

Clustering Citra Sertifikat Prestasi Berdasarkan Kemiripan Visual Menggunakan SIFT dan DBSCAN

Maulidya Prastita Syah¹, Wahyu Syaifullah Jauharis Saputra², Alfian Rizaldy Pratama³

^{1,2,3}*Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur*

¹22083010039@student.upnjatim.ac.id

²wahyu.s.j.saputra.if@upnjatim.ac.id

³alfan.fasilkom@upnjatim.ac.id

Corresponding author email: 22083010039@student.upnjatim.ac.id

ABSTRAK

Abstrak— Proses verifikasi dokumen sertifikat pada jalur Seleksi Nasional Berdasarkan Prestasi (SNBP) umumnya masih dilakukan secara manual, sehingga memerlukan waktu yang lama dan berpotensi menimbulkan ketidakkonsistenan dalam penilaian. Penelitian ini bertujuan untuk mengelompokkan citra sertifikat prestasi berdasarkan kemiripan visual guna mempermudah identifikasi pola desain yang serupa. Pendekatan yang digunakan melibatkan ekstraksi fitur visual dengan metode Scale Invariant Feature Transform (SIFT) yang mampu mengenali ciri lokal citra secara stabil terhadap perubahan rotasi, pencahayaan, dan skala. Hasil ekstraksi kemudian direduksi dimensinya menggunakan Principal Component Analysis (PCA) untuk memperoleh representasi yang lebih ringkas, sebelum dilakukan pengelompokan menggunakan algoritma Density Based Spatial *Clustering* of Applications with Noise (DBSCAN). Eksperimen dilakukan terhadap sekitar 7.000 citra sertifikat dalam berbagai format digital, yaitu JPG, PNG, dan PDF. Hasil pengelompokan menunjukkan terbentuknya sekitar 33 kluster utama dengan sejumlah citra yang dikategorikan sebagai outlier. Visualisasi memperlihatkan distribusi citra yang membentuk kelompok padat sesuai kemiripan elemen visual seperti logo, cap, dan tata letak teks. Hasil yang diharapkan dapat menjadi dasar bagi penelitian lanjutan dalam identifikasi kredibilitas sertifikat berbasis analisis pola visual secara otomatis.

Keywords: SIFT, PCA, DBSCAN, *clustering*, sertifikat prestasi

I. PENDAHULUAN

Bagian Perkembangan pemanfaatan dokumen digital dalam proses seleksi akademik, administrasi, dan rekrutmen telah meningkatkan kebutuhan akan sistem verifikasi dokumen yang cepat, konsisten, dan akurat. Proses verifikasi sertifikat prestasi masih banyak dilakukan secara manual dengan memeriksa elemen visual dokumen seperti tata letak, struktur teks, dan desain sertifikat [1]. Proses ini rentan terhadap inkonsistensi penilaian, kesalahan manusia, serta memerlukan waktu yang cukup panjang mengingat besarnya jumlah sertifikat yang harus diverifikasi setiap tahun [2]. Selain itu, variasi desain sertifikat yang sangat beragam antar lembaga menyebabkan proses identifikasi kemiripan visual semakin kompleks, terutama ketika desain yang berbeda memiliki struktur elemen yang hampir serupa [3].

Citra sertifikat sebagai dokumen visual umumnya memiliki karakteristik yang khas, seperti logo, cap, garis tepi, dan pola tata letak. Analisis citra dokumen digital sendiri telah menjadi bidang yang berkembang pesat dalam beberapa tahun terakhir. Menurut penelitian [4] melakukan survei komprehensif mengenai document *image analysis and recognition*, teknologi pemrosesan citra dokumen modern telah beralih dari pendekatan

berbasis aturan menuju metode ekstraksi fitur dan *machine learning* untuk menangani keragaman struktur visual dokumen. Ekstraksi fitur menjadi elemen fundamental dalam proses ini karena memungkinkan sistem untuk merepresentasikan dokumen secara numerik berdasarkan informasi visual yang dapat dibandingkan satu sama lain.

Salah satu pendekatan populer dalam ekstraksi fitur adalah *Scale Invariant Feature Transform* (SIFT), yang dikenal mampu mengidentifikasi titik-titik kunci lokal yang stabil terhadap perubahan skala, rotasi, dan pencahayaan [5]. Penelitian [6] menjelaskan mengenai pencocokan logo dokumen menunjukkan bahwa SIFT mampu menghasilkan kesesuaian fitur yang andal pada berbagai kondisi citra dan efektif untuk tugas seperti pencarian dan identifikasi dokumen berbasis logo. Kestabilan SIFT terhadap transformasi membuatnya unggul untuk menganalisis dokumen yang memiliki variasi desain namun tetap mempertahankan pola visual tertentu.

Selain ekstraksi fitur, proses pengelompokan (*clustering*) citra juga menjadi fokus penting dalam berbagai penelitian yang bertujuan untuk memetakan kemiripan visual. Pendekatan *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) merupakan salah satu algoritma yang banyak digunakan karena kemampuannya untuk menemukan pola berbasis kepadatan (*density based*) tanpa memerlukan penentuan jumlah kluster sejak awal serta mampu mengidentifikasi noise atau *outlier* secara otomatis. Dalam penelitian [7] yang membahas terkait deteksi *copy move forgery*, kombinasi SIFT dan DBSCAN terbukti efektif dalam mengelompokkan area citra yang memiliki kemiripan fitur signifikan, terutama ketika digabungkan dengan teknik pendukung untuk memperkuat segmentasi awal. Keberhasilan pendekatan tersebut menunjukkan bahwa integrasi metode SIFT dan DBSCAN memiliki potensi besar untuk berbagai aplikasi lain yang bergantung pada kemiripan visual, termasuk pengelompokan dokumen sertifikat.

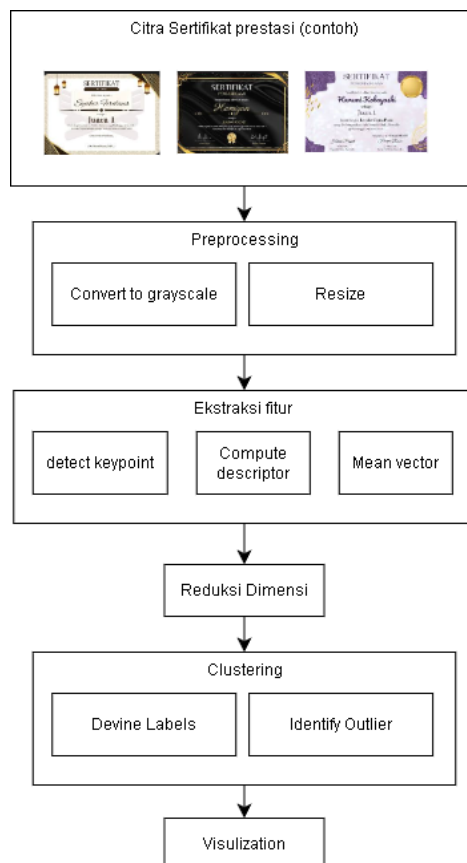
Meskipun berbagai penelitian telah dilakukan pada domain analisis citra dokumen, studi yang secara khusus memanfaatkan SIFT dan DBSCAN untuk mengelompokkan sertifikat prestasi berdasarkan pola visualnya masih sangat terbatas. Padahal, pengelompokan sertifikat dapat membantu proses verifikasi dengan mengidentifikasi pola kemiripan desain secara otomatis, mempercepat proses pemeriksaan dokumen, serta memberikan pemahaman mengenai distribusi dan variasi desain sertifikat pada skala besar. Melalui pengelompokan visual, proses verifikasi dapat dibantu dengan mengenali sertifikat yang berasal dari desain yang sama atau mirip, sehingga operator dapat memeriksa dokumen berdasarkan kelompok desain yang lebih terstruktur. Selain itu, pendekatan ini memungkinkan pemetaan distribusi visual pada dataset skala besar untuk memahami pola variasi desain yang muncul secara alami di dunia nyata.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan pendekatan berbasis ekstraksi fitur SIFT dan algoritma DBSCAN untuk mengelompokkan citra sertifikat prestasi dalam berbagai format digital. Ekstraksi fitur SIFT digunakan untuk memperoleh representasi lokal yang stabil dari setiap citra, untuk mereduksi kompleksitas fitur agar lebih efisien diproses menggunakan PCA, sementara DBSCAN dipilih karena kemampuannya mendeteksi kelompok berdasarkan kepadatan data tanpa memerlukan jumlah kluster awal. Pendekatan terpadu ini memungkinkan pemetaan sertifikat ke dalam kelompok visual yang konsisten, meskipun dataset terdiri dari ribuan citra dengan karakteristik yang sangat bervariasi. Hasil penelitian diharapkan dapat menghasilkan struktur kluster yang lebih informatif, membantu proses verifikasi dokumen dalam jumlah besar, dan menyediakan landasan teknis bagi pengembangan sistem deteksi atau penilaian kredibilitas dokumen secara otomatis di masa mendatang.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan data berupa kumpulan sertifikat prestasi mahasiswa jalur Seleksi Nasional Berdasarkan Prestasi (SNBP). Sertifikat prestasi ini diperoleh dari sumber resmi melalui panitia seleksi yang berpotensi mengandung variasi kualitas, resolusi, dan format. Format berkas sertifikat yang digunakan dalam penelitian ini mencakup PDF, JPG, dan PNG, sehingga memerlukan proses konversi dan penyeragaman terlebih dahulu sebelum diproses lebih lanjut. Sertifikat prestasi tersebut memiliki beragam tampilan visual, mulai dari logo instansi penerbit, cap lembaga, hingga elemen dekoratif lain yang menjadi ciri khas dokumen. Keberagaman inilah yang menjadikan penelitian ini menantang sekaligus penting, karena sistem otomatis harus mampu mengenali sertifikat yang benar-benar kredibel di tengah variasi desain yang begitu luas.

Penelitian ini dilakukan melalui serangkaian tahapan sistematis untuk mengelompokkan citra sertifikat prestasi berdasarkan kemiripan visual menggunakan kombinasi metode SIFT, PCA, dan DBSCAN.



Gambar 1. Diagram alir

II.1 Preprocessing

Tahap pertama dimulai dari proses *preprocessing* citra, di mana seluruh berkas sertifikat dalam berbagai format JPG, PNG, dan PDF diseragamkan agar dapat diproses secara konsisten. Pada tahap ini, setiap citra dikonversi terlebih dahulu ke format *grayscale* untuk menyederhanakan informasi visual tanpa menghilangkan karakteristik penting berupa tekstur dan pola struktural yang diperlukan pada tahap ekstraksi fitur. Citra kemudian diperkecil secara proporsional untuk mengurangi beban komputasi sekaligus menjaga konsistensi ukuran antar gambar. Khusus untuk format PDF, halaman pertama dirender menjadi gambar menggunakan beberapa tingkat DPI untuk mencegah *decompression bomb* dan memastikan file besar tetap dapat diproses dengan aman.

Preprocessing ini penting karena dataset sertifikat memiliki format, ukuran, dan resolusi yang sangat bervariasi, sehingga penyeragaman menjadi langkah dasar agar proses komputasi selanjutnya tetap efisien dan stabil.

II.2 Ekstraksi Fitur

Tahap berikutnya adalah ekstraksi fitur menggunakan metode *Scale Invariant Feature Transform* (SIFT). Algoritma SIFT merupakan algoritma dalam bidang visi komputer yang berfungsi untuk mendeteksi serta menggambarkan fitur lokal pada citra gambar [8]. Algoritma ini dipatenkan di Kanada oleh *University of British Columbia* dan diterbitkan oleh David Lowe pada tahun 1999. SIFT dipilih karena kemampuannya yang unggul dalam menangkap ciri-ciri lokal pada citra secara tahan terhadap perubahan rotasi, skala, dan intensitas pencahayaan karakteristik yang sangat penting mengingat sertifikat prestasi sering memiliki variasi orientasi, kualitas hasil scan, atau perbedaan warna. SIFT bekerja dengan membangun representasi scale-space menggunakan Difference of Gaussian (DoG) untuk mendeteksi titik-titik ekstrem yang potensial sebagai keypoint. DoG dihitung dengan mengurangkan dua hasil Gaussian blur pada skala berbeda, yaitu [9]:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \times I(x, y) \quad (1)$$

di mana $G(x, y, k\sigma)$ adalah fungsi Gaussian dan $I(x, y)$ adalah citra input. Keypoint yang stabil kemudian dianalisis untuk menentukan orientasi dominan berdasarkan magnitudo dan arah gradien, yang dihitung dengan [9]:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (2)$$

$$\theta(x, y) = \tan^{-1} \left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \right) \quad (3)$$

Setelah orientasi ditentukan, SIFT membangun deskriptor 128 dimensi dengan menyusun histogram gradien pada blok 4×4 sel, masing-masing berisi 8 arah orientasi gradien. Dalam konteks penelitian ini, deskriptor-deskriptor dari seluruh keypoint dalam satu citra kemudian dirata-ratakan untuk menghasilkan satu vektor fitur berdimensi 128 yang mewakili karakteristik visual global sertifikat. Pendekatan *mean descriptor* ini memungkinkan citra direpresentasikan secara ringkas dan stabil, tanpa kehilangan struktur visual utama seperti bentuk border, layout teks, logo, garis, maupun ornamen. Dengan demikian, fitur yang dihasilkan tidak bergantung pada variasi warna atau isi teks, tetapi murni pada pola desain, sehingga lebih sesuai untuk proses pengelompokan berbasis kemiripan tampilan sertifikat.

II.3 Reduksi Dimensi

Setelah fitur diekstraksi, penelitian ini melanjutkan proses dengan *Principal Component Analysis* (PCA) untuk mereduksi dimensi fitur. PCA adalah salah satu teknik untuk pengurangan dimensi dengan teknik statistik standar yang dapat digunakan untuk mengurangi dimensi suatu set data. PCA adalah metode analisis multivariat berdasarkan transformasi linear yang sering digunakan untuk mengurangi dimensi data, untuk mengambil informasi penting dari data besar, dan untuk menganalisis struktur variabel [10]. PCA digunakan untuk menurunkan dimensi vektor 128 dimensi menjadi bentuk yang lebih ringkas tanpa kehilangan struktur variansi penting dalam data. Tahapan utama PCA dimulai dengan membentuk matriks kovarians dari data terstandarisasi X , yang dihitung menggunakan:

$$C = \frac{1}{n-1} X^T X \quad (4)$$

Selanjutnya PCA mencari pasangan *eigenvalue* dan *eigenvector* dari matriks kovarians tersebut melalui dekomposisi:

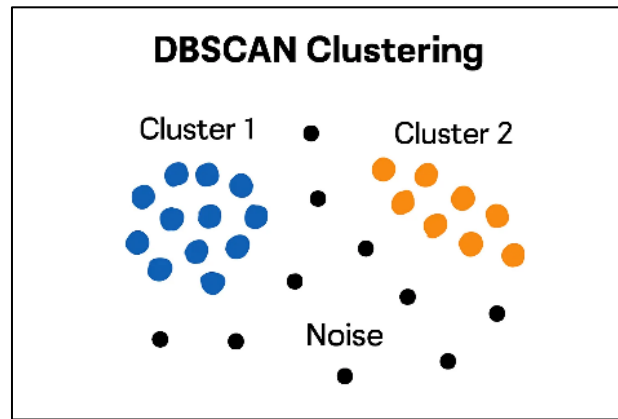
$$C_v = \lambda_v \quad (5)$$

di mana v adalah *eigenvector* yang merepresentasikan arah komponen utama, dan λ adalah *eigenvalue* yang menunjukkan besarnya variansi yang dijelaskan oleh komponen

tersebut. Komponen utama dengan nilai λ terbesar dipilih sebagai representasi data baru. Dalam penelitian ini, PCA digunakan untuk memproyeksikan vektor fitur SIFT berukuran 128 dimensi ke ruang berdimensi lebih rendah, baik untuk keperluan visualisasi dalam dua dimensi (PCA1–PCA2) maupun untuk mengurangi noise serta redundansi fitur sebelum proses *clustering* dilakukan. Reduksi dimensi ini tidak hanya menurunkan beban komputasi DBSCAN, tetapi juga membantu memperjelas struktur keterkelompokan citra, sehingga pola kemiripan desain sertifikat dapat diamati secara lebih jelas dan terstruktur dalam ruang fitur yang baru

II.4 Clustering

Tahap inti penelitian adalah proses pengelompokan menggunakan algoritma *Density Based Spatial Clustering of Applications with Noise* (DBSCAN). Metode DBSCAN merupakan suatu algoritma pengelompokan yang didasarkan pada kepadatan (density) data [11]. Algoritma ini dipilih karena memiliki sifat yang ideal untuk data visual sertifikat yang tidak memiliki pola klaster berbentuk bulat atau ukuran seragam.



Gambar 2. DBSCAN Clustering

DBSCAN bekerja berdasarkan kepadatan titik, sehingga mampu menemukan kelompok sertifikat yang sangat mirip satu sama lain tanpa memerlukan penentuan jumlah klaster di awal. Selain itu, DBSCAN secara otomatis menandai citra yang tidak memiliki kedekatan signifikan dengan kelompok manapun sebagai *outlier*, yang dalam konteks sertifikat dapat merepresentasikan desain unik, anomali, atau bentuk layout yang jarang ditemukan. DBSCAN bekerja dengan mendefinisikan *tetangga* suatu titik melalui radius ϵ (epsilon), yang disebut ϵ -neighborhood, secara matematis dituliskan sebagai:

$$N_{\epsilon}(p) = \{q | \text{dist}(p, q) \leq \epsilon\} \quad (6)$$

di mana $\text{dist}(p, q)$ adalah jarak Euclidean antar dua titik fitur. Sebuah titik dikategorikan sebagai *core point* jika jumlah titik di dalam radius tersebut memenuhi ambang minimal, dirumuskan sebagai:

$$|N_{\epsilon}(p)| \geq \text{minPts} \quad (7)$$

Konsep ini memungkinkan DBSCAN membentuk klaster berdasarkan hubungan *density reachability*, yaitu kondisi ketika sebuah titik q dapat dicapai dari titik inti p melalui jalur titik-titik yang saling berada dalam radius kepadatan yang sama.

Melalui mekanisme ini, DBSCAN mampu membentuk klaster dengan bentuk arbitrer, tidak terbatas pada struktur globular seperti pada K-Means. Pemilihan DBSCAN dalam penelitian ini didasarkan pada karakteristik citra sertifikat yang memiliki keragaman desain tinggi dan tidak terdistribusi secara merata pada ruang fitur SIFT, sehingga pendekatan berbasis kepadatan jauh lebih sesuai. DBSCAN juga secara otomatis menandai citra yang tidak memiliki kedekatan signifikan dengan kelompok mana pun sebagai *outlier*, yang dalam konteks verifikasi sertifikat dapat merepresentasikan desain

unik, format jarang digunakan, atau bentuk visual yang tidak sesuai dengan pola umum pada dataset.

Setelah proses pengelompokan selesai, seluruh label cluster disimpan dalam bentuk file CSV dan citra dikelompokkan ke dalam folder sesuai klasternya untuk mempermudah pemeriksaan manual. Hasil *clustering* juga diproyeksikan ke ruang dua dimensi menggunakan PCA untuk menampilkan peta persebaran visual, sehingga hubungan kedekatan antar kluster dapat diamati secara intuitif. Dengan menerapkan DBSCAN pada vektor fitur hasil SIFT yang telah direduksi dimensinya oleh PCA, penelitian ini menyediakan pendekatan yang fleksibel, adaptif terhadap bentuk kluster tidak beraturan, serta mampu menangani dataset citra sertifikat berukuran besar tanpa memerlukan jumlah kluster yang telah ditentukan sebelumnya.

III. HASIL DAN ANALISIS

III.1 *Preprocessing*

Tahap *preprocessing* menghasilkan data citra yang telah distandarkan dari berbagai format sumber, yakni JPG, PNG, dan PDF. Dataset awal terdiri dari 7.347 berkas, dan proses pemuatan citra berhasil dilakukan terhadap 7.345 berkas, sedangkan dua berkas gagal diproses. Seluruh citra yang berhasil dimuat kemudian dikonversi ke bentuk *grayscale* agar informasi visual lebih homogen dan bebas gangguan variasi warna. Setiap citra juga di-*downscale* secara proporsional dengan batas sisi maksimum 1600 piksel untuk mencegah *error* terkait ukuran berkas sangat besar pada PDF, termasuk peringatan *Decompression Bomb*. Untuk file PDF, sistem *rendering* menggunakan beberapa *fallback* DPI secara bertingkat (120, 96, dan 72) sehingga file beresolusi tinggi tetap dapat dibaca. Secara keseluruhan, tahap *preprocessing* berjalan dengan sukses dan menghasilkan kumpulan citra *grayscale* yang sudah distandarkan dan siap diekstraksi fiturnya. Contoh *preview* dari dua citra pertama menunjukkan bahwa proses konversi berlangsung konsisten tanpa distorsi, menghasilkan representasi visual awal yang layak untuk analisis lanjutan.

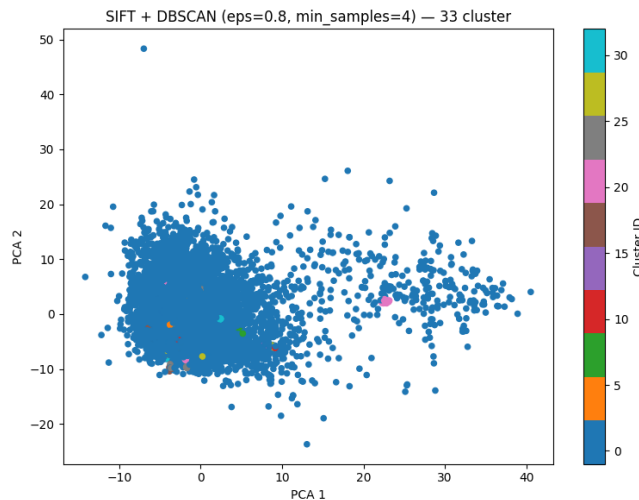
III.2 Ekstraksi Fitur

Pada tahap ekstraksi fitur, seluruh citra *grayscale* dianalisis menggunakan algoritma *Scale Invariant Feature Transform* (SIFT) untuk memperoleh representasi numerik yang stabil terhadap rotasi, skala, dan perubahan pencahayaan. Dari 7.345 citra yang berhasil diproses, semuanya berhasil menghasilkan fitur SIFT tanpa adanya *error* tambahan. Setiap citra menghasilkan sekumpulan keypoint dan deskriptor lokal, yang kemudian diringkas menjadi satu vektor fitur berukuran 128 dimensi melalui perhitungan rata-rata deskriptor. Vektor ringkas ini berfungsi sebagai identitas visual tiap citra dan menjadi dasar bagi proses *clustering*. Total waktu pemrosesan ekstraksi fitur mencapai sekitar 8.292 detik, yang wajar mengingat ukuran dataset sangat besar dan proses SIFT memang komputasional. Berdasarkan hasil tampilan citra sampel, fitur dapat diekstraksi dengan normal dan tidak ditemukan citra yang kosong atau rusak dalam kumpulan data yang berhasil diproses. Pada tahap ini, seluruh citra telah berhasil dikonversi ke dalam bentuk numerik sehingga dapat dianalisis secara komputasional melalui reduksi dimensi dan *clustering*.

III.3 Reduksi Dimensi

Setelah fitur SIFT diekstraksi, seluruh vektor 128 dimensi direduksi menggunakan *Principal Component Analysis* (PCA) menjadi dua komponen utama (PCA1 dan PCA2) untuk tujuan visualisasi dan pemahaman struktur data. Hasil

visualisasi PCA ditampilkan pada berikut, yang memperlihatkan persebaran seluruh 7.345 citra dalam ruang dua dimensi berdasarkan variansi terbesar dari fitur.

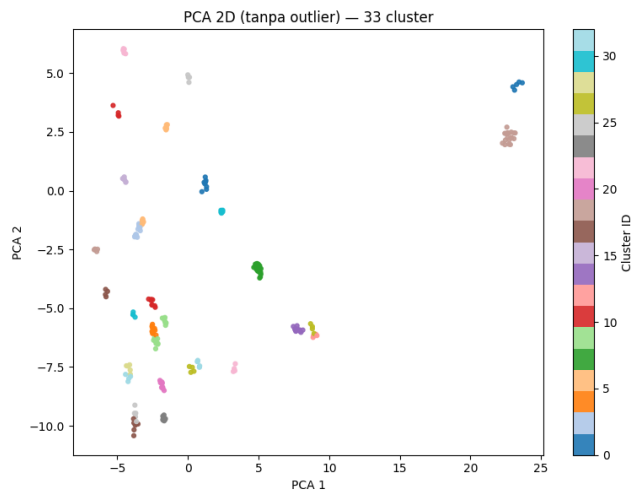


Gambar 3. Visualisasi Data dengan *outlier*

Gambar ini menunjukkan distribusi seluruh citra sertifikat dalam ruang PCA 2D. Warna biru menandakan *outlier* (label -1), sedangkan warna lainnya mewakili cluster kecil yang terbentuk. Terlihat bahwa lebih dari 95% data berada pada area biru, menandakan bahwa DBSCAN mengidentifikasi sebagian besar citra sebagai tidak memiliki kedekatan fitur yang cukup untuk membentuk kelompok padat. Persebaran yang sangat menyebar mengindikasikan tingginya variasi desain sertifikat, baik dari struktur ornamen, tata letak, hingga pola visual dominan. Visualisasi ini membantu memahami bagaimana citra-citra yang sangat berbeda saling menjauh satu sama lain dalam ruang fitur.

III.4 Clustering

Tahap utama penelitian adalah proses pengelompokan citra menggunakan DBSCAN. Dengan parameter $\text{eps} = 0.8$ dan minimal samples = 4, algoritma menghasilkan 33 cluster valid dengan total 239 citra anggota cluster, sementara 7.106 citra lainnya terdeteksi sebagai *outlier*. Hasil *clustering* divisualisasikan kembali dalam dua bentuk: (1) visualisasi PCA dengan seluruh data, dan (2) visualisasi PCA setelah menghapus *outlier* untuk melihat struktur cluster secara lebih jelas.



Gambar 4. Visualisasi tanpa *Outlier*

Setelah seluruh *outlier* dihapus, citra yang tersisa adalah 239 citra yang benar-benar membentuk cluster padat menurut DBSCAN. Visualisasi ini memperlihatkan bahwa setelah *outlier* dieliminasi, struktur klaster tampak lebih jelas, terpisah, dan rapat, menandakan bahwa citra-citra tersebut memiliki kemiripan visual kuat. Warna-warna yang berbeda menunjukkan ID cluster yang terbentuk. Dengan tidak adanya dominasi titik biru (*outlier*), gambar ini memperlihatkan bentuk alami cluster yang seharusnya dihasilkan model untuk kelompok citra yang benar-benar mirip.



Gambar 5. *Image Map*

Gambar ini memperlihatkan representasi PCA 2D yang titik-titiknya digantikan dengan thumbnail citra sertifikat sebenarnya. Peta ini memberikan gambaran visual yang lebih intuitif tentang bagaimana citra-citra dengan kemiripan desain berkelompok satu sama lain. Sertifikat dengan template yang sama, warna dominan serupa, dan struktur layout yang identik dapat terlihat membentuk kumpulan yang rapat. Sebaliknya, citra dengan desain berbeda tampak menyebar jauh. Visualisasi ini sangat berguna dalam memahami hubungan visual antar sampel tanpa harus membuka satu per satu berkas citra.

IV. KESIMPULAN

Hasil penelitian menunjukkan bahwa kombinasi metode preprocessing, ekstraksi fitur SIFT, reduksi dimensi PCA, dan pengelompokan DBSCAN mampu memberikan gambaran komprehensif mengenai struktur kemiripan visual dalam dataset sertifikat yang sangat beragam. Preprocessing berhasil menormalkan format dan kualitas citra, memungkinkan seluruh berkas diproses secara konsisten. Tahap SIFT kemudian berhasil merepresentasikan setiap citra dalam bentuk vektor fitur berdimensi 128 yang menggambarkan pola gradien lokal, sehingga variasi visual pada sertifikat dapat dianalisis menggunakan pendekatan numerik. Ketika fitur-fitur ini diproyeksikan ke ruang PCA 2 dimensi, tampak jelas bahwa dataset memiliki keragaman desain yang sangat tinggi, dibuktikan dengan penyebaran titik yang luas pada visualisasi awal serta dominasi besar titik-titik *outlier*. Hal ini mengindikasikan bahwa sebagian besar citra tidak memiliki kesamaan visual yang cukup kuat untuk membentuk cluster padat, meskipun semuanya merupakan sertifikat.

Setelah menerapkan DBSCAN, diperoleh 33 cluster valid dengan total 239 citra sebagai anggota cluster, sementara 7.106 citra lainnya dikategorikan sebagai *outlier*.

Visualisasi tanpa *outlier* memperlihatkan bahwa cluster-cluster yang terbentuk memiliki bentuk yang jelas dan terpisah, menandakan keberhasilan DBSCAN dalam menemukan kelompok citra yang benar-benar mirip secara visual. Peta PCA berbasis thumbnail semakin menguatkan validitas hasil *clustering*, karena citra-citra yang berada dalam cluster yang sama tampak memiliki template, pola, dan komposisi visual yang serupa. Secara keseluruhan, hasil analisis ini menunjukkan bahwa pendekatan yang digunakan dalam penelitian ini efektif untuk mengidentifikasi pola kemiripan desain sertifikat dalam dataset besar, serta memberikan dasar yang kuat untuk pengembangan sistem verifikasi otomatis berbasis kemiripan visual di masa mendatang.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada berbagai pihak yang telah mendukung terlaksananya penelitian ini. Penghargaan disampaikan kepada dosen pembimbing, yang telah memberikan arahan, bimbingan ilmiah, dan masukan berharga selama proses penelitian dan penyusunan artikel. Ucapan terima kasih juga ditujukan kepada Program Studi Sains Data serta seluruh pihak yang berkontribusi dalam penyediaan fasilitas dan lingkungan akademik yang mendukung pelaksanaan penelitian ini. Selain itu, penulis menyampaikan apresiasi kepada rekan-rekan yang telah memberikan bantuan teknis, diskusi konstruktif, serta motivasi selama proses pengerjaan. Penulis juga berterima kasih kepada Tim DISCOVER atas penyediaan template yang digunakan dalam penulisan artikel ini.

REFERENSI

- [1] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00948-z.
- [2] "Siaran Pers Nomor: 03/sipers/snpmb/II/2025 Penutupan Pendaftaran SNBP 2025 | Seleksi Nasional Penerimaan Mahasiswa Baru." Accessed: Oct. 16, 2025. [Online]. Available: <https://www.snpmb.id/blog/siaran-pers-penutupan-pendaftaran-snpb-2025>
- [3] N. Challa, A. Shende, and M. Mullanpudi, "Enhancing Document Verification Systems: A Review of Techniques, Challenges, and Practical Implementations," *International Journal of Computer Engineering and Technology (IJCET)*, vol. 15, no. 1, pp. 16–25, 2024, doi: 10.17605/OSF.IO/HVQ8E.
- [4] V. V. Arlazarov *et al.*, "Document image analysis and recognition: a survey," *Computer Optics*, vol. 46, no. 4, pp. 567–589, Jul. 2022, doi: 10.18287/2412-6179-CO-1020.
- [5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int J Comput Vis*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94/METRICS.
- [6] N. Akasha, "Logo Matching for Document Image Retrieval Using SIFT Descriptors," *Journal of Engineering Research and Application www.ijera.com*, vol. 7, no. 2, pp. 55–60, 2017, doi: 10.9790/9622-0702025560.
- [7] F. I. Rahma, E. Utami, and H. Al Fatta, "The Using of Gaussian Pyramid Decomposition, Compact Watershed Segmentation Masking and DBSCAN in Copy-Move Forgery Detection with SIFT," *2020 3rd International Conference on Information and Communications Technology, ICOLACT 2020*, pp. 325–330, Nov. 2020, doi: 10.1109/ICOLACT50329.2020.9332081.
- [8] R. Al Caruban, B. Sugiantoro, and Y. Prayudi, "ANALISIS PENDETEKSI KECOCOKAN OBJEK PADA CITRA DIGITAL DENGAN METODE ALGORITMA SIFT DAN HISTOGRAM COLOR RGB," *Cyber Security dan Forensik Digital*, vol. 1, no. 1, pp. 20–27, Jul. 2018, doi: 10.14421/csecurity.2018.1.1.1235.
- [9] L. Tang, S. Ma, X. Ma, and H. You, "Research on Image Matching of Improved SIFT Algorithm Based on Stability Factor and Feature Descriptor Simplification," *Applied Sciences 2022, Vol. 12, Page 8448*, vol. 12, no. 17, p. 8448, Aug. 2022, doi: 10.3390/APP12178448.
- [10] A. S. Ritonga and I. Muhandhis, "TEKNIK DATA MINING UNTUK MENGLASIFIKASIKAN DATA ULASAN DESTINASI WISATA MENGGUNAKAN REDUKSI DATA PRINCIPAL COMPONENT ANALYSIS (PCA)," *Jurnal Ilmiah Edutic : Pendidikan dan Informatika*, vol. 7, no. 2, pp. 124–133, May 2021, Accessed: Nov. 28, 2025. [Online]. Available: <https://journal.trunojoyo.ac.id/edutic/article/view/9247>
- [11] J. J. Psaila, "Image Enhancement for Improved Keypoint Detection and Matching Using Image Offsets with Adaptive Parameter Selection," *SSRN Electronic Journal*, Sep. 2022, doi: 10.2139/SSRN.4223290.