VERIFIKASI DOKUMEN CETAK MENGGUNAKAN METODE *EDGE DETECTION*-GLCM DAN K-*MEAN CLUSTERING*

Florentina Tatrin Kurniati[1], Roy Rudolf Huizen[2]
[1,2]Fakultas Teknologi Informasi, Institut Teknologi dan Bisnis STIKOM Bali
Email: [1]florentina.tatrin@gmail.com; [2]royrudolf.usm@gmail.com

Abstrak

Dokumen cetak merupakan data digital yang tervisualisasi di kertas. Dokumen cetak mudah dibuat sekaligus dipalsukan. Cara mengetahui keasliannya dengan verifikasi ciri. Setiap dokumen cetak terdapat ciri unik berasal dari printer yang digunakan. Mengembangkan metode untuk menentukan keaslian dokumen cetak dari printer jenis dan tipe sama menjadi tantangan untuk diselesaikan pada penelitian ini. Proses verifikasi menggunakan tahapan akuisisi, ekstraksi, pengenalan pola dan hasil verifikasi. Data uji menggunakan 6 printer jenis dan tipe sama, masing-masing menggunakan 3 lembar sampel, untuk karakter dipilih 8. Hasilnya pengujian menunjukkan nilai EER (equal error rate) sebesar 0,18 dari nilai tersebut digunakan menentukan nilai ambang verifikasi yaitu 80% dari nilai kedekatan. Verifikasi dengan kedekatan diatas 80% berarti dokumen cetak dinyatakan asli (accepted) sedangkan kedekatan dibawah 80% berarti dokumen cetak dinyatakan palsu (rejected). Pengujian menggunakan 20 varian dokumen cetak diperoleh nilai kedekatan tertinggi 97,7% dan terendah 84,1%. Berdasarkan pengujian diperoleh suatu kesimpulan bahwa meskipun pemalsuan menggunakan printer dengan jenis dan tipe yang sama, model mampu memilah dan menentukan suatu dokumen cetak asli ataupun palsu.

Kata Kunci : Verifikasi dokumen cetak, Edge Detection-GLCM, K-Mean Clustering

1. PENDAHULUAN

Masa sekarang ini dokumen digital telah digunakan secara luas, tetapi penggunaan dokumen cetak belum ditinggalkan [1]. Dokumen cetak merupakan data digital yang tervisualisasi di kertas menggunakan alat cetak. Berbagai dokumen cetak mudah dibuat, karena perkembangan teknologi alat cetak dan tools yang semakin maju dengan harga yang relatif murah. Kondisi baik ini sekaligus menjadi tantangan, karena potensi pemalsuan semakin besar [2]. Duplikasi atau pemalsuan suatu dokumen yang menggunakan alat cetak berkualitas tinggi, hasil cetaknya sulit dibedakan secara visual antara dokumen cetak asli dengan palsu meskipun pemalsuan menggunakan printer jenis dan tipe yang berbeda [3]. Kualitas tinggi dari alat cetak akan menghasilkan dokumen cetak palsu cenderung identik dengan dokumen cetak asli, membedakan secara visual sulit dilakukan. Perlu analisis yang mendalam untuk mengetahui dan membedakan kedua dokumen cetak tersebut, diantaranya menggunakan pengolahan gambar digital.

Pada dasarnya setiap jenis dan tipe printer mempunyai ciri unik, dengan analisis pengolahan gambar digital perbedaan atau cirinya dapat diketahui. Hal ini akan mempermudah proses verifikasi untuk menentukan suatu dokumen cetak asli atau palsu. Pemalsuan dengan menggunakan printer dengan jenis dan tipe yang berbeda dapat diketahui berdasarkan ciri dari jenis dan tipenya. Sedangkan jika pemalsuan menggunakan printer dengan jenis dan tipe sama, verifikasi untuk memilah dokumen cetak asli atau palsu sulit dilakukan, dikarenakan ciri cenderung sama. Tantangan lainnya adalah digitalisasi dokumen, penggunaan scan dengan kualitas rendah, fitur yang dihasilkan akan terdapat banyak noise. Hal ini berdampak pada hasil verifikasi menjadi tidak akurat [1], [4], [5]. Cara meminimalkan pengaruh noise dengan meningkatkan kualitas scan [6]. Proses verifikasi untuk menentukan keaslian dokumen cetak dari perangkat dengan jenis dan tipe yang sama menjadi tantangan yang akan diselesaikan pada

penelitian ini. Untuk melakukan verifikasi diperlukan ciri unik dari perangkat cetak yaitu ketidaknormalan hasil cetak pada tingkat piksel [7], [8]. Berdasarkan hal tersebut perlu dikembangkan metode ekstraksi yang mampu memperoleh ciri unik guna menentukan keaslian dokumen cetak.

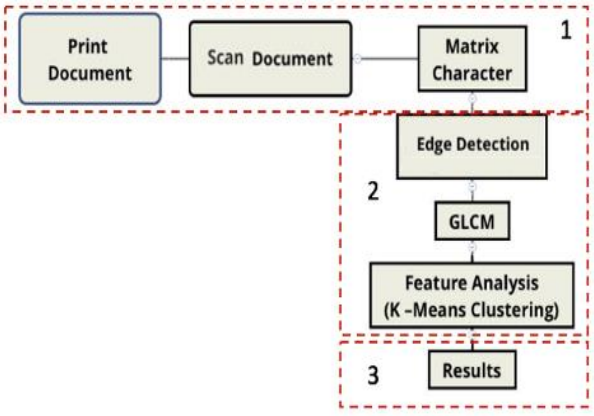
2. TINJAUAN PUSTAKA

Verifikasi adalah menguji keaslian dokumen cetak, prosesnya mencocokkan ciri dokumen cetak uji dengan ciri dokumen cetak asli [9]. Suatu dokumen cetak dinyatakan asli jika ciri dokumen cetak uji sama dengan dokumen cetak asli. Sebaliknya jika ciri berbeda maka dinyatakan palsu. Ciri diperoleh dengan mengekstraksi karakter yang terdapat pada dokumen cetak. Proses ekstraksi memunculkan suatu ciri unik. Ciri yang terdapat pada dokumen cetak bisa berasal dari kecacatan saat proses cetak (print defect) [7], [10]. Ketidaknormalan (*anomaly*) terjadi pada hasil cetak tingkat piksel. Print defect dapat digunakan sebagai ciri untuk menentukan asal dokumen cetak [11]. Print defect merupakan ketidaknormalan hasil cetak yang bersifat unik [6]. Anomaly print defect dapat diperoleh dari karakter-karakter yang membentuk lengkung, miring, garis ataupun berupa gambar [1], [11]. Print defect hasil cetak dapat diketahui dengan mengolah gambar menggunakan *edge detection*, selanjutnya dengan GLCM diekstraksi untuk memperoleh ciri.

Penggunaan metode *edge detection* digunakan untuk menemukan wilayah citra dengan intensitas yang berbeda. Intensitas suatu citra yang berbeda dapat terjadi pada periode curam ataupun landai. [12]. Hasil *edge detection* akan memperlihatkan adanya ketidaknormalan hasil cetak (*anomaly print defect*). Matrik *edge detection* sebagai distribusi piksel diekstraksi menggunakan metode GLCM (*Gray-Level Co-occurrence Matrices*) [13]. Proses ekstraksi ini untuk mengetahui distribusi piksel pada suatu matrik. Distribusi tersebut dapat menggunakan sudut 0^0, 45^0, 90^0, 135^0 [14]. Ciri yang diperoleh dianaliais menggunakan metode *K-mean Clustering* untuk diketahui pola masing-masing dokumen cetak [15]. Metode *K-mean clustering* dalam memodelkan suatu data ciri secara *unsupervised*. Proses pemodelannya mengelompokkan data ciri dengan sistem partisi. Pada tahap pertama pengelompokan, dipilih *centroid* acak dan untuk mengoptimalkannya *centroid* dihitung secara berulang, hingga diketemukan *centroid* dalam posisi stabil [16].

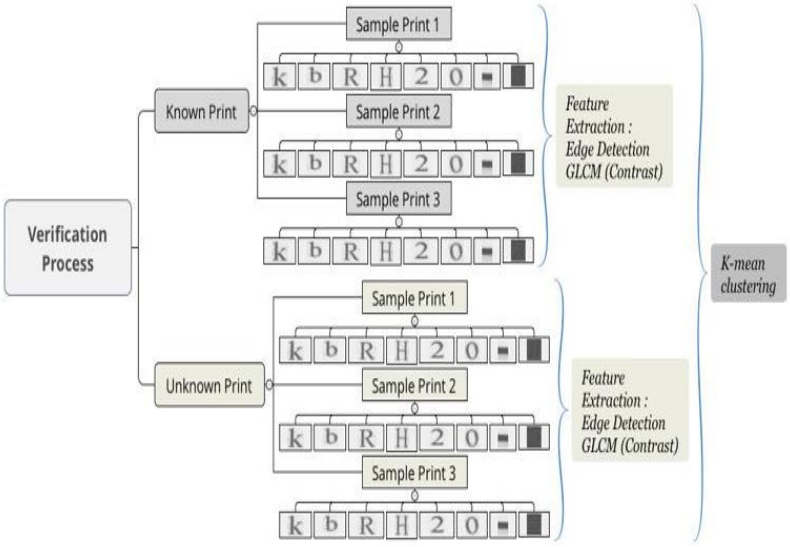
3. METODE PENELITIAN

Metodologi yang digunakan pada penelitian ini terdapat tiga tahapan, seperti ditunjukkan pada Gambar 1, yaitu (1). akuisisi, (2). Ekstraksi dan pola (3). hasil verifikasi. Berikut dijelaskan masing-masing bagian dari model tersebut;



Gambar 1. Model Verifikasi Dokumen Cetak

Akuisisi dokumen cetak terbagi menjadi 2 yaitu *known* dan *unknown*, untuk dokumen cetak asli dikategorikan sebagai *known* yaitu dokumen cetak yang telah diketahui sumber alat cetaknya. Sedangkan dokumen cetak uji dikategorikan sebagai *unknwon* yaitu dokumen cetak yang akan cocokkan dengan dokumen cetak *known*. Tahapan untuk menguji terlebih dahulu dengan digitalisasi dokumen cetak *known* dan *unknown* menggunakan perangkat *scanner* [11]. Karakter yang akan diekstraksi diambil dari masing-masing dokumen cetak (*known* dan *unknown*). Karakter yang dipilih mempunyai bentuk lengkung, miring dan lurus, seperti ditunjukkan pada Gambar 2. Pada gambar tersebut masing-masing known print dan *unknown* print dipilih karakter yang sama (*character dependent*).



Gambar 2. Akuisisi sampel karakter dokumen cetak asli (known) dan Uji (unknown)

Pengolahan gambar pada karakter menggunakan *edge detection*. Proses ini bertujuan untuk mengetahui adanya perubahan intensitas yang membentuk suatu tepi antar objek pada suatu karakter [17]. Tepi yang terbentuk terjadi karena adanya perubahan intensitas derajat keabuan yang besar dengan jarak yang pendek. Operasi untuk mendeteksi perubahan intensitas derajat keabuan menggunakan operator gradien derivatif pertama. Tahapan untuk memperoleh tepian pada suatu karakter dengan menggunakan operasi konvolusi. Operasi konvolusi pada matrik G_x dan G_y bertujuan memperoleh gradien dan mengetahui variasi intensitas [17] [18]. Fungsi kontinu f(x,y) pada gradien dinyatakan dalam vector ditunjukkan pada Persamaan (1) dan (2) ;

$$\nabla f(x,y) = \begin{bmatrix} G_x & G_y \end{bmatrix}^T \quad (1)$$

$$= \begin{bmatrix} \frac{\delta f}{\delta x} & \frac{\delta f}{\delta y} \end{bmatrix}^T \quad (2)$$

Berdasarkan arah gradien x dan y, maka untuk Gx dengan menggunakan pendekatan diferensial horizontal terhadap x maka fungsi f(x,y) diunjukkan pada Persamaan (3).

$$\frac{\partial f(x,y)}{\partial x} = f(x+1,y) - f(x,y) \quad (3)$$

matrik konvolusi ditunjukkan pada Persamaan (4) ;

$$G_x = \begin{bmatrix} 1 & -1 \end{bmatrix} \quad (4)$$

Gradien Gy diperoleh dengan pendekatan vertical terhadap y fungsi f(x,y) ditunjukkan pada Persamaan (5),

$$\frac{\partial f(x,y)}{\partial y} = f(x,1+y) - f(x,y) \quad (5)$$

matrik konvolusinya ditunjukkan pada Persamaan (6).

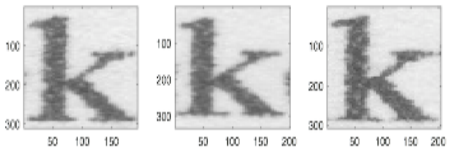
$$G_y = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (6)$$

Notasi ∇f menentukan kekuatan tepian yang lebih tajam dan arah, ditunjukkan pada Persamaan (7) dan (8).

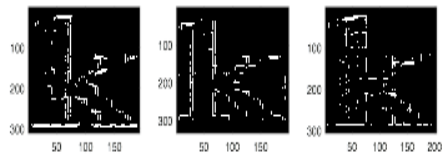
$$\nabla f = \sqrt{\left[\frac{\delta f}{\delta x}\right]^2 + \left[\frac{\delta f}{\delta y}\right]^2} = \sqrt{G_x^2 + G_y^2} \quad (7)$$

$$\text{Arah } \nabla f = tan^{-1}\left[\frac{G_y}{G_x}\right] \quad (8)$$

Sebagai ilustrasi suatu karakter yang akan diproses dengan *edge detection* ditunjukkan pada Gambar 3(a). Hasil pemrosesan ditunjukkan pada Gambar 3(b), dengan tepian ditunjukkan sebagai garis putih. Pada karakter, dengan *edge detection* terlihat adanya ketidaknormalan yang terbentuk saat proses cetak. Untuk mengetahui jumlah distribusi piksel pada karakter digunakan metode GLCM.



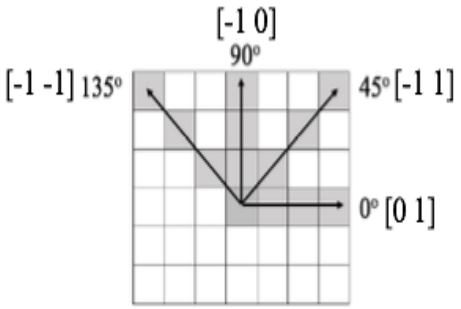
(a).



(b).

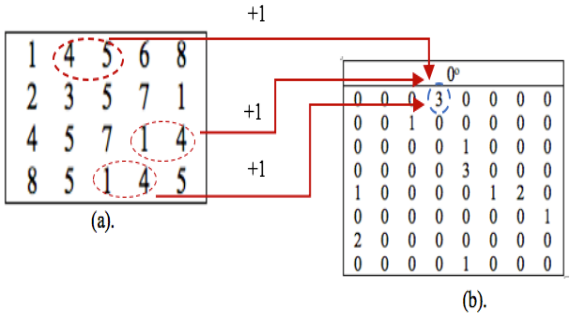
Gambar 3. (a). Karakter dari dokumen cetak (b). Hasil *edge detection*

Metode ini menghitung nilai keabuan dan frekuensi kemunculan suatu piksel pada citra. Proses ekstraksi GLCM berdasarkan distribusi statistik, dengan sudut 0º, 45º, 90º, 135º seperti ditunjukkan pada Gambar 4. Pada penelitian ini distribusi statistik yang digunakan untuk sudut 0º dan jarak d=1.



Gambar 4. Sudut dan arah

Pada Gambar 5.(a) diilustrasikan sebuah citra, dengan menggunakan sudut 0º dan jarak antar piksel adalah 1 (d=1). Maka setiap piksel dengan pasangannya (arah dan jarak) dihitung frekuensinya, hasil perhitungan tersebut menghasilkan matrik GLCM ditunjukkan Gambar 5.(b).



Gambar 5. (a). Matri suatu citra (b). GLCM dengan sudut 0^o

Pada penelitian ini analisa GLCM digunakan kontras, ditunjukkan pada Persamaan (10). Pemilihan kontras didasarkan pada variasi piksel, tanpa ada variasi maka nilai kontras akan bernilai 0. Semakin banyak variasi maka nilai kontras semakin besar. Matrik GLCM yang akan diekstraksi dinormalisasi, untuk memperoleh ciri kontras.

$$Kontras = \sum_{i}\sum_{j}(i-j)^2\, p(i,j) \tag{10}$$

Setiap karakter dari dokumen cetak dihitung nilai cirinya, dari keseluruhan ciri pada sampel membentuk suatu pola menggunakan algoritma *k-means clustering*. Metode ini dalam melakukan pengelompokan data, dengan menentukan terlebih dahulu jumlah claster. Setiap cluster terbentuk dengan partisi dengan centroid sebagai pusat cluster. Masing-masing data pada cluster dihitung ulang jarak antar objek, kemudian membentuk centroid baru. Proses tersebut berulang terus hingga diperoleh centroid (Persamaan 11), dengan nilai tidak berubah. Langkah-langkah pada K-mean clustering ditunjukkan sebagai berikut;

a. Menentukan jumlah claster
b. Menentukan nilai *centroid* secara acak

$$V_{ij} = \frac{1}{N_i} = \sum_{K=0}^{N_j} x_{kj} \tag{11}$$

c. Menghitung jarak data dan *centroid* ditunjukkan pada Persamaan

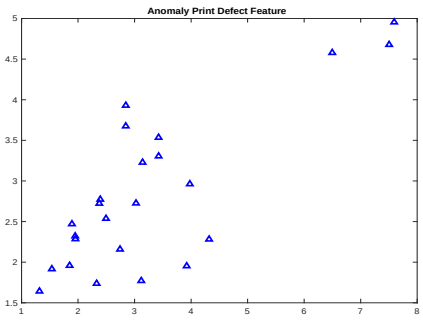
$$D = \sqrt{(x_i - a)^2 + (x_i - b)^2} \tag{12}$$

d. Pengelompokan untuk menentukan anggota cluster dengan menghitung jarak objek
e. Mengulang tahap dua hingga centroid tetap.

Hasil ekstraksi adalah *centroid* yang merupakan representasi dari ciri dokumen cetak. Dokumen cetak asli dan dokumen cetak uji nilai *centroid* dibandingkan untuk dihitung nilai prosentase kedekatan, semakin dekan berarti dokumen cetak asli sedangkan semakin jauh dokumen cetak dinyatakan palsu atau *accepted* dan *rejected*.

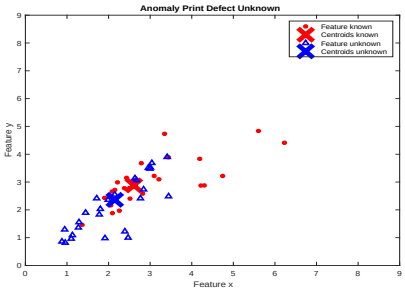
4. HASIL DAN PEMBAHASAN

Digitalisasi dokumen cetak menggunakan *scanner* dengan kualitas 600 dpi. Printer digunakan sebanyak 6, dengan jenis dan tipe sama. Dokumen cetak asli dipilih secara acak 4 printer dari 6 printer yang ada. Sedangkan dokumen cetak uji menggunakan 5 printer dari 6 printer yang ada. Setiap printer diambil sampel 3 lembar dokumen cetak dan 8 karakter huruf di setiap lembar. Alur ekstraksi menggunakan tahapan seperti Gambar 1 dan karakter yang diekstraksi pada Gambar 2. Hasil ekstraksi dengan *edge detection* dan GLCM, ciri kontras ditunjukkan pada Gambar 6.

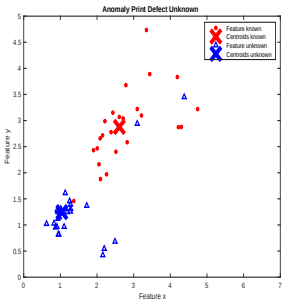


Gambar 6. Ekstraksi ciri dengan *edge detection* dan GLCM

Setiap karakter yang telah diekstraksi dihitung nilai centroid menggunaan *k-mean clustering*, hasilnya ditunjukkan Gambar 7. Pada Gambar 7(a). membandingkan dokumen cetak asli dengan dokumen cetak uji (*unknown* 1) hasilnya kedua centroid mempunyai kedekatan 84,5%. Ini berarti bahwa keduanya berasal dari printer yang sama (*accepted*). Pengujian berikutnya Gambar 7(b) memverifikasi dokumen cetak *unknown* 2 dengan dokumen asli hasilnya kedua *centroid* mempunyai kedekatan 39,7% ini berarti dokumen cetak tersebut dinyatakan palsu (*rejected).*



(a). verifikasi *accepted*



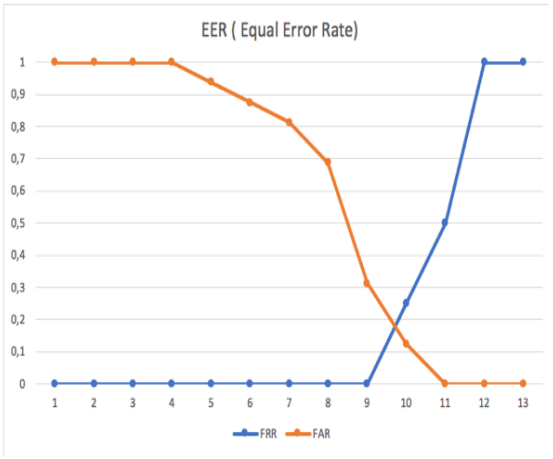
(b). verifikasi *rejected*

Gambar 7. Perbandingan ciri dokumen cetak

Verifikasi dokumen cetak menggunakan 20 varian, terdiri dari 4 dokumen cetak asli dan 5 dokumen cetak uji, pengujian hasilnya ditunjukkan pada Tabel 1. Pengujian diperoleh nilai EER *(equal error rate)* 0,18 yang merupakan pertemuan nilai FAR (*False Acceptance Rate*) dan FRR (*False Rejection Rate*). Hasilnya ditunjukkan pada Gambar 8, dari nilai EER ditentukan nilai ambang sebesar 80 %, yang digunakan untuk verifikasi hasil, ditunjukkan pada Tabel 2. Nilai kedekatan tertinggi adalah 97,7% sedangkan terendah adalah 84,1%.

Tabel 1. Pengujian kedekatan Dokumen Cetak

Dokumen Cetak Uji (*unknown Print)*	Dokumen Cetak Asli (*Known Print)* Nilai Kedekatan (%)			
	A	B	C	D
1	84,5	58,8	43,8	23,8
2	39,7	88,1	62,8	54,5
3	53,5	41,6	97,7	18,5
4	68,2	56,0	56,0	97,5
5	56,4	71,4	76,9	61,9



Gambar 8. EER Pengujian Model Verifikasi

Verifikasi dengan menggunakan nilai ambang 80% hasilnya ditujukkan pada Tabel 2. Setiap dokumen asli (*known print*) akan dibandingkan dengan lima dokumen cetak lainnya (*unknown print*). Hasil verifikasi menyatakan jika kedekatan diatas 80% dinyatakan *accept* yang berarti dokumen tersebut adalah asli. Sebaliknya jika nilai kedekatan antar centroid dibawah 80% maka dokumen cetak tersebut *rejected* atau dinyatakan palsu. Keseluruhan hasil pengujian terlihat pada Tabel 2.

Tabel 2. Hasil Verifikasi

Dokumen Cetak Asli	Dokumen Cetak Uji	**K-Mean**	
		Nilai Kedekatan (%)	**Hasil Verifikasi**
A	1	84,5	**Accept**
	2	39,7	Rejected
	3	53,5	Rejected
	4	68,2	Rejected
	5	56,4	Rejected
B	1	58,8	Rejected.
	2	88,1	**Accept**
	3	41,6	Rejected.
	4	56,0	Rejected
	5	71,4	Rejected.
C	1	43,8	Rejected
	2	62,8	Rejected
	3	97,7	**Accept**
	4	56,0	Rejected
	5	76,9	Rejected
D	1	23,8	Rejected
	2	54,5	Rejected
	3	18,5	Rejected
	4	97,5	**Accept**
	5	61,9	Rejected

5. KESIMPULAN

Berdasarkan hasil pengujian kesimpulan yang diperoleh adalah sebagai Model verifikasi dengan menggunakan *edge detection*-GLCM dan K-mean clustering diperoleh nilai EER sebesar 0,18. Nilai ambang optimum diperoleh 80% dengan nilai kedekatan tertinggi adalah 97,7% dan terendah adalah 84,1%. Pengujian menggunakan 20 varian dokumen cetak dari printer jenis dan tipe sama, hasilnya dapat digunakan untuk memilah dokumen cetak asli dan palsu.

DAFTAR PUSTAKA

[1] J. Gebhardt, M. Goldstein, F. Shafait, and A. Dengel, "Document Authentication Using Printing Technique Features and Unsupervised Anomaly Detection," *2013 12th Int. Conf. Doc. Anal. Recognit.*, pp. 479–483, Aug. 2013, doi: 10.1109/ICDAR.2013.102.

[2] R. Shao and E. J. Delp, "Forensic Scanner Identification Using Machine Learning," *Proc. IEEE Southwest Symp. Image Anal. Interpret.*, vol. 2020-March, pp. 1–4, 2020, doi: 10.1109/SSIAI49293.2020.9094618.

[3] D. G. Kim, J. U. Hou, and H. K. Lee, "Learning deep features for source color laser printer identification based on cascaded learning," *Neurocomputing*, 2019, doi: 10.1016/j.neucom.2019.07.084.

[4] Y. Wu, X. Kong, X. You, and Y. Guo, "Printer forensics based on page document's geometric distortion," pp. 2909–2912, 2009.

[5] M. Tsai and J. Liu, "Digital Forensics for Printed Source Identification," pp. 2347–2350, 2013.

[6] S. Elkasrawi and F. Shafait, "Printer Identification using Supervised Learning for Document Forgery Detection," *Doc. Anal. Syst. (DAS), IAPR Int. Work.*, 2014, doi: 10.1109/DAS.2014.48.

[7] F. T. Kurniati, A. J. Santoso, and Suyoto, "Printer Forensik Untuk Identifikasi Dokumen Cetak," *Semin. Nas. Teknol. Inf. dan Multimed.*, pp. 6–8, 2015.

[8] M. Bibi, A. Hamid, M. Moetesum, and I. Siddiqi, "Document Forgery Detection using Printer Source Identification—A Text-Independent Approach," pp. 7–12, 2019, doi: 10.1109/icdarw.2019.70134.

[9] O. Mayer and M. C. Stamm, "Forensic Similarity for Digital Images," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1331–1346, 2020, doi: 10.1109/TIFS.2019.2924552.

[10] A. C. Valente *et al.*, "Print defect mapping with semantic segmentation," *Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020*, pp. 3540–3548, 2020, doi: 10.1109/WACV45572.2020.9093470.

[11] F. T. Kurniati and R. R. Huizen, "Anomali Fitur Dokumen Cetak Untuk Verifikasi Di Printer Forensik," pp. 820–825, 2017.

[12] J. S. Owotogbe, T. S. Ibiyemi, and B. A. Adu, "Edge Detection Techniques on Digital Images - A Review," *Int. J. Innov. Sci. Res. Technol.*, vol. 4, no. 11, pp. 329–332, 2019.

[13] A. Hamid, M. Bibi, I. Siddiqi, and M. Moetesum, "Historical manuscript dating using textural measures," *Proc. - 2018 Int. Conf. Front. Inf. Technol. FIT 2018*, pp. 235–240, 2019, doi: 10.1109/FIT.2018.00048.

[14] R. Ghosh, C. Panda, and P. Kumar, "Handwritten Text Recognition in Bank Cheques," *2018 Conf. Inf. Commun. Technol. CICT 2018*, 2018, doi: 10.1109/INFOCOMTECH.2018.8722420.

[15] T. Gupta and S. P. Panda, "Clustering Validation of CLARA and K-Means Using Silhouette DUNN Measures on Iris Dataset," *Proc. Int. Conf. Mach. Learn. Big Data,*

Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 10–13, 2019, doi: 10.1109/COMITCon.2019.8862199.

[16] N. Sapkota, A. Alsadoon, P. W. C. Prasad, A. Elchouemi, and A. K. Singh, "Data Summarization Using Clustering and Classification: Spectral Clustering Combined with k-Means Using NFPH," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 146–151, 2019, doi: 10.1109/COMITCon.2019.8862218.

[17] J. X. Zhao and M. Y. Liu, "A color HSV image edge detection method based on gradient extreme value," *Proc. - 2008 2nd Int. Symp. Intell. Inf. Technol. Appl. IITA 2008*, vol. 3, pp. 381–384, 2008, doi: 10.1109/IITA.2008.9.

[18] Y. Y. Zheng, J. L. Rao, and L. Wu, "Edge detection methods in digital image processing," *ICCSE 2010 - 5th Int. Conf. Comput. Sci. Educ. Final Progr. B. Abstr.*, pp. 471–473, 2010, doi: 10.1109/ICCSE.2010.5593576.