Klasifikasi pada Citra Bunga dengan Ekstraksi Fitur Color Histogram

Muhammad Ghudafa Taufik Akbar[1], Henny Leidiyana[2]
[1]Fakultas Ilmu Komputer, Program Pasca Sarjana, Universitas Nusa Mandiri,
Fakultas Teknik dan Informatika, Program Studi Sistem Informasi[2]
[1]ghudafa.shino@gmail.com, [2]henny.hnl@bsi.ac.id

Abstract

Differences in flower characteristics cause differences in flower species so knowledge is needed to be able to classify them. Flowers can be characterized by their color. Flowers of the same type can even have several different colors. One of the techniques for obtaining patterns can be done through color feature extraction. This study attempts to extract images using the color histogram technique. Many studies have shown the effectiveness of color feature extraction before classifying. Color Histogram is the most widely used technique for extracting color features from an image because it represents images from different points of view. The dataset used in this study uses flower images with almost the same shape, namely sunflower, calendula, black eyed susan, common daisy. Making a classification model for these types of flowers uses machine learning algorithms Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Linear Discriminant Analysis, Naive Bayes, Support Vector Machine. Furthermore, the author uses the Confusion Matrix to evaluate the trained model and to produce accurate scores using 10-fold cross validation. The highest accuracy score obtained from the Random Forest model is 82%.

Keywords: Feature Extraction, Color Histogram, machine learning

Abstrak

Perbedaan ciri pada bunga menyebabkan perbedaan spesies bunga sehingga diperlukan pengetahuan untuk dapat mengklasifikasikannya. Bunga dapat dicirikan dari warnanya. Pada bunga dengan jenis yang sama bahkan bisa memiliki beberapa warna yang berbeda. Salah satu teknik untuk mendapatkan pola bisa dilakukan melalui ekstraksi fitur warna. Penelitian ini mencoba untuk melakukan ekstraksi citra dengan teknik color histogram. Penelitian yang menunjukkan efektifitas ekstraksi fitur warna sebelum melakukan klasifikasi sudah cukup banyak dilakukan. Color Histogram adalah teknik yang paling banyak digunakan untuk mengekstraksi fitur warna dari suatu citra karena mewakili gambar dari sudut pandang yang berbeda. Dataset yang digunakan dalam penelitian ini menggunakan citra bunga dengan bentuk yang hampir sama yaitu sunflower, calendula, black eyed susan, common daisy. Pembuatan model klasifikasi terhadap jenis-jenis bunga tersebut menggunakan algoritma pembelajaran mesin Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Linear Discriminant Analysis, Naive Bayes, Support Vector Machine. Selanjutnya Penulis menggunakan Confusion Matrix untuk mengevaluasi model terlatih dan untuk menghasilkan skor akurasi menggunakan 10-fold cross validation. Skor akurasi tertinggi diperoleh dari model Random Forest sebesar 82%.

Kata Kunci: Ekstraksi Fitur, Color Histogram, pembelajaran mesin

I. Pendahuluan

Bunga dengan aneka warna membuat kita takjub akan keindahannya. Banyaknya warna bunga yang ada merah, biru, dan ungu terutama berasal dari pigmen yang disebut klorofil, karotenoid, anthocyanin dan betalin dll (pigmen yang ada pada tumbuhan), yang termasuk dalam kelas bahan kimia yang disebut flavanoid. Keindahan bunga bisa

disaksikan melalui gambar-gambar yang merepresentasikan objek aslinya tanpa harus melihatnya langsung. Bukan hanya untuk dinikmati keindahannya saja, orang menggunakan citra bunga sebagai objek penelitian. [1][2][3][4]

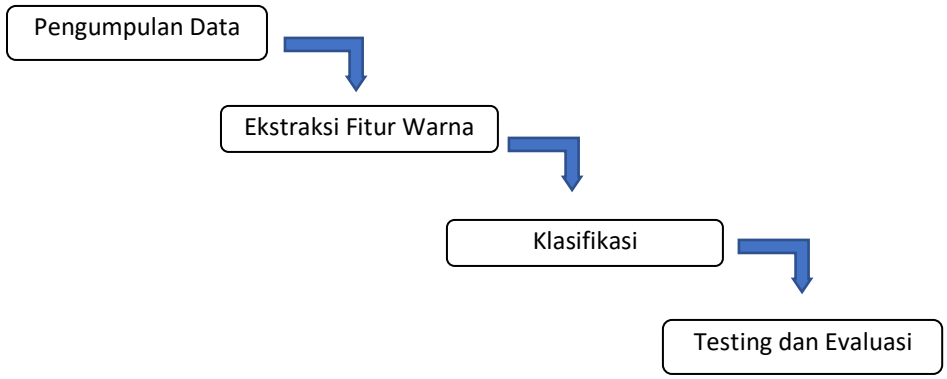
Sebuah citra terdiri dari berbagai warna dan setiap warna merupakan kombinasi dari Merah, Hijau, Biru. Histogram dapat memvisualisasikan berapa banyak proporsi warna RGB yang dimiliki dalam sebuah gambar. Histogram sebenarnya menunjukkan seberapa sering berbagai warna muncul dalam sebuah gambar tetapi bukan lokasi warna dalam sebuah gambar.

Klasifikasi dengan metode pembelajaran mesian terhadap objek bunga sudah pernah dilakukan [5][6] dengan mengkombinasikan Teknik ekstraksi fitur. Penelitian terkait dengan ekstraksi menggunakan fitur warna juga telah banyak dilakukan [7][8]

Tujuan utama dari penelitian ini adalah untuk mengklasifikasikan citra bunga ke dalam kelas warnanya menggunakan color histogram dan menilai kinerja klasifikasi berdasarkan akurasi. Algoritma pembelajaran mesin untuk klasifikasi citra bunga mengunakan algoritma Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Linear Discriminant Analysis, Naive Bayes, Support Vector Machine. Sebelum pembuatan model klasifikasi, terhadap citra bunga dilakukan ekstraksi fitur terlebih dahulu dengan teknik color histogram. Percobaan dilakukan menggunakan citra dengan bentuk yang mirip yang berasal dari kelompok bunga aster dengan jenis sunflower, calendula, black eyed susan, daisy. Setelah model dilatih, berikutnya adalah evaluasi untuk melihat kinerja model menggunakan Confusion Matrix dan 10 cross validation untuk untuk menghasilkan skor akurasi.

II. Metode Penelitian

Seperti ditunjukkan pada Gambar 1, tahapan penelitian ini dibagi menjadi 4 yaitu preprocessing, ekstraksi fitur dengan teknik color histogram, klasifikasi dengan metode pembelajaran mesin Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Linear Discriminant Analysis (LDA), Naive Bayes, Support Vector Machine (SVM)



Gambar 1. Tahapan Penelitian

Dataset diperoleh dari kaggle.com yaitu salah satu penyedia dataset yang bersifat publik. Citra yang dipilih memiliki bentuk yang mirip yang berasal dari kelompok bunga aster dengan jenis black eyed 1000 citra, calendula 978 citra, common daisy 980, sunflower 1027 citra, total dataset 3985 citra bunga.

Black eyed susan

Calendula

Sunflower

Common daisy

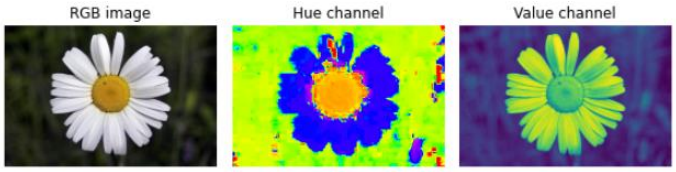
Gambar 2. Sampel dataset Citra Bunga

Tahap penelitian berikutnya adalah ekstraksi fitur warna. Ekstraksi Fitur Warna adalah Fitur warna yang diekstraksi dengan Color Histogram dan Color Descriptor. Histogram Warna menentukan distribusi piksel warna dalam gambar. Histogram warna menggunakan dua jenis ruang warna yaitu RGB dan HSV. Color Histogram (CH), memuat kemunculan setiap warna yang diperoleh dengan menghitung semua piksel citra yang memiliki warna tersebut. Setiap piksel dikaitkan ke bin histogram tertentu hanya berdasarkan warnanya sendiri, dan kesamaan warna di seluruh bin berbeda atau ketidaksamaan warna dalam bin yang sama tidak diperhitungkan. Karena piksel apa pun dalam gambar dapat dideskripsikan oleh tiga komponen dalam ruang warna tertentu (misalnya, komponen merah, hijau, dan biru dalam ruang RGB atau rona, saturasi, dan nilai dalam ruang HSV), sebuah histogram, yaitu distribusi dari jumlah piksel untuk setiap nampan terkuantisasi, dapat ditentukan untuk setiap komponen. Deskriptor warna terdiri dari color expectancy, color variance dan color skewness. Harapan warna adalah rata-rata atau rata-rata intensitas dalam gambar. Variasi warna adalah akar kuadrat dari standar deviasi. Kemiringan warna adalah ukuran asimetri distribusi probabilitas dari variabel acak bernilai nyata. Dua jenis kemiringan adalah kemiringan positif dan kemiringan negatif. [9]

Setelah dilakukan ekstraksi fitur berikutnya adalah klasifikasi citra. Klasifikasi citra adalah salah satu domain penelitian yang populer di bidang computer vision dan juga merupakan sistem klasifikasi citra dasar di bidang aplikasi citra lainnya. Klasifikasi dibagi menjadi tiga bagian penting: preprocessing citra, ekstraksi fitur citra, dan pengklasifikasi.[10]. Tahap terahkhir yaitu evaluasi untuk mengukur kinerja model menggunakan confusion matrix dan 10. Cross fold validation.

III. Hasil dan Pembahasan

Dalam penelitian diawali dengan melakukan pengumpulan data untuk kebutuhan percobaan. Percobaan dilakukan menggunakan bahasa pemrograman python yang ditulis pada google colaboratory. Karena dataset yang digunakan sudah siap untuk dilatih maka pada tahap preprocessing penulis hanya melakukan pengunggahan dataset ke dalam google drive agar dapat diakses melalui colaboratory. Setelah itu dilanjutkan dengan kegiatan ekstraksi fitur warna. fitur warna citra diekstrak untuk histogram warna. Dalam penelitian ini menggunakan HSV color histogram. Ruang warna HSV adalah salah satu ruang warna yang digunakan oleh manusia, dalam menentukan dan mendeskripsikan warna. Untuk mendapatkan karakteristik warna HSV dilakukan proses konversi warna dari RGB ke HSV.



Gambar 3. Konversi Citra Sampel RGB ke HSV

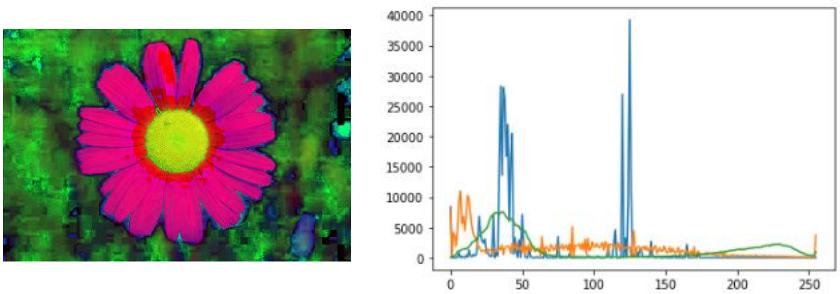
Fitur dihasilkan dengan memanfaatkan properti ruang warna HSV. HSV dapat menentukan intensitas dan variasi bayangan di dekat tepi objek, dengan demikian mempertajam batas dan mempertahankan warna informasi dari setiap piksel [11]

Algoritma ekstraksi fitur warna[12]
Input: RGB color citra bunga
Output: Color histogram
mulai:
1. baca citra dalam RGB color space.
2. konversi citra ke HSV
3. ekstrak fitur warna Hue(H), Saturation(S),
4. Intensity value(V)
5. Gunakan bin Hue, Saturation and Intensity bins untuk memperoleh histogram

Implementasi dalam python

```
def fitur_histogram(image, mask=None):
    # konversi image ke HSV color-space
    citra=cv2.cvtColor(image,cv2.COLOR_BGR2HSV)
    # menghitung color histogram
    hist=cv2.calcHist([citra],[0,1,2],None,[bin,bin,bin],[0,256,0,256,0,256])
    # menormalisasi histogram
    cv2.normalize(hist,hist)
    # return hasil
    return hist.flatten()
```

Percobaan dengan python menggunakan OpenCV dengan fungsi cv2.calcHist() untuk histogram. Berikut spesifikasi parameter yang terkait dengan fungsi cv2.calcHist():
1. citra: berisi data citra yang akan dibuat histogramnya.
2. channel: berbentuk list[]. Ini adalah indeks channel yang dihitung histogramnya. Karena citra berwarna maka menggunakan [0], [1] dan [2] untuk menghitung histogram.
3. mask: Untuk menemukan histogram citra, di sini bernilai NONE.histSize: ini menunjukkan jumlah "bin", di sini menggunakan list. Untuk skala penuh maka menggunakan nilai [256].
4. ranges: Normalnya adalah [0,256].



Gambar 4. Histogram HSV Citra Sampel

Fitur wana citra yang sudah diekstraksi dimasukkan ke pengklasifikasi dan kelas prediksinya dihitung. Hal ini dilakukan untuk setiap model. Untuk membuat model klasifikasi, data dipisah menjadi data training dan data testing, dalam penelitian ini menggunakan cross validation. Cross validation adalah salah satu metode resampling data yang paling banyak digunakan untuk memperkirakan prediksi kesalahan model yang sebenarnya dan untuk menyetel parameter model. [13] K-fold cross validation yang dipakai dalam penelitian ini menggunakan nilai k = 10. Meskipun sampai saat ini belum ada yang memastikan berapa nilai k yang harus dipilih untuk validasi silang k-fold namun biasanya nilai k yang digunakan yaitu 10 karena perkiraan kesalahan prediksi hampir tidak bias dalam 10 fold cross validation. [13]

Tabel 1 menunjukan kinerja model menggunakan metrik akurasi. Akurasi klasifikasi adalah rasio prediksi yang benar terhadap total prediksi yang dibuat. Akurasi klasifikasi juga dapat dengan mudah diubah menjadi tingkat kesalahan klasifikasi atau tingkat kesalahan dengan membalikkan nilainya. Dari Tabel 1 diperoleh informasi akurasi tertinggi dihasilkan oleh algoritma random forest, disusul oleh SVM, logistic regression, decision tree, LDA, KNN, dan paling kecil yaitu naïve bayes.

Tabel 1. Hasil Kinerja Model Klasifikasi

Model Klasifikasi	Akurasi
Logistic Regression	70%
Decision Tree	70%
Random Forest	82%
K Nearest Neighbors	64%
Linear Discriminant Analysis	66%
Naïve Bayes	43%
Support Vector Machine	70%

Untuk melihat relevansi data maka Gambar 5 menunjukan nilai presisi. Berapa banyak nilai positif aktual yang ditangkap model kami melalui pelabelan sebagai Positif (true) dapat dilihat dari nilai recall pada Gambar 5.Agar bisa dilihat keseimbangan antara precision dan recall maka Penulis juga menampilkan nilai F1 score.

Logistic Regression

	precision	recall	f1-score	support
black_eyed_susan	0.61	0.66	0.63	208
calendula	0.51	0.46	0.48	178
common_daisy	0.88	0.95	0.92	217
sunflower	0.76	0.70	0.73	194
accuracy			0.70	797
macro avg	0.69	0.69	0.69	797
weighted avg	0.70	0.70	0.70	797

Decision Tree

	precision	recall	f1-score	support
black_eyed_susan	0.70	0.68	0.69	208
calendula	0.57	0.61	0.59	178
common_daisy	0.89	0.83	0.86	217
sunflower	0.63	0.64	0.64	194
accuracy			0.70	797
macro avg	0.70	0.69	0.69	797
weighted avg	0.70	0.70	0.70	797

KNN

	precision	recall	f1-score	support
black_eyed_susan	0.53	0.63	0.58	208
calendula	0.43	0.46	0.44	178
common_daisy	0.86	0.88	0.87	217
sunflower	0.79	0.54	0.64	194
accuracy			0.64	797
macro avg	0.65	0.63	0.63	797
weighted avg	0.66	0.64	0.64	797

LDA

	precision	recall	f1-score	support
black_eyed_susan	0.57	0.69	0.62	208
calendula	0.48	0.48	0.48	178
common_daisy	0.85	0.86	0.85	217
sunflower	0.73	0.56	0.63	194
accuracy			0.66	797
macro avg	0.66	0.65	0.65	797
weighted avg	0.67	0.66	0.66	797

Naive Bayes

	precision	recall	f1-score	support
black_eyed_susan	0.49	0.44	0.46	208
calendula	0.31	0.28	0.29	178
common_daisy	0.70	0.23	0.34	217
sunflower	0.40	0.80	0.54	194
accuracy			0.43	797
macro avg	0.48	0.43	0.41	797
weighted avg	0.49	0.43	0.41	797

SVM

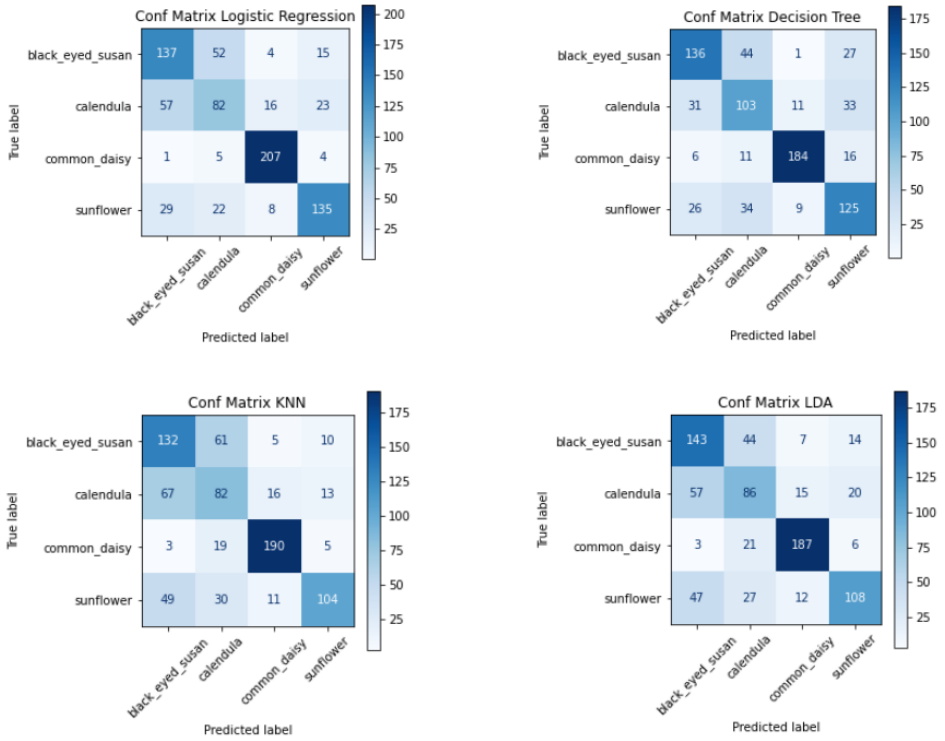
	precision	recall	f1-score	support
black_eyed_susan	0.65	0.63	0.64	208
calendula	0.50	0.57	0.53	178
common_daisy	0.88	0.94	0.91	217
sunflower	0.79	0.64	0.71	194
accuracy			0.70	797
macro avg	0.70	0.70	0.70	797
weighted avg	0.71	0.70	0.70	797

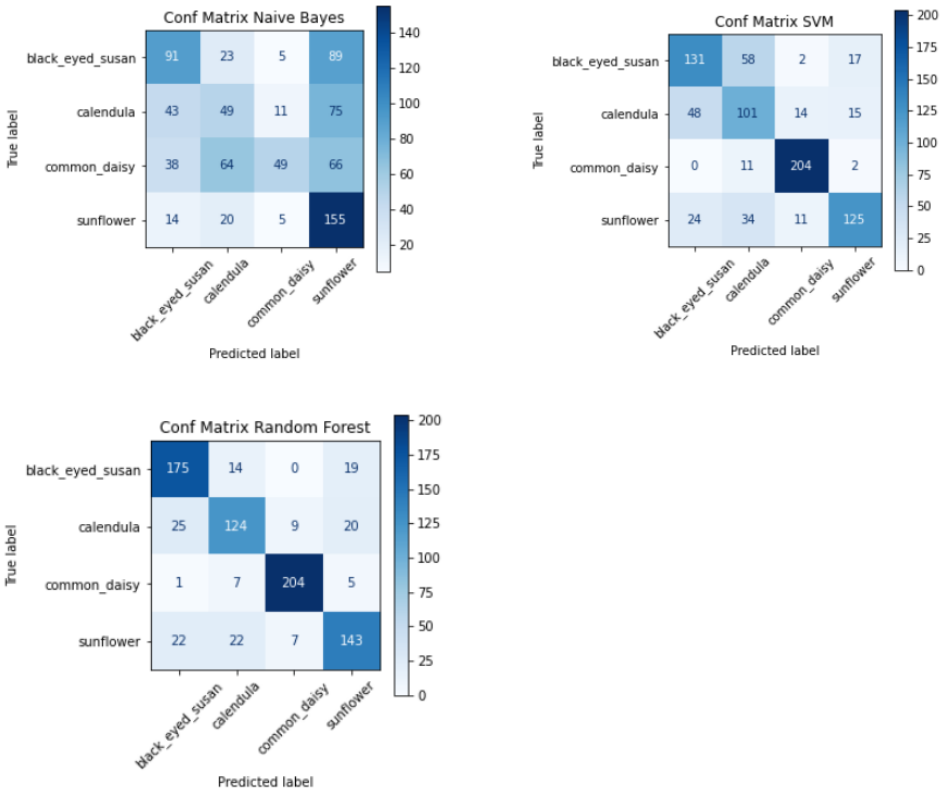
Random Forest

	precision	recall	f1-score	support
black_eyed_susan	0.77	0.82	0.80	208
calendula	0.74	0.69	0.71	178
common_daisy	0.93	0.95	0.94	217
sunflower	0.80	0.78	0.79	194
accuracy			0.82	797
macro avg	0.81	0.81	0.81	797
weighted avg	0.82	0.82	0.82	797

Gambar 5. Ringkasan Nilai Presisi, Recall, dan F1-Score Model Klasifikasi

Untuk menunjukkan rincian agar dapat memahami kinerja model klasifikasi yang dihasilkan, Penulis mengunakan confusion matrix karena alat ini digunakan secara luas untuk tugas klasifikasi. [14] Confusion matrix adalah ringkasan hasil prediksi pada suatu masalah klasifikasi. Jumlah prediksi benar dan salah dirangkum dengan nilai hitungan dan dipecah oleh masing-masing kelas. Confusion matrix menunjukkan bagaimana model klasifikasi bingung ketika membuat prediksi sehinga terlihat kesalahan yang dibuat oleh classifier dan jenis kesalahan yang dibuat.





Gambar IV. Confusion Matrix untuk Model Klasifikasi yang Digunakan

IV. Kesimpulan

Dalam penelitian ini dilakukan perbandingan algoritma pembelajaran mesin untuk klasifikasi citra bunga mengunakan algoritma Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Linear Discriminant Analysis, Naive Bayes, Support Vector Machine. Sebelum pembuatan model klasifikasi, terhadap citra bunga dilakukan ekstraksi fitur terlebih dahulu dengan teknik color histogram. Percobaan dilakukan menggunakan citra dengan bentuk yang mirip yang berasal dari kelompok bunga aster dengan jenis sunflower, calendula, black eyed susan, daisy. Evaluasi model klasifikasi menggunakan Confusion Matrix dan 10-fold cross validation. Skor akurasi tertinggi diperoleh dari model Random Forest sebesar 82%, SVM 70%, logistic regression 70%, decision tree 70%, LDA 66%, KNN 64%, dan paling kecil yaitu naïve bayes sebesar 43%. Penelitian selanjutnya bisa dilakukan percobaan menggunakan teknik ektraksi fitur bentuk atau tekstur untuk mengetahui bagaimana kinerja model klasifikasi. Untuk meningkatkan kinerja model klasifikasi bisa juga dengan menggunakan hyperparameter tuning.

Daftar Pustaka

[1] F. Muwardi *et al.*, “Pengolahan Citra Dan Pengklasifikasi Jarak,” *J. Ilmu Tek. Elektro Komput. dan Inform.*, vol. 3, no. 2, pp. 124–131, 2017.

[2] P. Rosyani and Oke Hariansyah, “Pengenalan Citra Bunga Menggunakan Segmentasi Otsu Treshold dan Naïve Bayes,” *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 1–7, 2020, doi: 10.30864/jsi.v15i1.304.

[3] H. Almogdady, S. Manaseer, and H. Hiary, “A flower recognition system based on image processing and neural networks,” *Int. J. Sci. Technol. Res.*, vol. 7, no. 11, pp. 166–173, 2018.

[4] N. Dayanand Lal, D. S. Sahana, R. C. Veena, S. H. Brahmananda, and D. S. Sakkari, “Image classification of the flower species identification using machine learning,” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, pp. 995–1007, 2019.

[5] P. Sharma, A. Aggarwal, A. Gupta, and A. Garg, *Leaf identification using HOG, KNN, and neural networks*, vol. 56. Springer Singapore, 2019.

[6] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh, "Comparing LBP, HOG and Deep Features for Classification of Histopathology Images," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, pp. 1–7, 2018, doi: 10.1109/IJCNN.2018.8489329.

[7] N. Kurnia Ningrum and E. Sasmita, "Ekstraksi Warna Berdasarkan Rgb Untuk Menentukan Tingkat Kematangan Daun Tembakau," *Udinus Jl. Imam Bonjol No*, vol. 207, p. 50131, 2015.

[8] A. R. Gala, "KLASIFIKASI APEL BERBASIS CITRA PENDAHULUAN Deteksi dan pengenalan pola pada citra sangat luas dan banyak dikembangkan dengan sejumlah pendekatan selama bertahun-tahun . Pengenalan pola adalah metode yang bekerja untuk menemukan pola pada data yang menunj," pp. 296–304.

[9] M. Poorani, T. Prathiba, and G. Ravindran, "Integrated Feature Extraction for Image Retrieval," no. November, 2017.

[10] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network," *Eurasip J. Image Video Process.*, vol. 2019, no. 1, 2019, doi: 10.1186/s13640-019-0417-8.

[11] S. Sural, G. Qian, and S. Pramanik, "Segmentation and Histogram Generation using HSV," pp. 589–592, 2002, [Online]. Available: http://www.cse.msu.edu/~pramanik/research/papers/papers/icip.hsv.pdf.

[12] B. S. Anami, S. S. Nandyal, and A. Govardhan, "A Combined Color, Texture and Edge Features Based Approach for Identification and Classification of Indian Medicinal Plants," *Int. J. Comput. Appl.*, vol. 6, no. 12, pp. 45–51, 2010, doi: 10.5120/1122-1471.

[13] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. April, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[14] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.