

Studi Perbandingan Algoritma Statistic, Ensemble, dan Neural Network Pada Kasus Image Classification

Salman Al Farizi Harahap¹, Muhammad Firly Rafliansyah², Ahmad Septyanto³, Ilham Pratama⁴

¹Teknik Elektro, Fakultas Teknik
Universitas Sriwijaya
Palembang, Indonesia

Penulis Korespondensi: 03041282025085@student.unsri.ac.id

Abstrak—*Machine learning* merupakan salah satu cabang *artificial intelligence* yang banyak digunakan untuk melakukan prediksi terhadap berbagai hal. Salah satu algoritma machine learning yang sudah digunakan sejak abad ke-18 adalah *Naïve Bayes*. *Naïve bayes* merupakan salah satu algoritma *machine learning* yang mengandalkan statistika dalam melakukan prediksi. Algoritma ini sangat baik dalam melakukan prediksi pada saat itu. Namun, sejak munculnya *trend Big Data* dan *Data Mining* membuat banyak sekali data dengan berbagai bentuk salah satunya adalah dalam bentuk gambar. Algoritma yang notabeneanya mengandalkan statistika ini perlahan mulai ditinggalkan dan orang lebih berminat kepada Algoritma baru dengan arsitektur baru. Oleh karena itu, peneliti ingin melakukan perbandingan antara Algoritma *Statistic* dengan algoritma-algoritma baru yaitu *Ensemble*, dan *Neural Network* Pada Kasus *Image Classification*. Peneliti menggunakan *dataset Bag Classes* dengan total data sebanyak 15.000 data gambar. Sebelum proses pelatihan, dilakukan *Image Augmentation* terlebih dahulu sebagai filter gambar agar proses pelatihan dapat dilakukan dengan lebih maksimal. Hasilnya, setiap algoritma memiliki keunggulannya masing masing. Algoritma *Statistic* kalah unggul dengan algoritma *Ensemble* dan *Neural Network* dari segi kecepatan dan performa namun masih mengungguli dari segi ukuran yaitu hanya sebesar 6,9 MB. Algoritma *Ensemble* mendapatkan hasil yang standar, Ukurannya tidak lebih besar dari *Neural Network* namun performanya jauh lebih baik jika dibandingkan dengan Algoritma *Statistic*. Algoritma ini cocok untuk project standar seperti *Mobile App* dan project standar lainnya. *Neural Network* menjadi algoritma paling baik untuk kasus *Image Classification* dengan akurasi mendekati hasil yang sempurna yakni 98,61%. Kekurangan algoritma ini hanya pada ukurannya yang terlalu besar. Algoritma ini sangat cocok digunakan untuk project-project besar seperti *Self Driving Car*, *Seat Belt Detection*, dan project besar lainnya.

Kata Kunci—*Machine learning, Algoritma Statistic, Algoritma Ensemble, Neural Network, Image Classification*

Abstrak—*Machine learning* is a branch of artificial intelligence that is widely used to predict things. One of Machine learning algorithm that has been used since the 18th century is *Naïve Bayes*. *Naïve Bayes* is a Machine learning algorithm that relies on statistics to make predictions. This algorithm is very good at making predictions at that time. However, since the emergence of the trend of *Big Data* and *Data Mining*, there has been a lot of data in various forms, one of which is in the form of images. Algorithms which in fact rely on statistics are slowly starting to be abandoned and people are more interested in new algorithms with new architecture. Therefore, we want to do a comparison between the Statistical Algorithm with the new algorithms, namely *Ensemble*, and *Neural Network* in the Case of *Image Classification*. We use the *Bag Classes* dataset with a total of 15,000 image data. Before the pelatihan process, *Image Augmentation* is done first as an image filter so that the pelatihan process can be carried out more optimally. As a result, each algorithm has its own advantages. The Statistical algorithm is inferior to the *Ensemble* and *Neural Network* algorithms in terms of speed and performance but still outperforms in terms of size, which is only 6,9 MB. The *Ensemble* Algorithm gets standard results, its size is not bigger than the *Neural Network* but its performance is much better than the Statistical Algorithm. This algorithm is suitable for standard projects such as *Mobile App* and other standard projects. *Neural Network* is the best algorithm for the *Image Classification* case with near perfect accuracy of 98.61%. The only drawback of this algorithm is that its size is too large. This algorithm is very suitable for large projects such as *Self Driving Car*, *Seat Belt Detection*, and other large projects.

Keywords—*Machine learning, Statistical Algorithm, Ensemble Algorithm, Neural Network, Image Classification*

I. PENDAHULUAN

Machine learning merupakan sebuah cabang *artificial intelligence* yang digunakan oleh para insinyur kendali dan juga insinyur *computer science*. *machine learning* pertama kali diperkenalkan oleh seorang psikologis asal Cornell University, Frank Rosenblatt dan timnya. Mesin ini mereka namai sebagai *perceptron* karena mesin ini meniru sistem syaraf pada manusia [1]. Pada tahun 1960-an *machine learning* meledak. Banyak ilmuwan data dan komputer mengembangkan konsep *perceptron* ini. Beberapa pendekatan mereka lakukan untuk membuat *machine learning* menjadi lebih baik lagi diantaranya adalah pendekatan *Deterministic* dan *Stochastic*. Hingga abad ke-21 ini, *machine learning* semakin berkembang. Banyak algoritma-algoritma baru yang muncul seperti *Decision*

Tree, *XGBoots*, dan banyak algoritma lain. Selain itu, dengan adanya *trend Big Data* dan *Data Mining* membuat *machine learning* semakin populer dan banyak dipelajari oleh para mahasiswa dan dosen di seluruh dunia.

Trend Big Data dan *Data Mining* membuat banyak sekali data dengan berbagai bentuk mulai dari data tabel, gambar, suara dan lain sebagainya. Dari sini timbul permasalahan untuk mengelola data-data ini agar bisa dimanfaatkan dengan baik. Data visual seperti video dan juga gambar adalah data yang paling umum kita temui di Internet.

Oleh Karenanya, muncul berbagai algoritma-algoritma machine learning dengan arsitektur yang berbeda-beda dan untuk menangani data dengan berbagai bentuk tersebut. Salah satu algoritma yang muncul adalah Naïve Bayes. Naïve Bayes merupakan sebuah algoritma yang muncul dari rumus probabilitas pada statistika. Naïve Bayes sendiri ditemukan oleh Thomas Bayes asal University of Edinburgh pada tahun 1722 [2]. Kelebihan dari algoritma ini terletak pada kesederhanaannya. Algoritma ini menggunakan rumus statistika seperti mean, deviasi standar, serta distribusi normal untuk menghasilkan output. Selain itu, algoritma ini juga sangat cepat untuk menghasilkan output. Ini karena kesederhanaannya membuatnya cepat dalam menentukan output. Namun, algoritma ini tidak dapat menunjukkan performa terbaiknya pada data yang berukuran besar. Algoritma ini selalu kalah dengan algoritma yang terbaru untuk masalah performa.

Karena Naïve Bayes tidak terlalu baik dalam menangani data berbentuk visual yang berukuran besar [3], peneliti ingin menguji coba beberapa algoritma-algoritma machine learning keluaran terbaru yang dapat menangani data berbentuk visual ini. Algoritma yang akan peneliti uji coba adalah algoritma-algoritma terbaru Ensemble dan juga Neural Network. Peneliti juga akan menguji coba Naïve Bayes sebagai algoritma Statistic untuk melihat seberapa baik Naïve Bayes dalam menangani data yang berbentuk visual.

II. STUDI PUSTAKA

A. Machine Learning

Machine learning merupakan program dari komputer yang dapat belajar seperti manusia namun tidak memiliki bentuk yang biologis. Machine learning dapat dikatakan sebagai otak dalam proses prediksi [4]. machine learning dapat dibagi menjadi 2 kategori yakni Supervised Learning dan Unsupervised Learning.

Supervised Learning merupakan salah satu metode prediksi pada machine learning yang sudah memiliki data koreksi. Kasus Supervised Learning dapat kita temukan pada klasifikasi dan juga regresi. Sementara itu Unsupervised Learning adalah metode prediksi yang berkebalikan dengan Supervised Learning dimana kita tidak memiliki data koreksi. Kasus Unsupervised Learning dapat kita temukan pada klusterisasi [5]. Kita dapat dengan mudah untuk menentukan apakah kita berurusan dengan Supervised atau Unsupervised Learning dengan melihat apakah data kita memiliki label atau tidak. Jika tidak, maka kita akan berurusan dengan Unsupervised Learning begitupun sebaliknya.

B. Algoritma Statistic

Algoritma *Statistic* adalah algoritma yang menggunakan ilmu statistika dalam permasalahan *machine learning*. Algoritma *Statistic* yang digunakan pada uji coba ini adalah *Naïve Bayes*. *Naïve Bayes* merupakan algoritma statistik yang menggunakan statistika inferensial serta mengacu pada pembangunan model probabilistik *Bayesian* yang memberikan probabilitas kelas *posterior* ke instance [6]. Pengklasifikasi *Naïve Bayes* sederhana menggunakan probabilitas ini untuk menetapkan instance ke kelas. Menerapkan teorema Bayes (Persamaan 1) dan menyederhanakan notasi sedikit maka kita dapatkan:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (1)$$

dimana $P(y|x)$ adalah probabilitas *posterior* label atas fitur, $P(x)$ adalah probabilitas munculnya fitur, $P(y)$ adalah probabilitas munculnya kelas, dan $P(x|y)$ adalah probabilitas *posterior* fitur atas label

C. Algoritma Ensemble

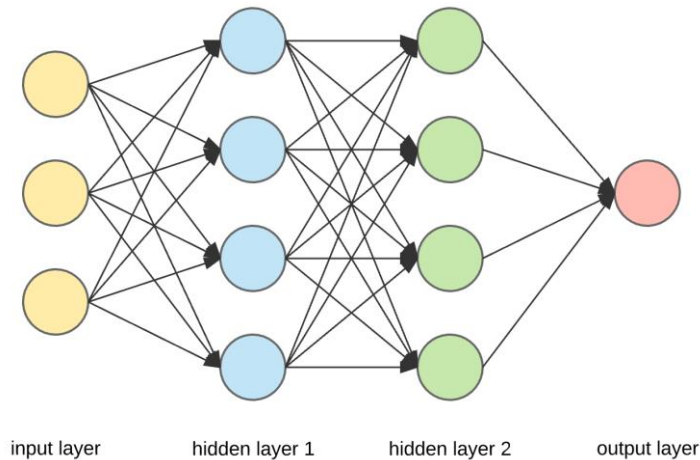
Algoritma *Ensemble* adalah sebuah algoritma *machine learning* yang dapat melakukan banyak komputasi statistik dan mengambil hasil terbaik dari komputasi tersebut. Algoritma *Ensemble* yang akan digunakan pada uji coba ini adalah Random Forest Classifier Classifier. Random Forest Classifier Classifier adalah algoritma yang berbasis pada *Decision Tree* di mana setiap *Tree* bergantung pada kumpulan variabel acak.[7]. Algoritma ini didasarkan pada pengambilan nilai maksimal dari banyaknya hasil komputasi *Decision Tree* (Persamaan 2).

$$F(x) = \arg \max \sum_1^j 1(y = h_j(x)) \quad (2)$$

dimana j adalah jumlah *Decision Tree* dan $h(x)$ adalah *Decision Tree*

D. Neural Network

Neural Network adalah sebuah algoritma *machine learning* yang meniru cara kerja dari sistem syaraf manusia [8]. Algoritma ini memiliki arsitektur yang paling kompleks daripada algoritma-algoritma sebelumnya. *Neural Network* terdiri atas 3 buah layer (Gambar 1) yaitu *input layer*, *hidden layer*, dan *output layer* dan setiap *layer* terdiri atas satu atau banyak channel. Setiap channel akan melakukan komputasi matematik atau bisa disebut juga dengan *activation function* pada setiap input yang masuk. Komputasi matematik ini akan membentuk bobot yang digunakan sebagai output pada setiap channel.



Gambar 1. Arsitektur Neural Network

Neural Network yang akan digunakan sebuah *Deep Learning Model* yang diberi nama *EfficientNetV2*. *EfficientNetV2* merupakan model dari keluarga *EfficientNet* yang memiliki ukuran yang lebih kecil namun lebih cepat dibandingkan dengan model lainnya [9]. *EfficientNetV2* terdiri dari 42 layer dengan 7 stage seperti yang tampak pada tabel I. Layer-layer yang digunakan oleh *EfficientNetV2* ini merupakan *layer* reduksi, dimana *output* pada layer akan lebih kecil dibandingkan dengan *input* pada layer sehingga parameter yang digunakan pada model ini lebih sedikit dibandingkan model pendahulunya. Oleh karena itu, *EfficientNetV2* lebih ringan dan lebih cepat dibandingkan model pendahulunya.

TABLE I. LAYER MODEL EFFICIENTNETV2

Stage	Operator	Stride	Jumlah Channel	Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	272	15
7	Conv1x1 & Pooling & FC	-	1792	1

Algoritma *Neural Network* menggunakan *hub layer* dari *Tensorflow Hub* dan juga *dense layer* sebagai *output layer*. Total parameter yang akan masuk pada algoritma ini adalah sebanyak 208.896.832 angka pada setiap epoch. 3.003 diantaranya merupakan *trainable parameter* dan sisanya *non-trainable-parameter*.

III. METODOLOGI PENELITIAN

A. Metode Penelitian

Paper ini memuat sebuah experiment yang dimulai dengan eksplorasi data dengan cara menampilkan data dari setiap kelas secara acak menggunakan *Numpy*, *Pandas*, dan *IPython*. Selanjutnya akan dilakukan *Image Augmentation* sebagai filter gambar agar proses pelatihan dapat dilakukan dengan lebih maksimal. Selanjutnya dilakukan proses pelatihan dimana peneliti akan merekam kecepatan, ukuran, serta performa setiap algoritma *machine learning* yang diuji. Rekaman ini akan dijadikan tolak ukur algoritma mana yang lebih baik digunakan untuk kasus *Image Classification*.

B. Metrics

Metrics atau tolak ukur yang digunakan sebagai hasil penelitian adalah kecepatan, ukuran, dan performa. Data kecepatan akan peneliti ambil dari jumlah waktu yang dibutuhkan algoritma pada saat proses pelatihan. Semakin sedikit waktunya artinya semakin cepat algoritma itu memproses data pada proses pelatihan begitupun sebaliknya. Data ukuran peneliti dapatkan setelah proses pelatihan selesai yaitu ketika peneliti menyimpan

algoritmanya. Dan untuk performa peneliti akan diwakilkan oleh akurasi akan menunjukkan seberapa akurat algoritma tersebut dalam memprediksi data pada dunia nyata.

C. Pengumpulan Dataset

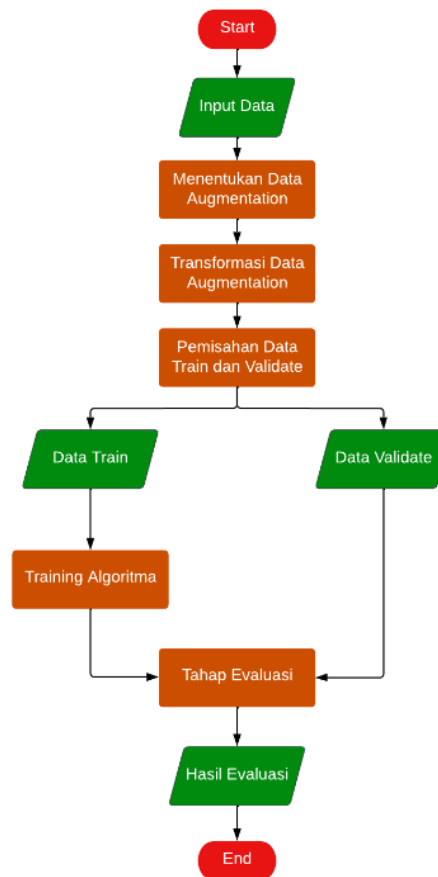
Dataset peneliti dapatkan dari *Kaggle* [10] yang merupakan komunitas *machine learning engineer* dan *data scientist*. Dataset ini di upload oleh seseorang dengan nama VinceVence, seorang Data Scientist asal Filipina pada bulan Agustus tahun 2022. Dataset ini sudah mendapatkan lisensi yakni CC BY-SA 4.0, Lisensi yang dikeluarkan oleh *Creative Common*. Dataset ini terdiri dari 3 kelas yang dapat dilihat pada tabel II.

TABLE II. DATASET BAG CLASSES

Kelas	Stride
<i>Paper Bag Images</i>	5.000
<i>Plastic Bag Images</i>	5.000
<i>Garbage Bag Images</i>	5.000
Total	15.000

D. Rancangan Uji Coba

Penelitian ini bertujuan untuk menentukan algoritma *machine learning* apa yang cocok untuk menangani masalah *Image Classification*. Oleh karena itu, peneliti menetapkan diagram alir (Gambar 2) bagaimana penelitian ini akan berlangsung. Diagram alir ini akan mewakili 1 experiment dari awal pemrosesan data hingga mendapatkan hasil.



Gambar. 2. Diagram Alir Uji Coba

IV. HASIL DAN PEMBAHASAN

Percobaan ini dilakukan dengan menggunakan tiga algoritma yakni algoritma *Statistic*, *EnsembleModel*, dan *Neural Network*. Adapun rincian prosedur percobaan ini adalah sebagai berikut.

A. Data Augmentation

1) Normalisasi Data

Normalisasi data merupakan proses yang sangat penting dalam *data preprocessing*. Karena melibatkan normalisasi dan agregasi. Normalisasi data adalah tahap *preprocessing* data dimana data diskalakan atau diubah untuk mendapatkan kontribusi yang sama dari setiap fitur [11]. Tahap normalisasi dapat dilakukan dengan berbagai cara. Dalam penelitian ini, peneliti menggunakan *rescale* untuk data gambar sebesar 1/255.

2) Pemisahan Dataset

Dataset yang telah diproses sebelumnya dibagi menjadi dua bagian dengan ukuran berbeda pada waktu yang berbeda dan digunakan sebagai dataset pelatihan dan pengujian. Data pelatihan yang mana peneliti menggunakan sebesar 75% dari dataset dan data pengujian digunakan sebesar 25% dalam algoritma klasifikasi perbandingan metode yang berbeda untuk menentukan algoritma mana yang menghasilkan hasil terbaik dari segi *accuracy*, *loss*, dan *Confusion Matrix*.

3) Transformasi Data

Transformasi memainkan peran penting dalam aplikasi pengolahan citra untuk peningkatan kualitas gambar, analisis data gambar seperti analisis perilaku fungsi gelombang [12]. Transformasi data dilakukan secara acak pada setiap data. Transformasi secara acak ini dilakukan agar algoritma tidak terkesan menghafal dalam proses pelatihan.

- *Rotasi*: Rotasi adalah teknik pengolahan citra dengan mengubah koordinat piksel dengan sudut tertentu. Rotasi dilakukan sebesar 20° searah jarum jam.
- *Horizontal Flip*: *Horizontal flip* adalah teknik pengolahan citra dengan memutar gambar atau citra ke cerminannya. Cara kerja operasi horizontal flip dengan transformasi pembalikan frame berdasarkan y-axis (sumbu y),
- *Shear Range*: *Shear range* menerapkan n metode shear transformation sebanyak 0.2.
- *Height Shift*: *Height Shift* ini akan melakukan pergeseran pada citra secara vertikal atau dari atas ke bawah. Rentang nilai pergeseran ditentukan berdasarkan persentase dari total tinggi citra yaitu sebesar 10.
- *Width Shift*: *Weight shift* ini transformasi dengan melakukan pergeseran pada citra secara horizontal atau dari kanan ke kiri. Rentang nilai pergeseran ditentukan berdasarkan persentase dari total lebar citra yaitu sebesar 10.

B. Data Cleaning

Data yang peneliti ambil melalui *Kaggle* ini harus dibersihkan untuk memastikan bahwa data memenuhi standar kualitas sebelum pemodelan dimulai. Data yang dibersihkan ini hanya untuk Algoritma *Statistic* dan *Ensemble*. Pembersihan data ini dilakukan dengan merubah bentuk tensor menjadi matrix dengan fungsi *reshape*. Hal ini dilakukan karena algoritma *Statistic* dan *Ensemble* tidak dapat menerima masukan berupa tensor.

C. Evaluasi

1) Ukuran

Ukuran menunjukkan seberapa berat sebuah algoritma setelah dilatih dengan 15.000 data gambar. Data ukuran didapat dengan menyimpan algoritma dalam komputer dan melihat sizenya pada *properties*. Ukuran dari ketiga algoritma dapat dilihat pada tabel III. Dimana algoritma *statistic* memiliki ukuran yang paling kecil disusul oleh algoritma *Ensemble* dan yang paling besar adalah *Neural Network*.

TABLE III. UKURAN ALGORITMA

Algoritma	Ukuran (Dalam MB)
<i>Statistical</i>	6,9
<i>Ensemble</i>	25
<i>Neural Network</i>	749,38

2) Kecepatan

Kecepatan menunjukkan seberapa cepat sebuah algoritma mempelajari pola pada gambar saat proses pelatihan. Data kecepatan didapat ketika proses pelatihan selesai. Jika dilihat pada tabel IV, algoritma *statistic* paling cepat dalam mempelajari pola gambar pada dataset disusul oleh *Neural Network* dan yang paling lama adalah algoritma *Ensemble*. Kecepatan ini dapat dipengaruhi oleh perangkat keras, dimana *Neural Network* memiliki keunggulan karena dapat menggunakan GPU pada proses pelatihannya.

TABLE IV. KECEPATAN ALGORITMA

Algoritma	Kecepatan (Dalam detik)
<i>Statistical</i>	364
<i>Ensemble</i>	36531
<i>Neural Network</i>	10825

3) Performa

Performa menunjukkan tingkat keakuratan sebuah algoritma dalam memprediksi data pada dunia nyata. Performa diwakili oleh akurasi dimana akurasi merupakan perbandingan antara hasil prediksi benar dan salah (persamaan 3), akurasi dari ketiga algoritma dapat dilihat pada tabel V.

$$Accuracy(y, \hat{y}) = \arg \max \sum_{i=0}^{n-1} 1(\hat{y} = y) \quad (3)$$

dimana y adalah jumlah hasil prediksi algoritma dan \hat{y} adalah label sebenarnya

TABLE V. AKURASI SETIAP ALGORITMA

Algoritma	Akurasi (Dalam persen)
<i>Statistical</i>	34,42
<i>Ensemble</i>	74,18
<i>Neural Network</i>	98,61

V. KESIMPULAN

Dari hasil perbandingan ketiga algoritma, dapat disimpulkan bahwa Naïve Bayes sebagai algoritma Statistic tidak dapat menangani kasus image classification. Namun, algoritma ini masih memiliki keunggulan dibandingkan dengan dua algoritma. Algoritma Statistic unggul pada size yang kecil dan cocok digunakan untuk penelitian kecil dengan data yang kecil. Selanjutnya untuk algoritma Ensemble unggul dalam size yang kecil namun performanya yang jauh melampaui algoritma statistic. Algoritma Ensemble cocok digunakan dalam project standar seperti mobile app. Dan terakhir, Neural Network unggul dalam kecepatan dan performa. Neural Network dapat dikatakan algoritma paling cocok dalam menangani kasus image classification. Satu-satunya kelemahan dari algoritma ini adalah pada ukurannya yang terlalu besar jika dibandingkan dengan ketiga algoritma yang dibandingkan. Neural Network cocok digunakan untuk project besar yang mementingkan keakuratan dalam prediksinya seperti Self Driving Car, Auto Pilot, dan project besar lainnya.

DAFTAR PUSTAKA

- [1] A. L. Fradkov, "Early history of machine learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1385–1390, 2020, doi: 10.1016/j.ifacol.2020.12.1888.
- [2] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 713–714. doi: 10.1007/978-0-387-30164-8_576.
- [3] M. Makhtar, H. Nawang, and S. N. W. Shamsuddin, "Analysis on students performance using naïve Bayes classifier," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 16, pp. 3993–4000, 2017.
- [4] H. Wehle, "ML – AI- COGNITIVE," no. August, 2017.
- [5] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *Int. J. Adv. Res. Artif. Intell.*, vol. 2, no. 2, 2013, doi: 10.14569/ijarai.2013.020206.
- [6] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 403–412, 2018, doi: 10.1016/B978-0-12-809633-8.20473-1.
- [7] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Mach. Learn.*, no. January, 2012, doi: 10.1007/978-1-4419-9326-7.
- [8] R. Dastres and M. Soori, "Artificial Neural Network Systems," *Int. J. Imaging Robot.*, vol. 2021, no. 2, pp. 13–25, 2021, [Online]. Available: www.ceserp.com/cp-jour
- [9] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," 2021, [Online]. Available: <http://arxiv.org/abs/2104.00298>
- [10] VinceVence, "Plastic - Paper - Garbage Bag Synthetic Images," 2022. <https://www.kaggle.com/datasets/vencerlanz09/plastic-paper-garbage-bag-synthetic-images?select=ImageClassesCombinedWithCOCOAnnotations>
- [11] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.
- [12] A. Andrian, "Perancangan Perangkat Lunak Kompresi Citra Menggunakan Transformasi Wavelet dan PCA," *Algoritma. J. Ilmu Komput. Dan Inform.*, vol. 3, no. 1, p. 1, 2019, doi: 10.30829/algoritma.v3i1.4430.