Klasifikasi Gender Berdasarkan Sidik Jari Menggunakan Principal Component Analysis dan Support Vector Machine

Gian Nathan Christyo Nugroho[1]
[1]*Informatika,Universitas Kristen Duta Wacana*
Jl. dr. Wahidin Sudirohusodo no. 5-25, Yogyakarta
gian.nathan@ti.ukdw.ac.id

***Abstract*—Mayoritas penelitian klasifikasi sidik jari menggunakan ciri-ciri seperti *core* dan *delta* sebagai basisnya. Sebelum mengekstraksi fitur-fitur sidik jari, berbagai tahapan preprosesing biasanya dilakukan terlebih dahulu. Penelitian ini berbeda dengan penelitian lain karena klasifikasi dilakukan langsung pada citra sidik jari tanpa melalui tahap preposesing yang detail dan hanya dilakukan pengubahan ukuran piksel menjadi 96x103. Ciri-ciri sidik jari tidak ditentukan secara manual, melainkan diekstraksi secara otomatis menggunakan metode *Principal Component Analysis* (PCA) yang menghasilkan 4200 ciri-ciri yang terbaik. Untuk alasan kesempurnaan fitur telah dilakukan normalisasi fitur menggunakan *StandardScaler*. Klasifikasi dari penlitian ini menggunakan metode *Support Vector Machine* (SVM) nonlinear dengan kernel *Polynomial*. Penelitian ini menggunakan 6000 sampel data dari *database* SOCOFing. Model ini memperoleh akurasi klasifikasi hingga 88,75%.**

***Kata Kunci*— sidik jari, klasifikasi gender, PCA, SVM**

I. PENDAHULUAN

Pada penelitian sidik jari, dapat dibagi menjadi dua cakupan bagian yaitu cakupan besar dan cakupan kecil dari sidik jari, yang dimana untuk penelitian ini akan berfokus pada cakupan kecil. Hal ini dikarenakan oleh beberapa faktor yaitu dataset yang digunakan berasal dari Kaggle. Dan juga tujuan akhir dari penelitian ini yaitu, untuk mencari tingkat akurasi dengan menggunakan metode klasifikasi SVM (*Support Vector Machine*) dan pengurangan dimensi dengan bantuan PCA (*Principal Component Analysis*).

Sebelum dilakukannya klasifikasi gender, akan dilakukan proses ekstraksi fitur dengan menggunakan teknik *Principal Component Analysis*. Penggunaan PCA sebagai teknik penentuan fitur-fitur yang dibutuhkan dinilai keefektifannya. Sebagai contoh, [1] menggunakan PCA dalam analisis gambar sidik jari untuk klasifikasi gender dan menunjukan tingkat akurasi hingga 95%. Demikian pula, [2] menggunakan PCA dalam penelitian mereka tentang identifikasi gender berbasis sidik jari dan mencapai tingkat akurasi keseluruhan sebesar 80% untuk pria dan 90% untuk wanita. Temuan ini memberikan bukti kuat akan kesesuaian PCA dalam mengklasifikasikan gender secara akurat berdasarkan fitur sidik jari.

Klasifikasi gender berdasarkan sidik jari telah banyak dilakukan. Beberapa penelitian menunjukkan hasil dari penggunaan *Support Vector Machine* yang menjanjikan untuk klasifikasi gender berbasis sidik jari. Ditunjukkan oleh [3] yang menggunakan *Support Vector Machine* (SVM) dengan fitur *Discrete Wavelet Transform* (DWT) dan mencapai tingkat klasifikasi 89% dan 91% dengan klasifikasi SVM yang berbeda. Bukti ini mendukung bahwa SVM adalah metode yang tepat dan efektif untuk klasifikasi sidik jari, terutama dalam konteks identifikasi gender.

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi gender berdasarkan sidik jari yang akurat dan efisien. Sistem tersebut akan menggunakan metode *Principal Component Analysis* (PCA) untuk mengurangi dimensionalitas data sidik jari dan metode *Support Vector Machine* (SVM) untuk melakukan klasifikasi.

A. Tinjauan Pustaka

Menurut penelitian [4], menyatakan bahwa klasifikasi gender dari sidik jari memainkan peran penting dalam mengidentifikasi gender seorang pelaku kejahatan dan mempersempit daftar tersangka. Penelitian sebelumnya tentang klasifikasi gender dari sidik jari telah berfokus pada berbagai fitur seperti *ridge density, ridge breadth, ridge count, ridge count asymmetry, pattern type concordance, Ridge Thickness to Valley Thickness Ratio (RTVTR), dan white lines count*.

Menurut penelitian [5], menyatakan bahwa klasifikasi sidik jari berdasarkan gender merupakan area penelitian yang signifikan dalam ilmu forensik. Pengembangan algoritma untuk klasifikasi gender berdasarkan fitur sidik jari memiliki potensi yang sangat bermanfaat dalam investigasi forensik. Salah satu tantangan dalam klasifikasi gender adalah proses ekstraksi fitur, yang bergantung pada kualitas gambar sidik jari dan akurasi perhitungan.

Menurut penelitian [6], menyatakan bahwa klasifikasi gender dari sidik jari telah mendapatkan perhatian yang signifikan dalam bidang biometrik karena potensinya dalam penyelidikan kejahatan, otentikasi, dan analisis statistik. *Ridge count* dan ukuran ujung jari telah diidentifikasi sebagai parameter utama untuk klasifikasi gender. Penelitian sebelumnya telah menunjukkan bahwa *ridge pattern* menunjukkan variasi yang signifikan secara statistik antara pria dan wanita, sehingga cocok untuk penentuan gender.

Menurut penelitian [7], menyatakan bahwa klasifikasi gender berdasarkan analisis sidik jari telah mendapatkan perhatian yang signifikan dalam beberapa tahun terakhir. Detektif dan agen keamanan sering menemukan sidik jari di tempat kejadian perkara dan dapat secara akurat menentukan gender dimana akan sangat membantu dalam mempersempit ruang pencarian. Dalam hal ekstraksi fitur, satu set fitur sidik

jari yang efisien untuk pengenalan gender. Fitur-fitur ini termasuk *ridge count, minutiae points, discrete cosine transform, entropy, local binary pattern, dan ridge thickness valley thickness ratio*. Kombinasi dari fitur-fitur ini mencapai akurasi pengenalan gender sebesar 91% ketika diterapkan pada setiap jari secara individual.

Akurasi klasifikasi gender dari sidik jari telah meningkat secara signifikan dalam beberapa tahun terakhir. Penelitian terbaru telah mencapai akurasi hingga 95%. Klasifikasi gender dari sidik jari memiliki potensi untuk digunakan dalam berbagai aplikasi, seperti penyelidikan kejahatan, otentikasi, dan analisis statistik.

II. LANDASAN TEORI

Berikut merupakan penjelasan untuk masing-masing teori dan metode yang digunakan pada penelitian ini.

A. Sidik Jari

Sidik jari adalah sistem otentikasi terbaik yang digunakan saat ini. Sistem biometrik ditemukan aplikasinya di bidang-bidang seperti paspor, perbankan, ponsel atau laptop, rumah, kantor, kartu, departemen forensik, dan lain-lain. Hal ini dikarenakan sidik jari memiliki karakteristik yang unik, permanen, dan sulit dipalsukan. Sidik jari setiap orang berbeda, bahkan pada kembar identik. Sidik jari juga tidak akan berubah selama hidup, kecuali karena cedera atau penyakit [8].

B. Faktor Biologis

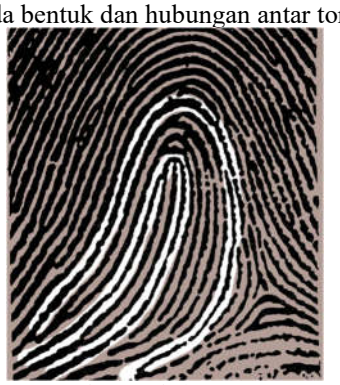
Latar belakang biologis pembentukan sidik jari masih belum sepenuhnya dipahami, dan ada beberapa teori yang diajukan untuk menjelaskan perkembangan pola sidik jari. Salah satu teori adalah hipotesis pelipatan, yang menyatakan bahwa sidik jari terbentuk melalui pelipatan lapisan basal epidermis. Hipotesis ini didukung oleh pemodelan matematis dan simulasi komputer, yang menunjukkan bahwa proses tekuk pada lapisan basal dapat menghasilkan pola sidik jari yang menyerupai sidik jari sebenarnya. Namun, penelitian lebih lanjut diperlukan untuk mengidentifikasi sumber tegangan pertumbuhan dan untuk mempelajari distribusi tegangan dan pola yang sesuai [9].

C. Faktor Genetik

Penelitian ini mengajukan pendekatan untuk menyelidiki hubungan genetik manusia menggunakan sidik jari. Meskipun pola umum sidik jari tampaknya diturunkan melalui keluarga, setiap sidik jari unik karena perbedaan dalam berbagai fitur tingkat. Ini berlaku untuk kloning dan kembar identik juga. Sebagian besar penelitian sidik jari hingga saat ini berfokus pada kemungkinan penerapan teknologi dalam biometrik, peradilan pidana, dan keamanan. Belakangan ini, banyak pembahasan tentang metode untuk menegakkan hubungan genetik berdasarkan sidik jari. Terdapat 75 individu dari 25 keluarga (dari 3 generasi) diambil sidik jarinya. Ciri sidik jari dari ibu jari, telunjuk, dan jari manis tangan kanan diperhitungkan. Studi tentang vektor fitur sidik jari dari anggota keluarga intra-kelas menunjukkan korelasi yang lebih besar daripada anggota keluarga inter-kelas [10].

D. Pola Sidik Jari

Pola sidik jari, seperti sebuah jejak yang tertinggal saat jari yang diberi tinta ditekan ke kertas yang sebenarnya adalah pola tonjolan gesekan pada jari tersebut. Pola tonjolan gesekan ini dikelompokkan menjadi tiga jenis utama *loop, whorl, dan arch* dimana masing-masing dengan variasi unik, tergantung pada bentuk dan hubungan antar tonjolan.

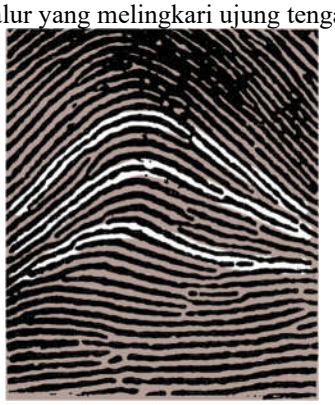


Gbr. 1. Pola *loop* [11]

Gambar 1 merupakan contoh dari sidik jari yang berbentuk pola *loop* dimana sidik jari ini terbentuk ketika alur masuk jari dari satu sisi, melengkung, dan kemudian meninggalkan jari di sisi yang sama.

Gbr. 2. Pola *whorl* [11]

Gambar 2 merupakan contoh dari sidik jari yang berbentuk pola *whorl* dimana sidik jari ini mengacu pada pembentukan alur yang melingkari ujung tengah jari.



Gbr. 3. Pola *arch* [11]

Gambar 3 merupakan contoh dari sidik jari yang berbentuk pola *arch* dimana sidik jari ini terbentuk ketika alur masuk jari dari satu sisi, naik di tengah untuk membentuk *arch*, dan kemudian meninggalkan jari di sisi yang berlawanan.

Para peneliti telah menemukan bahwa pola sidik jari generik sering kali dimiliki oleh anggota keluarga, yang mendukung teori bahwa pola-pola ini diwariskan [12].

E. Fitur Sidik Jari

Sidik jari memiliki fitur yang disebut *minutiae*, yang digunakan untuk mengidentifikasi seseorang. Terdapat beberapa *minutiae* dalam sidik jari antara lain *core* sebagai lingkaran dalam pola *ridge*, *delta* sebagai pertemuan *ridge* berbentuk Y, *ridge* sebagai garis yang menciptakan pola, *ridge ending* sebagai ujung *ridge* yang terputus, *bifurcation* sebagai *ridge* tunggal yang terbagi menjadi dua, dan *spur* sebagai *bifurcation* dengan *ridge* pendek yang bercabang dari *ridge* yang lebih panjang. *Minutiae* unik untuk setiap individu dan tetap tidak berubah sepanjang hidup mereka. Selain itu juga sangat sulit untuk dipalsukan. Ini membuatnya ideal untuk digunakan dalam sistem identifikasi sidik jari [13].

F. Ekstraksi Fitur

Ekstraksi fitur dalam analisis gambar dan pengenalan pola yang melibatkan proses identifikasi dan isolasi informasi yang relevan dari data mentah, seperti gambar, untuk memfasilitasi analisis atau klasifikasi tersebut. Aplikasinya dapat mencakup pengidentifikasian tepi, kontur, tekstur, bentuk, atau fitur khas lainnya dalam data. Berbagai algoritma dan teknik digunakan untuk ekstraksi fitur, termasuk *Principal Component Analysis*. Tujuannya adalah untuk mengurangi dimensi data sambil mempertahankan informasi penting yang dapat digunakan untuk pengenalan pola selanjutnya atau pemrosesan gambar [14].

G. Korelasi dengan Gender

Menurut penelitian [15], menunjukkan adanya korelasi antara karakteristik sidik jari dan gender. Penelitian telah menunjukkan bahwa wanita cenderung memiliki detil *ridge* yang lebih halus dibandingkan pria, dengan perbedaan ketebalan dan lebar *ridge* menjadi lebih signifikan pada subjek Kaukasia dan Afrika Amerika. Penerapan teorema *Bayes* lebih lanjut mendukung korelasi ini, menunjukkan bahwa sidik jari dengan kepadatan *ridge* 11 *ridge*/25 mm^2 atau kurang lebih mungkin berasal dari pria, sementara yang dengan kepadatan 12 *ridge*/25 mm^2 atau lebih mungkin berasal dari wanita, tanpa memandang ras.

H. Principal Component Analysis

Principal Component Analysis (PCA) adalah metode statistik yang digunakan untuk mengidentifikasi pola dalam data dan mengurangi dimensi data sambil tetap mempertahankan informasi penting. Metode PCA adalah dengan mengubah variabel asli menjadi serangkaian variabel tak berkorelasi baru yang disebut komponen utama. Komponen-komponen ini diurutkan sedemikian rupa sehingga beberapa komponen pertama mempertahankan jumlah variasi maksimum yang ada dalam data asli. PCA banyak digunakan untuk eksplorasi data, visualisasi, dan pengurangan *noise*, dan memiliki aplikasi di berbagai bidang seperti pengolahan citra dan sinyal, keuangan, dan genetika [16].

I. Support Vector Machine

Dalam bidang *machine learning*, *Support Vector Machine* adalah model *supervised learning* yang dapat menganalisis data dan mengidentifikasi pola. Teori ini digunakan untuk klasifikasi dan analisis regresi yang terkait dengan algoritma pembelajaran. SVM adalah metode *machine learning* baru berdasarkan teori pembelajaran statistik, dan telah menjadi topik penelitian hangat di bidang *machine learning* karena kinerja pembelajarannya yang luar biasa. SVM juga merupakan mesin pembelajaran berbasis fungsi *kernel*, dan kemampuan generalisasinya sangat bergantung pada fungsi *kernel* yang dipilih [17].

SVM sangat baik untuk menangani dataset berukuran besar ketika digunakan. Teknik SVM lain menggunakan teknik *kernel* untuk memindahkan dataset dari dimensi awal ke dimensi yang lebih besar. Dengan menggunakan metode SVM, generalisasi metode klasifikasi menjadi lebih baik karena metode ini memaksimalkan nilai batas *hyperplane* sambil memaksimalkan nilai *margin*. Pencarian lokasi *hyperplane* adalah inti dari proses pembelajaran atau pelatihan SVM. Dengan data yang paling dekat dengan *hyperplane* (*support vector*) pada masing-masing kelas, garis *hyperplane* dapat ditemukan dengan menghitung jarak (*margin*) terbesar antara garis *hyperplane*. SVM *linear* dan SVM *nonlinear* adalah metode *machine learning* yang digunakan untuk klasifikasi. SVM *linear* dapat digunakan untuk data yang dapat dipisahkan dengan garis *linear*, sedangkan SVM *nonlinear* dapat digunakan untuk data yang tidak dapat dipisahkan dengan garis *linear*.

Decision function untuk *Support Vector Machine* (SVM) digunakan untuk mengklasifikasikan *instance* baru x. Fungsi ini mengukur jarak dari x ke *decision boundary*. Jika *output*-nya positif, x diklasifikasikan sebagai kelas positif (1), sebaliknya diklasifikasikan sebagai kelas negatif (-1). Nilai absolut dari *output* menunjukkan tingkat *confidence* terhadap klasifikasi. Rumus dari *decision function* untuk SVM seperti berikut:

$$f(x) = \sum_{i}^{\square} \square \alpha_i y_i \, (x_i{}^T \, x) + b \qquad (1)$$

SVM *nonlinear* membutuhkan fungsi *kernel* untuk pemetaan fitur lama ke fitur baru. *Kernel trick* digunakan untuk memulai proses perhitungan SVM *nonlinear*. Berikut rumus dari *kernel trick*:

$$K\,(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \qquad (2)$$

Fungsi *kernel* adalah fungsi matematis yang digunakan dalam SVM untuk memetakan data ke ruang fitur yang lebih tinggi. Dalam SVM *nonlinear*, fungsi *kernel* digunakan untuk membuat data yang terlihat tidak terpisahkan secara *linear* dalam ruang asli menjadi terpisahkan secara *linear* dalam ruang fitur yang lebih tinggi. Berikut ini adalah beberapa opsi untuk menentukan fungsi *kernel* yang akan digunakan pada perhitungan SVM *nonlinear*:

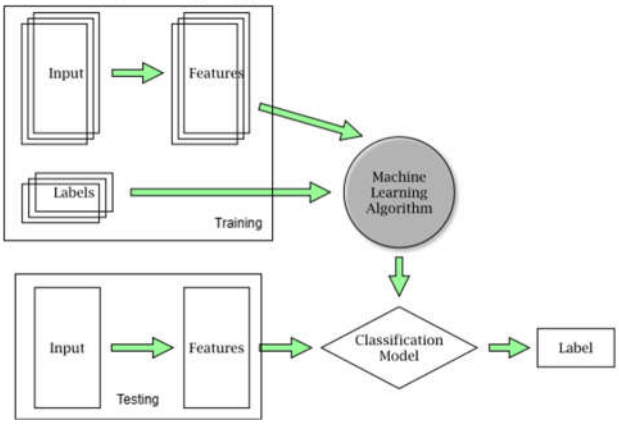
$$K\,(x, y) = exp\,(\frac{-\| \, x - y \, \|^2}{(2 \cdot \, \sigma\text{^}2\,)}) \qquad (3)$$

III. METODOLOGI PENELITIAN

A. Prinsip Klasifikasi

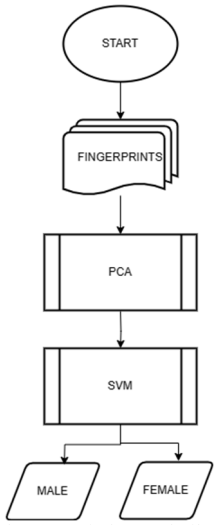
Gambar 4 merupakan diagram umum prinsip kerja *machine learning* untuk proses klasifikasi berdasarkan fitur yang ditentukan dengan metode *Principal Component Analysis* dan label yang ditentukan dengan metode klasifikasi oleh *Support Vector Machine*. Model pembelajaran *supervised learning* adalah model yang dapat digunakan untuk membuat prediksi berdasarkan data pelatihan. Model ini terdiri dari empat komponen utama

yaitu input, fitur, label, dan algoritma. Proses pembelajaran *supervised learning* bekerja dengan cara model dilatih pada data pelatihan, algoritma pembelajaran mesin mempelajari hubungan antara fitur dan label dari data pelatihan, dan model kemudian dapat digunakan untuk mendapatkan label baru dan juga dapat membuat prediksi pada data baru yang dinamakan proses pengujian.



Gbr. 4. Diagram alir prinsip klasifikasi

B. Prinsip Klasifikasi



Gbr. 5. Arsitektur pelatihan

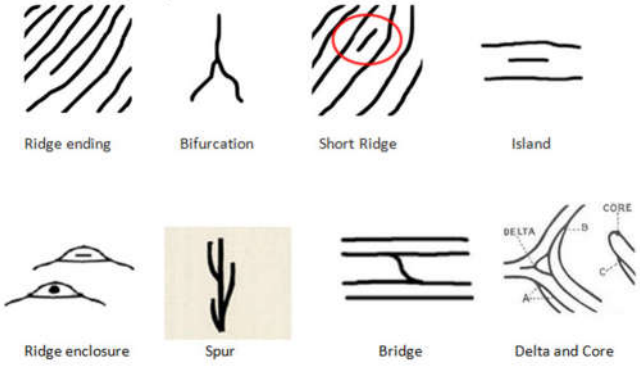
Gambar 5 merupakan diagram pelatihan klasifikasi gender berdasarkan sidik jari dengan metode *Principal Component Analysis* dan *Support Vector Machine*. Sistem ini terdiri dari tiga komponen utama yaitu preprosesing gambar, ekstraksi fitur, dan klasifikasi gender. Proses pengenalan sidik jari akan menggunakan dataset SOCOFing kemudian dilakukan prosesing gambar dengan dibantu oleh PCA akan dilakukan ekstraksi fitur-fitur penting dari gambar sidik jari, seperti *core, delta, ridge, ridge ending, bifurcation, dan spur*. Fitur-fitur tersebut kemudian diklasifikasikan oleh SVM ke dalam dua kelas, yaitu pria dan wanita.

C. Pengumpulan Data

Data yang digunakan untuk penelitian ini adalah *Sokoto Coventry Fingerprint Dataset* (SOCOFing) dimana termasuk ke dalam data sekunder yang diambil melalui situs Kaggle. Dataset ini berisi 6.000 gambar sidik jari dari 600 subjek Afrika, dengan sidik jari setiap subjek diambil sepuluh kali. SOCOFing juga berisi atribut unik berupa label untuk gender, nama tangan dan jari. Semua gambar dikumpulkan dengan pemindai sensor *Hamster Plus* (HSDU03PTM) dan *SecuGen* SDU03PTM.

D. Ekstraksi Fitur

Fitur-fitur yang paling berpengaruh terhadap klasifikasi gender telah ditentukan. Fitur-fitur ini dapat diperoleh dari struktur sidik jari, seperti *core, delta, ridge, ridge ending, bifurcation, dan spur*.



Gbr. 6. Fitur sidik jari

E. Pembuatan Model Klasifikasi

Setelah fitur-fitur dipilih, model SVM dibentuk menggunakan fitur-fitur tersebut. Model SVM dibentuk dengan cara mencari *hyperplane* yang dapat memisahkan dua kelas data kedalam pria dan wanita dengan jelas. Pada model SVM sendiri memiliki beberapa jenis *kernel* yang ada seperti kernel *linear*, *polynomial*, *gaussian*, dan *sigmoid*. Pada penelitian ini akan dilakukan proses klasifikasi menggunakan model SVM yang nantinya akan menghasilkan dua kelas yaitu pria dan wanita. Pemilihan fungsi *kernel* ini sangat penting karena berpengaruh terhadap penentuan fitur baru, yang akan berdampak pada fungsi klasifikasi yang akan dicari, serta *hyperplane* yang akan dicari. Pada penelitian ini akan digunakan SVM *nonlinear* dengan *kernel* yang akan ditentukan dengan perbandingan *kernel* yang terbaik..

F. Perancangan Pengujian Sistem

Setelah model SVM terbentuk, akurasi model SVM diuji menggunakan set pengujian. Set pengujian terdiri dari data sidik jari yang berasal dari dataset SOCOFing pada bagian *altered* (sudah dilakukan augmentasi). Akurasi model SVM diukur menggunakan beberapa metrik yaitu *precision, recall*, dan *confusion matrix*.

IV. IMPLEMENTASI DAN PEMBAHASAN

A. Preprosesing Data

Pada proses preprosesing gambar, ditemukan bahwa terdapat 44 gambar yang memiliki dimensi dan *bit depth* berbeda dari mayoritas dataset. Gambar-gambar tersebut memiliki dimensi dan *bit depth* yang tidak sesuai dengan standar yang telah ditetapkan, yaitu 96x103 dan 32, dimana gambar-gambar tersebut memiliki dimensi 241x298 dan *bit depth* 24. Hal ini menimbulkan permasalahan karena model klasifikasi yang dirancang tidak dapat memproses gambar dengan dimensi dan *bit depth* yang berbeda. Perbedaan tersebut bisa diketahui dikarenakan ada kesalahan bentuk yang tidak standar dari gambar sehingga program terhenti. Oleh karena itu, dilakukan pencarian setiap gambar yang

memiliki dimensi dan *bit depth* yang berbeda dari ukuran mayoritasnya.

Dibuat kode yang bertujuan mencari gambar dalam direktori "Real" yang memiliki dimensi dan *bit depth* berbeda dari standar (96x103, 32). Kode ini menggunakan fungsi *_get_image_dimension* dan *_get_image_bit_depth* untuk mendapatkan dimensi dan *bit depth* gambar. Jika gambar tidak sesuai standar, nama filenya ditambahkan ke variabel *result*. Kode ini kemudian mencetak nama file gambar yang tidak sesuai standar dan jumlahnya.

```
44 gambar
226__M_Left_index_finger.BMP
226__M_Left_little_finger.BMP
226__M_Left_middle_finger.BMP
226__M_Left_ring_finger.BMP
226__M_Left_thumb_finger.BMP
226__M_Right_index_finger.BMP
226__M_Right_little_finger.BMP
226__M_Right_middle_finger.BMP
```

Gbr. 7. Daftar gambar inkonsisten

Gambar 7 merupakan hasil dari kode sebelumnya yang menampilkan daftar gambar yang harus dilakukan preposesing. Keberadaan gambar-gambar tersebut dapat memengaruhi performa model klasifikasi secara keseluruhan. Modelnya akan kesulitan untuk mempelajari pola dan karakteristik sidik jari dengan baik karena terdapat variasi dimensi dan *bit depth* yang tidak terduga. Hal ini dapat menyebabkan penurunan akurasi dan kemampuan model dalam klasifikasi gender. Oleh karena itu, diperlukan preprosesing tambahan pada gambar-gambar yang memiliki dimensi dan *bit depth* berbeda. Dilakukan perubahan pada gambar-gambar dengan mengubah ukurannya menjadi 96x103 dan dikonversi ke format "RGBA" agar sesuai dengan standar dataset.

Dibuat kode yang mendefinisikan fungsi *resize_and_save_images* yang merubah ukuran dan format semua gambar dalam satu folder. Fungsi ini pertama-tama membuka setiap gambar dan kemudian mengubah ukurannya menjadi 96x103 piksel. Gambar yang diubah ukurannya kemudian diubah ke format "RGBA" dan disimpan dengan nama baru yang menyertakan "_resize" dan ".png". Fungsi ini berguna untuk mempersiapkan gambar untuk analisis atau pemrosesan lebih lanjut dengan memastikan semua gambar memiliki ukuran dan format yang standar. Selanjutnya beberapa gambar yang telah diubah ke ukuran yang benar akan digabungkan dengan dataset awal untuk dilakukan proses klasifikasi.

B. Pengolahan Data

Pada proses ini dilakukan pengambilan data berupa gambar sidik jari yang dimuat dari folder “data”. Tujuan dari proses ini adalah untuk dilakukan proses pelabelan dari setiap gambar dengan nilainya sebagai gender berupa “Male” dan “Female”. Selain itu, juga akan dilakukan proses *flatenning* dimana gambar tersebut akan dikonversikan ke dalam bentuk *array* agar nantinya bisa dilakukan ekstraksi fitur dengan metode PCA.

Dibuat kode yang bertujuan untuk mengekstraksi label dari nama file gambar menggunakan *regular expression*. Dengan menggunakan *regex* “(\d+)__[MF]”, maka dicocokkan pola dalam nama file. Bagian “(\d+)” akan mencocokkan satu atau lebih digit berturut-turut, yang kemungkinan adalah nomor gambar. Bagian “__” harus cocok secara persis dengan string harfiah dalam nama file. Dan bagian “[MF]” akan mencocokkan karakter “M” atau “F”, yang mewakili gender (implementasinya pria atau wanita). Kemudian, dengan “group(0)”, didapatkan seluruh *string* yang cocok dengan pola *regular expression* tersebut. Selanjutnya adalah mengambil karakter terakhir dari *string* yang cocok, yang dilakukan dengan “[-1]” dimana akan menentukan “M” atau “F”.

Dikarenakan representasi biner membuat interpretasi hasil model menjadi lebih mudah maka akan dikonversikan ke dalam bentuk biner. Nilai label ditetapkan berdasarkan karakter terakhir yang ditemukan. Jika karakter terakhir adalah “M”, maka label diberi nilai 0, sedangkan jika karakter terakhir adalah “F”, maka label diberi nilai 1.

Proses *flatenning* sangat diperlukan karena proses ekstraksi fitur tidak bisa memproses dalam bentuk gambar. Jadi gambar tersebut kemudian diubah menjadi *array NumPy* menggunakan *np.array()*, yang memungkinkan untuk memanipulasi gambar dalam bentuk *array*. Kemudian, gambar dilakukan *flatten* menjadi satu dimensi menggunakan metode *flatten()*. Ini menghasilkan *array* satu dimensi yang berisi piksel-piksel dari gambar yang sudah diubah.

C. Implementasi Ekstraksi Fitur

Proses ekstraksi fitur yang dilakukan tidak dengan menentukan fiturnya dan nilainya melainkan diberikan pada metode PCA sendiri yang akan menentukan fiturnya. Namun tujuan penggunaan PCA ini adalah untuk mereduksi dimensionalitas dari fitur-fitur tersebut. Sehingga hasil dari proses ini adalah *principal components* yang digunakan untuk proses klasifikasi selanjutnya.

Langkah awal sebelum dilakukan proses ekstraksi fitur dengan PCA perlu dilakukan normalisasi fitur. Tujuannya adalah untuk mengubah skala fitur dalam dataset sehingga memiliki skala yang serupa atau berada dalam rentang yang sama. Normalisasi fitur penting karena untuk menghindari bias.

StandardScaler adalah salah satu teknik dalam preprosesing data yang digunakan untuk mengubah distribusi nilai fitur sehingga memiliki *mean* 0 dan varian standar 1. Selanjutnya menerapkan normalisasi pada *images* menggunakan metode *fit_transform* dari *StandardScaler*. Metode *fit_transform* melakukan dua hal yaitu. Pertama, dihitung rata-rata dan deviasi standar dari setiap fitur dalam *images* (melalui metode *fit*), lalu digunakan informasi tersebut untuk menormalisasi *images* (melalui metode *transform*), menghasilkan *normalized_images*.

Dilakukan implementasi dari metode PCA. Dengan membuat objek PCA dengan menyertakan parameter n_components=4200. Parameter ini menentukan jumlah komponen utama (*principal components*) yang akan dihasilkan oleh PCA. Dalam kasus ini, dipilih untuk 4200 komponen utama.

Gbr. 8. Hubungan antara akurasi dengan jumlah komponen

Pada gambar 8 merupakan alasan penggunaan dari jumlah n_component senilai 4200. Dimana pada gambar tersebut menunjukan akurasi tertinggi terdapat pada jumlah komponen 4200 dengan akurasi 83.17%.

Selanjutnya menerapkan PCA ke dataset *normalized_images*. Dengan metode *fit_transform()* dari objek PCA yang telah dibuat dipanggil. Metode ini melakukan dua hal. Pertama, mempelajari pola dalam data *normalized_images* dan menghitung komponen-komponen utama dari data tersebut. Kedua, menerapkan transformasi PCA ke data, menghasilkan dataset yang baru dengan dimensi yang telah direduksi berdasarkan komponen-komponen utama yang dipilih.

Terakhir, dengan menjumlahkan semua nilai dalam *array explained_variance_ratio_*, yang memberikan jumlah kumulatif dari rasio varian yang dijelaskan oleh semua komponen utama. Ini adalah metrik yang berguna untuk memahami seberapa banyak variasi dalam data yang berhasil dipertahankan setelah reduksi dimensi menggunakan PCA. Semakin tinggi jumlah kumulatif ini, semakin banyak informasi yang dipertahankan dalam dataset yang direduksi.

```
Variance ratio: 1.00
```

Gambar 9. Rasio varian

Gambar 9 menunjukan rasio varian sebesar 1 (atau 100%) berarti komponen PCA menangkap semua varian dalam kumpulan data asli. Hal ini terjadi karena jumlah komponen PCA hampir sama dengan jumlah data yang ada, dengan demikian sedikit informasi yang dihilangkan.

D. Implementasi Klasifikasi

Proses klasifikasi ini akan menghasilkan model yang dapat memprediksi suatu gambar sidik jari berasal dari pria atau wanita. Proses ini terdiri dari dua tahap dimana tahap pertama adalah dengan menyiapkan data yang dibutuhkan dengan pembagian dataset menjadi set pelatihan dan set pengujian. Tahap selanjutnya dilakukan klasifikasi dengan metode *Support Vector Machine* secara *nonlinear* menggunakan kernel *polynomial*.

Langkah awal dari klasifikasi adalah dengan dilakukan pemisahan data antara set pelatihan dan set pengujian menggunakan fungsi *train_test_split*. Penjelasan untuk setiap parameternya antara lain, *principalComponents* adalah matriks fitur yang sudah direduksi dimensinya menggunakan PCA pada langkah sebelumnya. *labels* adalah *array* yang berisi label yang sesuai dengan setiap sampel dalam dataset. Label ini digunakan untuk melatih model dan juga untuk mengukur kinerja model dengan nilai "Male" sebagai 0 dan "Female" sebagai 1. *test_size=0.2* adalah proporsi dataset yang akan dialokasikan untuk subset pengujian. Dalam kasus ini, 20% dari data akan digunakan untuk pengujian, sedangkan 80% akan digunakan untuk pelatihan. Terakhir, *random_state=42* adalah untuk menetapkan nilai yang memastikan bahwa pembagian dataset menjadi subset pelatihan dan pengujian akan konsisten setiap kali kode dijalankan. Hasil dari kode tersebut, dataset akan terbagi menjadi empat subset:

- X_train: Subset pelatihan dari fitur dataset.
- X_test: Subset pengujian dari fitur dataset.
- y_train: Subset pelatihan dari label dataset.
- y_test: Subset pengujian dari label dataset.

Langkah selanjutnya adalah proses klasifikasi dengan algoritma Support Vector Machine. Baris pertama membuat objek svm_classifier yang merupakan model SVM dengan kernel polynomial. Pemilihan SVM nonlinear dengan kernel polynomial adalah keputusan yang didasarkan pada beberapa pertimbangan, terutama terkait dengan efisiensi komputasi dan performa model dibandingkan penggunaan SVM linear yang memerlukan waktu komputasi yang sangat lama. Parameter pemilihan kernel (kernel="poly") menentukan bahwa digunakan kernel polynomial untuk menentukan decision boundary antara kelas "Pria" dan "Wanita". Baris kedua melatih model SVM yang telah dibuat dengan menggunakan subset pelatihan dari dataset, X_train dan y_train. Metode fit() dari objek model SVM digunakan untuk mempelajari pola dari data pelatihan dan menyesuaikan modelnya agar sesuai dengan data. Setelah pelatihan selesai, model SVM siap digunakan untuk membuat prediksi terhadap data baru.

TABEL. 1.

PERBANDINGAN KERNEL SVM NONLINEAR

	Polynomial	Radial Basis Function	Sigmoid
Akurasi	84.17 %	83.17 %	65.83 %

Alasan penggunaan *kernel polynomial* untuk klasifikasi SVM secara *nonlinear* karena penggunaan *kernel polynomial* menunjukan hasil yang terbaik. Dengan dilakukan percobaan dengan tiga jenis *kernel* SVM *nonlinear* yaitu *polynomial*, RBF, dan *sigmoid* yang dibandingkan dengan hasil akurasi akhir. Pada Tabel 1 menunjukan bahwa penggunaan *kernel polynomial* menghasilkan akurasi tertinggi sebesar 84.17% lebih tinggi daripada *kernel* RBF dan *sigmoid*. Oleh karena itu, penelitian ini menggunakan klasifikasi dengan SVM *nonlinear* dengan *kernel polynomial*.

E. Pengujian dan Analisis

Untuk pengujian digunakan model SVM yang telah dilatih dengan metode *svm_classifier* untuk membuat prediksi terhadap subset pengujian *X_test*. Metode predict() digunakan untuk menghasilkan prediksi kelas untuk setiap sampel dalam *X_test*.

Dilanjutkan dengan menghitung akurasi dengan membandingkan prediksi yang dihasilkan (*predictions*)

dengan label sebenarnya dari subset pengujian *y_test*. Fungsi *accuracy_score()* dari *scikit-learn* digunakan untuk menghitung akurasi, yang merupakan rasio antara jumlah prediksi yang benar dan jumlah total prediksi.

Sekaligus menunjukan laporan klasifikasi yang menyediakan informasi rinci tentang kinerja model, termasuk *precision, recall*, dan *f1-score* untuk setiap kelas, serta rata-rata dari masing-masing metrik.

Gambar 10 merupakan output dari evaluasi model SVM pada subset pengujian, termasuk akurasi dan laporan klasifikasi. Akurasi mengukur seberapa sering model melakukan prediksi yang benar. Dalam hal ini, akurasi model menunjukan 84.17%, yang berarti sekitar 84.17% dari sampel dalam subset pengujian diklasifikasikan dengan benar oleh model.

```
Accuracy: 84.17%
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.86      0.84       598
           1       0.86      0.82      0.84       602

    accuracy                           0.84      1200
   macro avg       0.84      0.84      0.84      1200
weighted avg       0.84      0.84      0.84      1200
```

Gbr. 10. Akurasi dan laporan klasifikasi

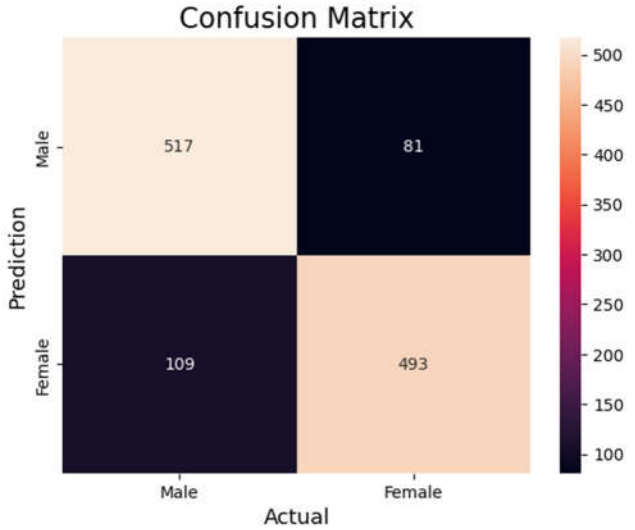
Laporan klasifikasi memberikan informasi tentang kinerja model pada setiap kelas target dalam dataset. Ini mencakup beberapa metrik, termasuk *precision*, *recall*, dan *f1-score*. *Precision* mengukur seberapa banyak dari prediksi positif yang sebenarnya benar. Di sini, *precision* untuk kelas 0 (Pria) adalah 0.83, yang berarti sekitar 83% dari prediksi yang diklasifikasikan sebagai kelas Pria oleh model benar-benar benar. Namun, *precision* untuk kelas 1 (Wanita) adalah 0.86, yang menunjukkan bahwa sampel yang diklasifikasikan sebagai kelas 1 (Wanita) yang sebenarnya benar hampir sama namun sedikit lebih tinggi.

Recall mengukur seberapa banyak dari instance positif yang sebenarnya diklasifikasikan dengan benar oleh model. *Recall* untuk kelas 0 (Pria) adalah 0.86, yang menunjukkan bahwa sebanyak 86% sampel kelas 0 (Pria) yang ada dalam subset pengujian diklasifikasikan dengan benar oleh model. Namun, *recall* untuk kelas 1 (Wanita) adalah 0.82, yang menunjukkan bahwa 82% sampel kelas 1 (Wanita) yang diklasifikasikan dengan benar oleh model dimana lebih rendah.

F1-score adalah rata-rata harmonik dari *precision* dan *recall*, memberikan keseimbangan antara keduanya. *F1-score* untuk kelas 0 (Pria) adalah 0.84, yang merupakan rata-rata harmonik dari *precision* dan *recall* untuk kelas 0 (Pria). Dan juga *f1-score* untuk kelas 1 (Wanita) adalah 0.84.

Dilanjutkan dengan menghitung dengan membandingkan label sebenarnya (*y_test*) dengan prediksi yang dihasilkan oleh model (*predictions*). Fungsi *confusion_matrix()* dari *scikit-learn* digunakan untuk menghasilkan *confusion matrix*. Lalu dengan menggunakan fungsi *heatmap()* dari pustaka *Seaborn* untuk menampilkan visualisasi *heatmap* dari *confusion matrix*.

Gambar 11 menunjukkan *confusion matrix* untuk model klasifikasi. Baris pertama merepresentasikan kelas aktual. Dalam konteks ini, kelas "Pria" merupakan label 0 dan kelas "Wanita" merupakan label 1. Kolom pertama merepresentasikan prediksi yang dibuat oleh model. Dalam konteks ini, nilai 517 di sudut kiri atas menunjukkan bahwa 517 sampel dari kelas "Pria" diklasifikasikan dengan benar sebagai "Pria" oleh model. Nilai 109 di sudut kiri bawah menunjukkan bahwa 109 sampel dari kelas "Wanita" diklasifikasikan secara tidak benar sebagai "Pria" oleh model. Baris kedua merepresentasikan kelas "Wanita". Kolom kedua merepresentasikan prediksi mengenai kelas "Wanita". Dalam konteks ini, nilai 81 di sudut kanan atas menunjukkan bahwa 81 sampel dari kelas "Wanita" diklasifikasikan dengan salah karena diprediksi sebagai "Pria" oleh model. Nilai 493 di sudut kanan bawah menunjukkan bahwa 493 sampel dari kelas "Wanita" diklasifikasikan secara benar sebagai "Wanita" oleh model.



Gbr. 11. *Confusion matrix*

F. Perbandingan Akurasi

Semua metode penelitian telah dilakukan dari semua tahap dari normalisasi, ekstraksi fitur, hingga akhirnya proses klasifikasi. Namun untuk melihat keefektifan penggunaan metode normalisasi dan metode PCA maka dilakukan penelitian untuk membandingkan hasil akurasi jika tidak digunakan kedua proses normalisasi dan PCA dimana dalam hal ini akan digunakan tiga kasus untuk penggunaan kernel *polynomial,* RBF, dan *sigmoid*. Dari hasil penelitian yang disajikan dalam Tabel 2, 3, dan 4, terlihat bahwa penggunaan metode normalisasi dan PCA memiliki pengaruh yang signifikan terhadap akurasi klasifikasi dengan berbagai jenis kernel (*polynomial*, RBF, dan *sigmoid*).

TABEL. 2.

PERBANDINGAN AKURASI KERNEL POLYNOMIAL

	Normalisasi	Tanpa Normalisasi
PCA	84.17 %	88.75 %
Tanpa PCA	84.17 %	84.58 %

Analisis berdasarkan Tabel 2 dengan metode PCA akurasi sedikit lebih rendah dengan normalisasi (84.17%) dibandingkan tanpa normalisasi (88.75%). Hal ini menunjukkan bahwa normalisasi mungkin tidak terlalu membantu ketika PCA sudah diterapkan dalam kasus kernel polynomial. Sedangakn tanpa menggunakan metode PCA akurasi hampir sama baik dengan normalisasi (84.17%)

maupun tanpa normalisasi (84.58%). Ini menunjukkan bahwa normalisasi tidak memberikan perbedaan yang signifikan ketika PCA tidak digunakan.

TABEL. 3.

PERBANDINGAN AKURASI KERNEL RBF

	Normalisasi	Tanpa Normalisasi
PCA	83.17 %	85.75 %
Tanpa PCA	83.17 %	76.08 %

Analisis berdasarkan Tabel 3 dengan metode PCA akurasi lebih rendah dengan normalisasi (83.17%) dibandingkan tanpa normalisasi (85.75%). Seperti pada kernel polynomial, normalisasi tampaknya tidak terlalu efektif ketika PCA digunakan. Sedangkan tanpa menggunakan PCA akurasi jauh lebih tinggi dengan normalisasi (83.17%) dibandingkan tanpa normalisasi (76.08%). Ini menunjukkan bahwa normalisasi sangat penting ketika PCA tidak digunakan untuk kernel RBF, membantu meningkatkan akurasi secara signifikan.

TABEL. 4.

PERBANDINGAN AKURASI KERNEL SIGMOID

	Normalisasi	Tanpa Normalisasi
PCA	65.83 %	62.65 %
Tanpa PCA	65.75 %	51.50 %

Analisis berdasarkan Tabel 4 dengan metode PCA akurasi lebih tinggi dengan normalisasi (65.83%) dibandingkan tanpa normalisasi (62.65%). Ini menunjukkan bahwa normalisasi bermanfaat ketika PCA digunakan dalam kasus kernel sigmoid. Sedangkan tanpa penggunaan metode PCA akurasi jauh lebih tinggi dengan normalisasi (65.75%) dibandingkan tanpa normalisasi (51.50%). Ini menunjukkan bahwa normalisasi sangat penting ketika PCA tidak digunakan untuk kernel sigmoid, membantu meningkatkan akurasi secara signifikan.

Penggunaan PCA tanpa normalisasi kadang-kadang memberikan hasil yang lebih baik atau setidaknya sebanding dengan penggunaan normalisasi. Namun, dalam kasus kernel sigmoid, normalisasi tetap penting meskipun PCA digunakan. Normalisasi sangat penting untuk meningkatkan akurasi di semua jenis kernel. Tanpa normalisasi, akurasi cenderung turun, terutama untuk kernel RBF dan sigmoid. Efektivitas normalisasi dan PCA dapat bervariasi tergantung pada jenis kernel yang digunakan. Untuk kernel polynomial, PCA lebih berpengaruh. Untuk kernel RBF dan sigmoid, normalisasi memainkan peran yang lebih besar, terutama ketika PCA tidak digunakan.

Dengan demikian, penelitian ini menunjukkan pentingnya mempertimbangkan baik normalisasi maupun PCA dalam proses klasifikasi, dan pilihan penggunaannya harus disesuaikan dengan jenis kernel yang digunakan untuk mencapai akurasi terbaik.

V. KESIMPULAN

Penelitian ini berfokus pada klasifikasi gender berdasarkan sidik jari dengan menggunakan metode *Principal Component Analysis* (PCA) dan *Support Vector Machine* (SVM). Penggunaan PCA untuk mereduksi dimensi fitur terbukti tidak efektif dalam mengurangi kompleksitas data tanpa menghilangkan informasi penting karena banyak fitur yang sangat diperlukan untuk proses klasifikasi sehingga jika dilakukan pengurangan dimensi secara signifikan akan mengurangi akurasi model. Dengan dilakukan klasifikasi model SVM dengan *kernel polynomial* menunjukkan kinerja yang efektif dalam mengklasifikasikan gender berdasarkan gambar sidik jari ditunjukan dengan akurasi sebesar 88.75%. Namun jika dilakukan normalisasi dengan *StandardScaler* telah menurunkan akurasi menjadi 84.17%. Dengan analisis *precision*, *recall*, dan f1-s*core* menunjukkan bahwa model ini mampu mengenali pola yang relevan dalam data untuk menentukan gender. Selain itu, evaluasi dengan menggunakan *confusion matrix* memberikan jumlah yang lebih eksplisit dimana model dapat memprediksi dengan benar untuk sidik jari "Pria" sebanyak 517 sedangkan model juga dapat memprediksi dengan benar untuk sidik jari "Wanita" sebanyak 493. Hasil tersebut dari 1200 data yang temasuk dalam data pengujian yang menunjukan 1010 data diprediksi benar dari kelas "Pria" dan "Wanita", lalu sisanya sebanyak 190 data telah salah diprediksi oleh model.

DAFTAR PUSTAKA

[1] S. S. Gornale dan G. C. D, "ANALYSIS OF FINGERPRINT IMAGE FOR GENDER CLASSIFICATION USING SPATIAL AND FREQUENCY DOMAIN ANALYSIS," *American International Journal of Research in Science*, hlm. 13–212, 2013, [Daring]. Tersedia pada: http://www.iasir.net

[2] R. Kaur dan S. G. Mazumdar, "FINGERPRINT BASED GENDER IDENTIFICATION USING FREQUENCY DOMAIN ANALYSIS," 2012. [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:212512523

[3] S. S. Gornale, "Gender Classification Using Fingerprints Based On Support Vector Machines (SVM) With 10-Cross Validation Technique," 2015. [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:212549222

[4] A. Badawi, M. R. Mahfouz, dan R. Jantz, "Fingerprint-Based Gender Classification," 2006. [Daring]. Tersedia pada: https://www.researchgate.net/publication/220809166

[5] S. F. Abdullah, A. F. N. A. Rahman, Z. A. Abas, dan W. H. M. Saad, "Development of a Fingerprint Gender Classification Algorithm Using Fingerprint Global Features," 2016. [Daring]. Tersedia pada: www.ijacsa.thesai.org

[6] P. Gnanasivam dan R. Vijayarajan, "Gender classification from fingerprint ridge count and fingertip size using optimal score assignment," *Complex and Intelligent Systems*, vol. 5, no. 3, hlm. 343–352, Okt 2019, doi: 10.1007/s40747-019-0099-y.

[7] S. Jalali, R. Boostani, dan M. Mohammadi, "Efficient fingerprint features for gender recognition," *Multidimens Syst Signal Process*, vol. 33, no. 1, hlm. 81–97, Mar 2022, doi: 10.1007/s11045-021-00789-6.

[8] H. Agrawal, "Fingerprint Based Gender Classification using multi- class SVM," 2014. [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:212487544

[9] M. Kücken, "Models for fingerprint pattern formation," *Forensic Science International*, vol. 171, no. 2–3.

Elsevier Ireland Ltd, hlm. 85–96, 13 September 2007. doi: 10.1016/j.forsciint.2007.02.025.

[10] P. Naseeda dan M. AnzarS., "Combining features and ancillary measures for exploring hereditary trends in fingerprints," *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, hlm. 1626–1631, 2019, [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:209335708

[11] J. P. Black dan S. Campbell, "Principles of Fingerprint Analysis," Scientific Working Group on Friction Ridge Analysis, Study and Technology.

[12] G. Langenburg, "Are one's fingerprints similar to those of his or her parents in any discernable way?," *Scientific American*, 24 Januari 2005.

[13] D. Maltoni, D. Maio, A. K. Jain, dan S. Prabhakar, *Handbook of Fingerprint Recognition*, 2 ed. Springer Science & Business Media, 2009.

[14] S. Dara dan P. Tumma, "Feature Extraction By Using Deep Learning: A Survey," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, hlm. 1795–1801, 2018, [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:52913157

[15] M. A. Acree, "Is there a gender difference in fingerprint ridge density?," 1999.

[16] J. Le-Rademacher dan L. Billard, "Principal component analysis for histogram-valued data," *Adv Data Anal Classif*, vol. 11, no. 2, hlm. 327–351, Jun 2017, doi: 10.1007/s11634-016-0255-9.

[17] H. Wang, H. Liu, dan X. Zhang, "Development Trend of Support Vector Machine and Applications on the Field of Computer Science," 2016.