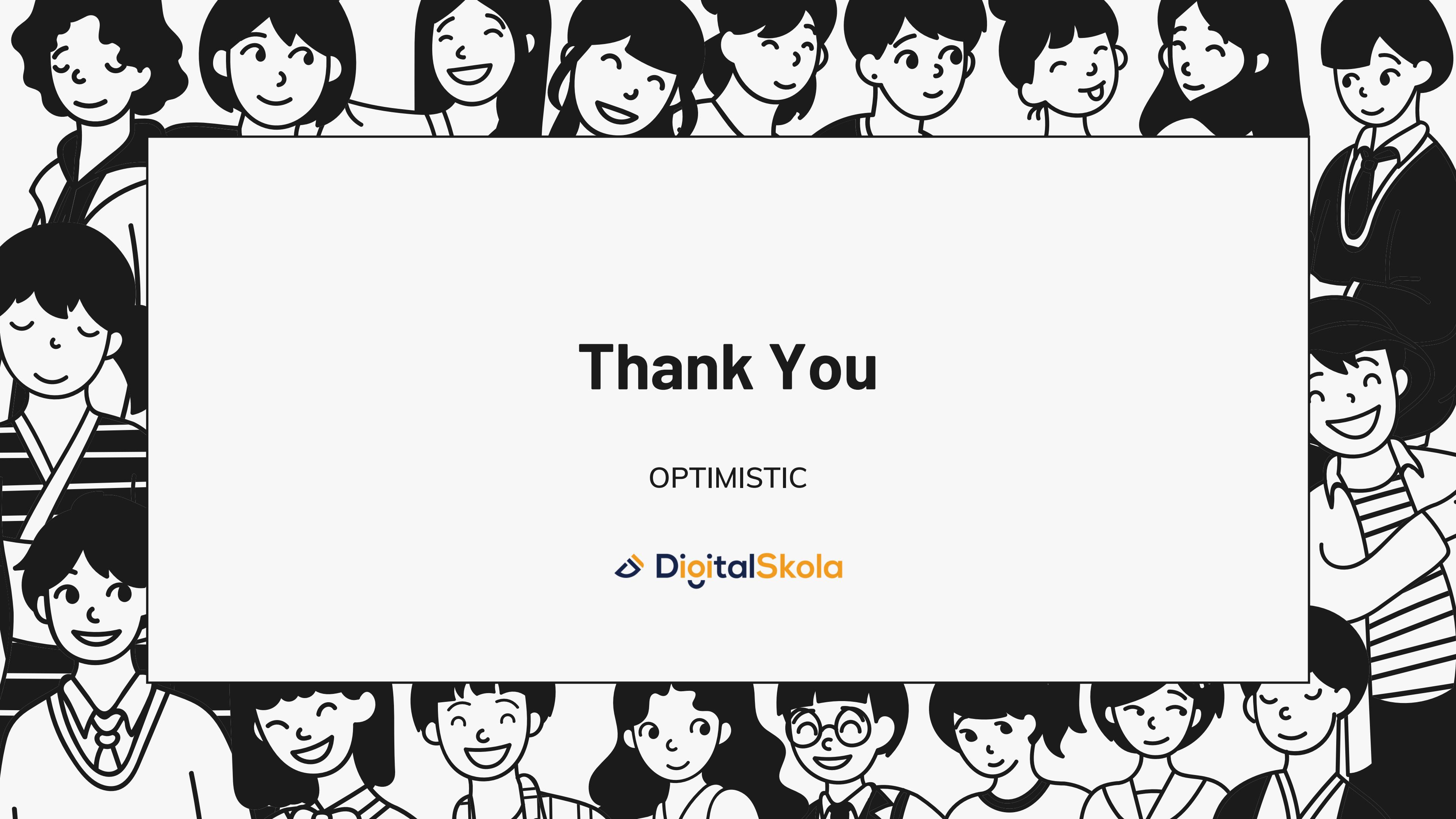
1. Mampu mengakuisisi pengetahuan walau tidak ada kepastian
2. Mampu melakukan generalisasi dan ekstraksi dari suatu pola data tertentu ANN dapat menciptakan suatu pola pengetahuan melalui pengaturan diri atau kemampuan belajar (self organizing).
3. Memiliki fault tolerance, gangguan dapat dianggap sebagai noise saja.
4. Kemampuan perhitungan secara paralel sehingga proses lebih singkat.



KEKURANGAN DARI ANN

1. Kurang mampu untuk melakukan operasi operasi numerik dengan presisi tinggi.
2. Kurang mampu melakukan operasi algoritma aritmatik, operasi logika dan simbolis.
3. Lamanya proses training yang mungkin terjadi dalam waktu yang sangat lama untuk jumlah data yang besar.





Thank You

OPTIMISTIC

