SAINTEKS
Volume 19 No 1, April 2022
p-ISSN: 0852-1468; e-ISSN: 2686-0546 (53 – 68)

Klasifikasi Jenis Buku Berdasarkan *Cover* dan Judul Buku Menggunakan Metode *Support Vector Machine* dan *Cosine Similarity*

Classification of Book Types Based on Book Cover and Title Using Support Vector Machine and Cosine Similarity Methods

Sri Winiarti[1*], Desy Widayanti [2], Ulaya Ahdiani[3], Taufiq Ismail[4]

[1,2,3,4]*Jurusan Teknik Informatika Fakultas Teknologi Industri Universitas Ahmad Dahlan Yogyakarta*

*corr_author: Sri.winiarti@tif.uad.ac.id

ABSTRAK

Penelitian ini dibuat untuk mengklasifikasi jenis buku berdasarkan citra cover pada buku. Kategori buku yang digunakan dalam penelitian ini adalah Agama, Kesehatan, Pendidikan, Sastra dan Teknik. Permasalahan yang sering terjadi yaitu semakin banyaknya jenis buku maka akan semakin sulit dalam melakukan pengkategorian, maka memerlukan sebuah sistem untuk mengklasifikasi buku berdasarkan kategori secara otomatis. Penelitian ini melakukan ekstraksi fitur pada buku menggunakan metode *Gray Level Co-occurance Matrix* (GLCM). Metode tersebut mengekstraksi empat fitur yaitu *Homogenity*, *Correlation*, *Energy* dan *Contras*, ke empat fitur tersebut digunakan pada proses ekstraksi nilai pada citra. Setelah itu dilakukan proses klasifikasi jenis buku berdasarkan sampul dan judul buku dengan menerapkan metode *Support Vector Machine* (SVM) dan *Cosine Similarity*. Penerapan SVM untuk mengklasifikasikan jenis buku berdasarkan citra cover buku dan *Cosine Similarity* untuk mengukur tingkat kemiripan dari suatu dokumen dengan menggunakan judul buku. Pada proses pengujian dilakukan menggunakan metode *Confusion Matrix* untuk mendapatkan nilai akurasi. Objek penelitian ini adalah citra sampul buku di perpustakaan Universitas Ahmad Dahlan Kampus 4 dan Kampus 3. Tujuan dari penelitian ini menghasilkan suatu software untuk mengidentifikasi jenis buku berdasarkan judul dan gambar pada Sampul buku, sehingga pembaca lebih cepat tahu informasi jenis buku. Berdasarkan penelitian yang dilakukan metode *Cosine Similarity* menghasilkan nilai akurasi sebesar 93.3% dan metode *Support Vector Machine* menghasilkan nilai akurasi sebesar 86.67%. Dari pengujian yang dilakukan disimpulkan bahwa persentase yang didapat dari perbandingan citra sampul buku asli dengan hasil citra sampul buku yang didapat dari system telah diidentifikasi dengan baik menggunakan metode *Cosine Similiairy*. Hasil uji akurasi dengan metode *Support Vector Machine* menghasilkan nilai yang rendah. Sebagai salah satu penyebab jumlah data yang sedikit dengan model arsitektur VGG16 yang digunakan. Dengan demikian perlu diujicobakan dengan mengubah model arsitektur dengan model arsitektur lainnya, misalnya dengan *Convolutional Neural Network* atau *Tranfer Learning.*

Kata-kata kunci: *Cosine Similarity*, GLCM, jenis buku, *Support Vector Machine*

ABSTRACT

Research are Religion, Health, Education, Literature, and Engineering. The problem that often occurs is that the more types of books it will be more difficult to categorize, it requires a system to classify books based on categories automatically. This study performs feature extraction on books using the Gray Level Co-occurrence Matrix (GLCM) method, The method extracts four features namely Homogeneity, Correlation, Energy, and Contras, the four features are used in the value extraction process in the image. After that, the process of classifying the types of books based on the cover and title of the book is carried out by applying the Support Vector Machine (SVM) and Cosine Similarity methods. The application of SVM to classify types of books based on the image of the book cover and Cosine Similarity to measure the level of similarity of a document using the title of the book. In the testing process, the Confusion Matrix method is used to get the accuracy value. The object of this research is the image of the cover of the book in the library of Ahmad Dahlan University Campus 4 and Campus 3. The purpose of this research is to produce software to identify the type of book based on the title and the image on the cover of the book so that the reader can know information about the type of book more quickly. Based on research conducted by the Cosine Similarity method, the accuracy value is 93.3% and the Support Vector Machine method produces an accuracy value of 86.67%. From the tests conducted, it is concluded that the percentage obtained from the comparison of the original book cover image with the book cover image obtained from the system has been identified properly using the Cosine Similarity method. The results of the accuracy-test using the Support Vector Machine method produce a low value. As one of the causes of the small amount of data with the VGG16 architectural model used. Thus, it is necessary to experiment with changing the architectural model with other architectural models, for example with Convolutional Neural Networks or Transfer Learning.

Keywords: Cosine Similarity, GLCM, Book type, Support Vector Machine

PENDAHULUAN

Buku merupakan jendela bagi setiap orang mendapatkan informasi maupun pengetahuan di bidang politik, ekonomi, maupun aspek kehidupan lainnya. Fungsi utama buku adalah sebagai media informasi yang pada awalnya dalam bentuk tulisan tangan, kemudian cetakan, dan belakangan ini dalam bentuk elektronik. Selain memiliki fungsi sebagai media informasi, buku juga memiliki bagian sampul dan judul yang berbeda-beda. Semakin banyak jenis buku yang berbeda tidak menutup kemungkinan memiliki kemiripan pada sampul dan judul buku. Judul buku yang memiliki kemiripan belum tentu ditempatkan pada kategori yang sama, sehingga hal tersebut bisa menjadi kesalahan dalam melakukan pengkelompokan buku. Begitupun dengan perpustakaan yang digunakan untuk menyimpan buku dan terbitan lainnya yang biasanya disimpan menurut tata susunan tertentu yang digunakan pembaca bukan untuk dijual (Basuki, 1991). Setiap perpustakaan memiliki banyak buku yang di kelompokan berdasarkan kategori buku seperti buku agama, teknologi, politik dan lain-lain. Pada perpustakaan sering terjadi kesalahan dalam mengelompokan buku berdasarkan kategori judul tertentu, tidak memungkinkan bagi perpustakaan dapat mengetahui setiap detial buku yang di miliki, sehingga hal tersebut membuat pengelompokan buku hanya dilakukan berdasarkan judul dan sampul. Permasalahan lainnya bisa terjadi pada kemiripan yang terdapat pada judul buku.

Berdasarkan hasil survei yang dilakukan di lapangan dengan adanya kemiripan pada judul buku dapat mempengaruhi proses pengkategorian buku yang menjadi tidak efesien

sehingga membutuhkan waktu yang lama dan sulit dalam melakukan pengkategorian jenis buku. Dengan adanya sebuah sistem yang dapat mengelompokan buku berdasarkan judul dan sampul buku dapat mempermudah petugas perpustakaan dalam melakukan pengkategorian jenis buku dengan tingakat kemiripan yang sesuai. Maka dari itu, perpustakaan membutuhkan sistem yang dapat mengelompokan jenis buku berdasarkan judul dan sampul buku menggunakan suatu metode yang diterapkan yaitu *Support Vector Machine* dan *Cosine similarity*. *Cosine similarity* adalah sebuah metode yang menghitung tingkat kemiripan (Similarity) antar dua objek atau lebih (A. A. Puspitasari, E. Santoso, & Indriati, 2018). Penelitian ini juga menggunakan *Support Vector Machine* (SVM). SVM dikenal sebagai metode yang memiliki nilai akurasi yang sangat baik untuk pengklasifikasian data.

Penelitian terkait klasifikasi objek ramai dilakukan oleh para peneliti untuk bidang kajian *Artificial Intelligence* (AI), baik penelitian nasional maunpun internasional. Penerapan konsep klasifikasi untuk identifikasi citra maupun identifikasi teks. Penelitian internasional untuk klasifikasi citra untuk identifikasi jenis objek telah dilakukan untuk identifikasi bangunan tradisional (Winiarti, Pramono, & Pranolo, 2022) (Winiarti, Andi Saputro, & Sunardi, Deep Learning dalam Mengindetifikasi Jenis Bangunan Heritage, 2021) dan klasifikasi untuk identifikasi jenis teks juga sudah dilakukan oleh para peneliti (Zhao , Ying, Ju Shuaichen, & LuYonghe, 2021) (Sifeng Jing, et al., 2022). Kedua penelitian tersebut menggunakan konsep deep learning dalam mengidentifikasi objek.

Para peneliti nasional telah melakukan klasifikasi objek untuk identifikasi jurnal berdasarkan abstrak menggunakan metode *cosine similarity* dan *support vector machine*, dengan menggunakan 210 data jurnal untuk membuat model klasifikasi yang dapat mengklasifikasikan jurnal secara otomatis. Berdasarkan pengujian metode ini didapatkan hasil akurasi pengujian *cosine similarity* sebesar 61% dan metode *support vector machine* menghasilkan akurasi sebesar 75% (M. Habibi & P. W. Cahyo, 2020). Penelitian nasional yang mengidentifikasi objek berbentuk teks mengenai klasifikasi jurnal berdasarkan judul dan abstrak menggunakan metode *cosine similarity* dalam penelitian menyatakan bahwa penelitiannya bertujuan untuk menghapus atau eliminasi kata yang tidak signifikan yang akan memberikan hasil yang kurang relevan. Penelitian ini menghasilkan nilai akurasi sebesar 64.28% dengan waktu eksekusi yang di butuhkan saat proses klasifikasi yaitu 0.05033s. (P. dwi Nurfadila, A. P. Wibawa, & I. A. E. Zaeni, 2019).

Berikut disampaikan beberapa teori terkait dengan penelitian ini yaitu *supervised learning*, *Gray Level Co-Ocurrence Matrix* (GLCM), *Support Vector Machine*, dan *Cosine Similarity*.

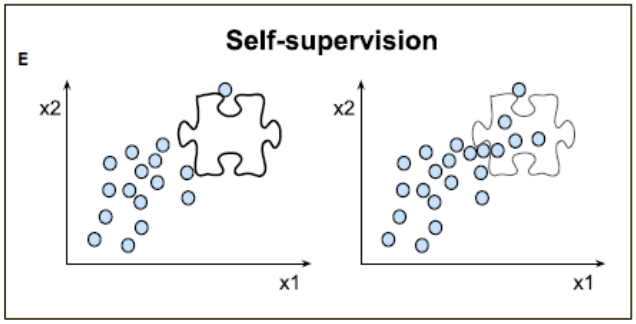
1. Supervised Learning

Supervised learning merupakan salah satu metode yang ada pada *machine learning*. *Supervised learning* menggunakan data training yang diberi label untuk melakukan sebuah pembelajaran mesin, sehingga sistem dapat mengidentifikasikan label input untuk melakukan klasifikasi. *Supervised learning* bisa diartikan juga sebagai pembelajaran terarah atau terawasi (Suyanto, 2018). Pembelajaran dengan pengawasan sendiri mempelajari representasi yang relevan (atau penyematan) tanpa anotasi manual apa pun. Itu bergantung pada proses dua langkah:

a. tugas dalih (atau tugas yang diawasi sendiri) digunakan untuk mempelajari representasi yang bermakna dari anotasi yang melekat pada data,
b. Representasi yang dipelajari dimanfaatkan untuk mengatasi hilir tugas-tugas yang ada, dengan lebih sedikit data berlabel.

Gambar 1 menunjukkan proses kedua langkah untuk proses *Supervised Learning*. Proses pendekatan belajar mandiri yang dimaksud di sini, tugas melukis jigsaw, yang tidak

terkait dengan label (yang disebut tugas dalih) digambarkan sebagai contoh (Gambar 1). Teka-teki tugas dalih dirumuskan secara otomatis dan memungkinkan pembelajaran representasi dari data.



Gambar 1. Pola Kerja *Supervised Learning* (Suyanto, 2018)

2. *Gray Level Co-Ocurrence Matrix (GLCM)*

Gray Level Co-Occurrence Matrix (GLCM) merupakan metode ekstraksi ciri yang digunakan untuk analisis tekstur, dibentuk dari suatu citra dengan melihat pada piksel-piksel yang berpasangan yang memiliki intensitas tertentu. *Co-Occurrence* dapat diartikan sebagai kejadian bersama, berarti banyaknya kejadian pada satu level piksel yang bertetanggan dengan nilai piksel yang lainnya berdasarkan jarak (d) dan orientasi suatu sudut (θ), jarak direpresentasikan sebagai piksel dan orientasi direpresentasikan dalam derajat, orientasi sudut terbentuk dari empat arah yaitu 0º, 45º, 90º dan 135º dengan d=1 sebagai jarak antar piksel (R. A. Surya, A. Fadlil, & A. Yudhana, 2017) (Bakti, et al., 2021). Dalam penggunaan GLCM terdapat beberapa konsep yang perlu dipertimbangkan, yaitu; kontras, korelasi, *energy* dan *Homogeneity*. Kontras merupakan ukuran keberadaan variasi aras keabuan piksel citra, yang dihitung dengan menggunakan persamaan (1), sedangkan untuk menghitung korelasinya menggunakan persamaan (2) (Kadir & Susanto, 2013).

$$Contrast = \sum_{ij=1}^{n} P(i,j)(i-j)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Dimana:
i,j = koordinat di dalam *matrix co-occurrence*.
P(i,j) = *matrix co-occurrence* nilai pada koordinat i,j.

$$Correlation = \sum_{i,j=1}^{n} \frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i \sigma_j} \dots\dots\dots\dots\dots\dots\dots (2)$$

Dimana:
μ_i = nilai rata-rata elemen kolom matrik.
μ_j = nilai rata-rata elemen baris matrik.
σ_i = standar deviasi elemen kolom matrik.
σ_j = standar deviasi elemen baris matrik.

Untuk Korelasi didefinisikan merupakan ukuran ketergantungan linear antar nilai aras keabuan dalam citra, yang dapat dicari dengan menggunakan persamaan 2, sedangkan untuk energi Energi merupakan distribusi intensitas piksel terhadap jangkauan aras keabuan (Widyaningsih, 2017). Persamaan (3) digunakan untuk menghitung Energi. Homogenitas merupakan ketidaksamaan suatu citra dan kontras menghasilkan angka yang lebih besar. Jika bobot berkurang jauh dari diagonal, ukuran tekstur yang di hitung akan lebih besar. Homogenitas dihitung dengan menggunakan persamaan (4) (Hall-Beyer, 2017).

$$Energy = \sum_{t,j=0}^{n} P(i,j)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Dimana: i,j = koordinat di dalam *matrix co-occurrence*.
P(i,j) = *matrix co-occurrence* nilai pada koordinat i,j.
Dimensi N pada matrix *co-occurrence*.

$$Homogenity = \sum_{i,j=0}^{n} \frac{P_{i,j}}{1+(i,j)^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4)$$

Dimana:
i,j = koordinat di dalam *matrix co-occurrence*.
$P_{i,j}$ = *matrix co-occurrence* nilai pada koordinat i,j.
Dimensi N pada *matrix co-occurrence*.

3. *Support Vector Machine*

Pada tahun 1992, pertamakalinya *Support Vector Machin*e (SVM) diperkenalkan oleh Vapnik di *Annual Workshop on Computational Learning Theory*. Konsep SVM berawal dari masalah klasifikasi dua kelas sehingga membutuhkan training set positif dan negatif, *support vector machine* berusaha menemukan *hyperplane* terbaik untuk memisahkan dan memaksimalkan ke dalam dua kelas. Pada SVM data dapat diklasifikasi menggunakan *Lineary Separable Data* dan *Nonlineary Separable Data*. *Lineary separable* data merupakan data yang dapat dipisahkan secara *linear* sedangkan *Nonlineary separable* data merupakan data yang dapat dipisahkan secara *non-linear*. (Susilowati, E., Sabariah, M. K., & Gozali, A. A., 2015). Fungsi persamaan SVM mrnggunakan persamaan (5).

$$f(x) = w * x + b \; atau \; f(x) = \sum_{i=1}^{m} a_i y_i K(x, x_i) + b \quad \dots\dots\dots\dots\dots\dots\dots\dots (5)$$

Keterangan:

w = parameter *hyperplane* yang dicari (garis tegak lurus antara garis hyperplane dan titik support vector)
x = titik data masukan *support vector machine*
a_i = nilai bobot setiap titik data
(x,x_i) = fungsi kernel
b = parameter hyperplane yang dicari (nilai bias)

4. *Cosine Similarity*

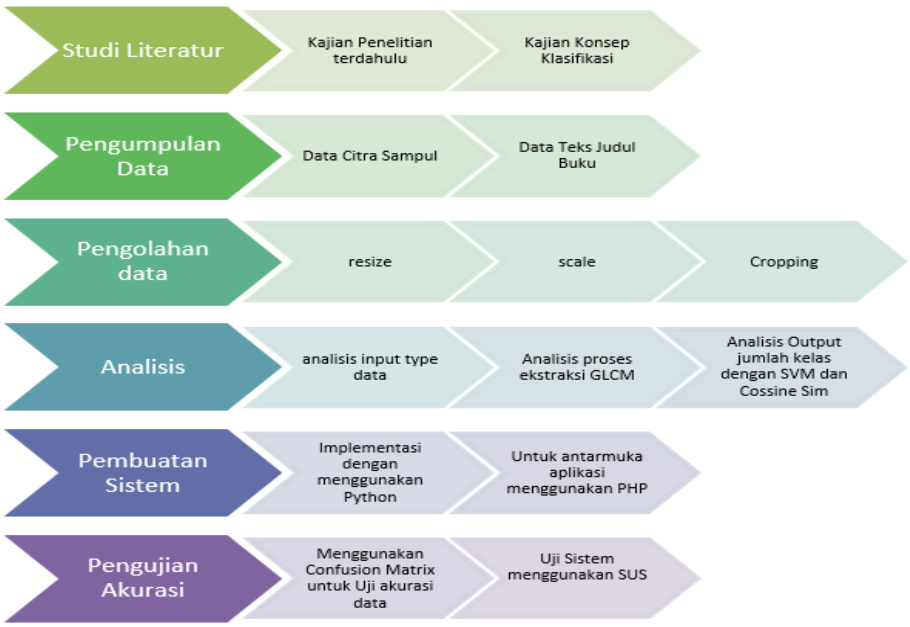
Metode *Cosine similarity* adalah sebuah metode yang menghitung tingkat kemiripan (*similarity*) antar dua objek atau lebih, perhitungan *Cosine similarity* menggunakan objek dokumen untuk menghitung sebuah *similarity* antar dokumen yang dinyatakan dalam sebuah *vector* dengan menggunakan kata kunci sebagaimana dalam persamaan (6) (Sya'bani & R. Umilasari,, 2018).

$$similarity\left(\vec{d_j},\vec{q}\right) = \frac{\vec{d_j}.\vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}\left(W_{ij} \, . W_{iq}\right)}{\sqrt{\sum_{i=1}^{t} W_{ij}^2} \cdot \sum_{i=1}^{t} W_{iq}^2} \dots\dots\dots (6)$$

Keterangan:
D = Dokumen
Q = *query*
T = *term*
W_{ij} = *TF-IDF* kata ke-i dari dokumen ke-j
W_{iq} = *TF-IDF* kata ke-i dari *query*

METODE PENELITIAN

Pada penelitian ini diawali dengan melakukan kajian penelitian terdahulu untuk memperkuat konsep klasifikasi yang dipakai dalam penelitian ini. Karena menggunakan dua algoritma untuk mengidentifikasi jenis buku berdasarkan gambar dan judul buku pada sampul, maka di gunakan algoritma untuk identifikasi citra pada sampul buku dengan menggunakan *algoritma Support Vector Machine* (SVM) dan untuk klasifikasi teks menggunakan *Cosine Simmilarity* dan TF IDF sebagai metode yang digunakan untuk menghitung bobot setiap kata daru judul teks yang dipakai untuk menghitung kemiripan. Selanjutnya dilakukan pengumpulan *dataset* dengan cara mengambil data buku yang terdapat di Toko buku Gramedia Yogyakarta dan buku teks di Perpustakaan Universitas Ahmad Dahlan jumlah data berupa cover dan judul sebanyak 150 data. Untuk Pengolahan data dilakukan pembagian data menjadi 2 bagian yaitu data *training* (latih) dan data *testing* (uji). Dalam proses pengolahan data dilakukan proses *resize*, *scale* dan *cropping* sehingga diperoleh ukuran data 128x128. Data training digunakan untuk melatih algoritma atau model *machine learning* untuk memprediksi hasil yang akan di prediksi, sedangkan data *testing* digunakan untuk mengukur kinerja seperti akurasi dari algoritma yang digunakan, yaitu dengan metode *Confusion matrix*. Penelitian ini menggunakan 70% data *training* dengan 105 jumlah sampul buku, 20% data *validasi* dengan jumlah 30 sampul buku dan 10% *testing* untuk uji akurasi dengan 15 jumlah sampul buku. Proses analisis dilakukan untuk menganalisa hasil data yang dinputkan untuk dilakukan ekstraksi data dengan menggunakan Metode GLCM. Selanjutnya analisa output yang dilakukan untuk memperoleh jumlah kelas yang ditargetkan, untuk sampul buku menggunakan Algoritma SVM dan untuk judul buku menggunakan algoritma *Cosine Similarity.* Algoritma SVM berfungsi untuk memisahkan dua kelas pada data input. Pada kasus ini membagi data ke dalam Lima kelas yaitu Agama, Kesehatan, Pendidikan, Sastra dan Teknik**.** Hal ini dilakukan dengan menemukan garis yang dapat memisahkan kelima kelompok data. Separator *hyperplane* terbaik dapat diperoleh dengan mengukur margin hyperplane dan titik maksimumnya. Margin adalah jarak antara *hyperplane* dan data terdekat dari masing-masing kelas. Data yang paling dekat dengan *hyperplane* terbaik disebut *Support Vector* (Nuraedah, M. Bakri, , & A. A. Kasim, 2018). Gambar 2 menunjukkan tahapan dalam penelitian.



Gambar 2. Tahap Penelitian

HASIL DAN PEMBAHASAN

1. Pengambilan Data

a. Proses Mengambil Citra sampul Buku

Pengambilan citra sampul buku (akuisisi citra buku) dengan sumber pengambilan citra buku didapatkan dari perpustakaan Universitas Ahmad Dahlan kampus 4 dan kampus 3. Menggunakan *Smartphone* Samsung Galaxy J7 Duo dengan resolusi yang digunakan 13 MP Akuisisi data citra sampul buku dilakukan sebanyak 150 pengambilan, dengan pembagian data citra sampul buku untuk data latih yaitu sebanyak 70 %, sedangkan untuk data uji sebanyak 30 %. Pada data uji di bagi lagi menjadi 2 bagian yaitu data uji validasi 20% dan data uji akurasi sebanyak 10% mendapatkan akurasi dari sitem, dimana proses pengambilan data citra buku di bedakan dengan lima kategori yaitu Teknik, Kesehatan, Sastra, Agama dan Pendidikan.

b. Pengelompokan Citra sampul Buku

Pengelompokan citra sampul buku dibagi menjadi lima kategori yaitu Teknik, Kesehatan, Sastra, Agama dan Pendidikan. Dasar pengelompokan ini berdasarkan jenis buku yang paling sering dicari oleh pengunjung perpustakaan serta variasi buku yang paling banyak terdapat di perpustakaan UAD.

c. Pengelompokan Deteksi Kemiripan Cosine Similarity

Dalam jurnal (Herqutanto, 2013) membagi plagiarisme menjadi beberapa jenis salah satunya yaitu Klafisikasi berdasarkan proporsi atau persentasi kata, kalimat, paragraf yang dibajak. Oleh karena itu kemiripan data dari teks citra sampul buku dalam penelitian ini dikategorikan 4 jenis, yaitu; Tidak mirip, Kemiripan ringan, kemiripan sedang dan kemiripan Tinggi.

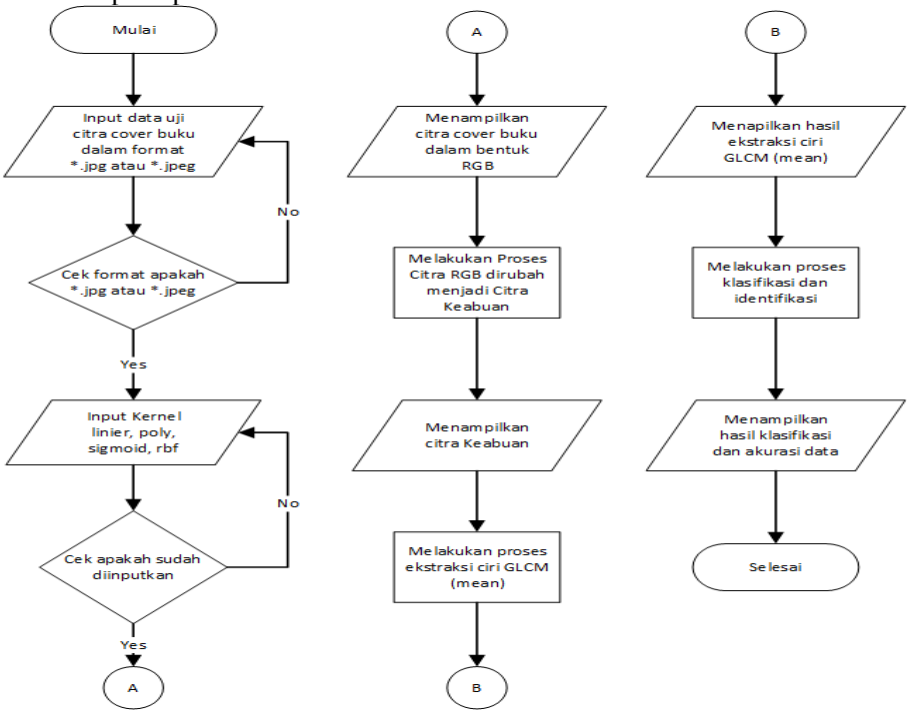
2. *Preprocessing*

Proses pengolahan, setiap citra dirubah menjadi data yang berupa fitur. Ketika pengolahan selesai data fitur dibagi menjadi dua bagian yaitu sebagai data latih dan data uji, untuk selanjutnya dilakukan klasifikasi. Tahapan *preprocessing support vector machine* dilakukan dengan cara menginputkan data latih maupun data uji dengan format *jpg atau *jpeg untuk menentukan setiap label dari buku yang di inputkan. Setiap *cover* buku akan diubah menjadi citra keabuan yang bertujuan untuk mendapatkan nilai ekstraksi fitur citra. Pada tahapan *preprocessing cosine similarity* proses pengolahan, setiap citra diambil teks yang berupa judul dan penerbit untuk dikadikan *document* atau corpus. Ketika proses pengambilan teks judul dan penerbit dari citra sampul buku selesai data *text* dibagi menjadi dua bagian yaitu sebagai data *document* atau corpus dan *query*, untuk selanjutnya dilakukan proses perhitungan bobot tiap kata dengan *term frequency–inverse document frequency* (tf-idf) dan *cosine similarity* untuk mencari nilai kemiripan buku berdasarkan judul.

3. Perancangan Alur Sistem

a. Alur Sistem Klasifikasi

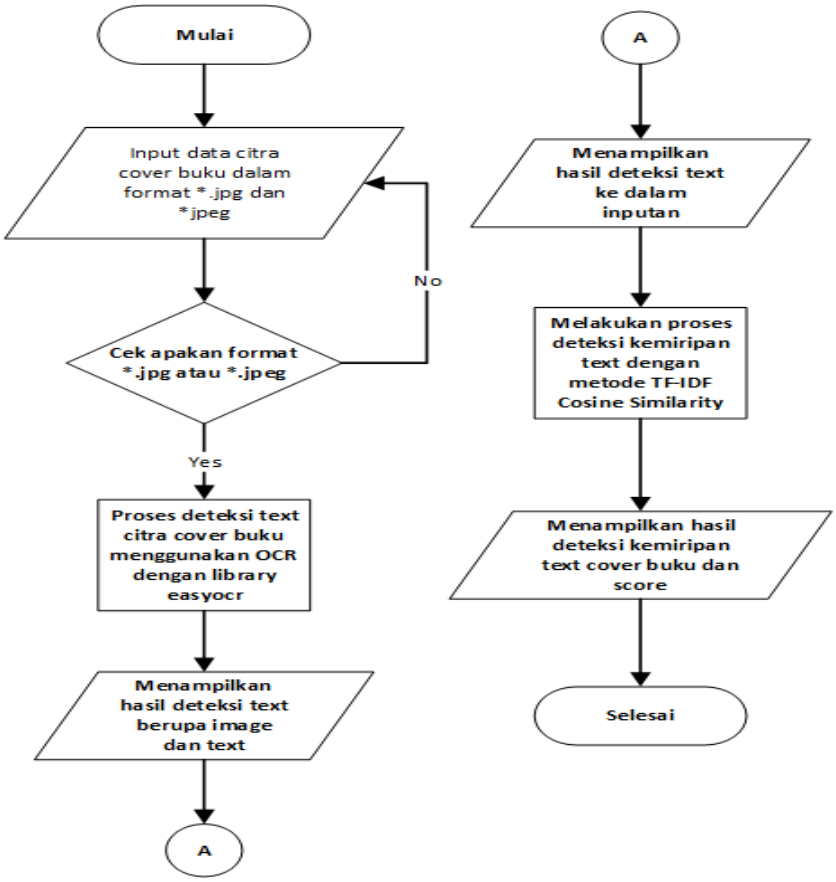
Perancangan pada *flowchart system* klasifikasi dibuat untuk mempermudah dalam menjelaskan proses klasifikasi dari data latih dan data uji yang akan dibangun dalam penelitian ini seperti pada Gambar 3.



Gambar 3. *Flowchart System* Klasifikasi

b. Perancangan Sistem Tf-Idf Cosine Similarity

Pada perancangan *flowchart system* proses tf-idf *cosine similarity text cover* buku dibuat untuk mempermudah dalam menjelaskan proses tf-idf *cosine similarity text cover* buku yang akan dibangun dalam penelitian ini seperti pada Gambar 4.



Gambar 4. *Flowrchart* Sistem proses *cosine similarirty text* citra sampul buku

c. Pembahasan Hasil Sistem

Implementasi sistem untuk mengidentifikasi jenis buku ini menggunakan dua algoritma, yaitu algoritma SVM untuk deteksi Gambar dan Algoritma *Cossine Similairity* untuk judul buku berupa teks.

Pengujian citra sampul buku ke data latih citra sampul buku dengan banyaknya citra uji sampul buku yaitu 15 citra sampul buku uji sedangakan data latih citra sampul buku sebanyak 105 data dari hasil ekstraksi fitur menggunakan GLCM dengan mengambil nilai rata-rata yang sudah diberi label. Data uji yang akan dilakukan klasifikasi dengan data latih menggunakan metode SVM dimana pada proses klasifikasi menggunakan kernel linear yang akan menghasilkan label atau kategori dan akurasi data. Data uji menggunakan sempel sebanyak 15 data citra sampul buku yang akan diujikan kedalam *system* dapat menghasilkan label atau kategori berupa Teknik, Kesehatan, Pendidikan, Sastra, dan Agama dengan menggunakan klasifikasi SVM. Tabel 1 menunjukkan data hasil klasifikasi.

Tabel 1. Kategori klasifikasi buku

Nilai	Kategori hasil klasifikasi
1	Teknik
2	Kesehatan
3	Sastra
4	Agama
5	Pendidikan

Penjelasan metode SVM digambarkan sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah kelima kelas yang ditargetkan. Gambar 3 memperlihatkan bagaimana data citra diolah dari data asli bertipe RGB dan diubah menjadi Citra bertipe keabuan. Selanjutnya dilakukan ektraksi fitur dengan menggunakan metode GLCM dengan menggunakan persamaan (7). Dengan demikian akan didapat nilai entropy, homogeneity, energy dan nilai contras. Dimana masing-masing nilai ini akan dipakai untuk yang digunakan untuk mengukur konsentrasi pasangan intensitas pada matriks GLCM. Masing-masing nilai tersebut dapat menggunakan persamaan (8) – (11) (Musrini B, Andriana, & Hidayat, 2017).

$$I(i, y) = \alpha . R + \beta . G + \gamma . \beta \quad \text{..... (7)}$$

Dimana:

I(x,y) = level keabuan pada suatu koordinat
R = nilai warna merah
G = nilai warna hijau
B = nilai warna biru

Entropy menggunakan persamaan: $E_2 = \sum_i^m \sum_j^n p(i,j) log\{p(i,j)\}$ …….. (8)

Dimana:

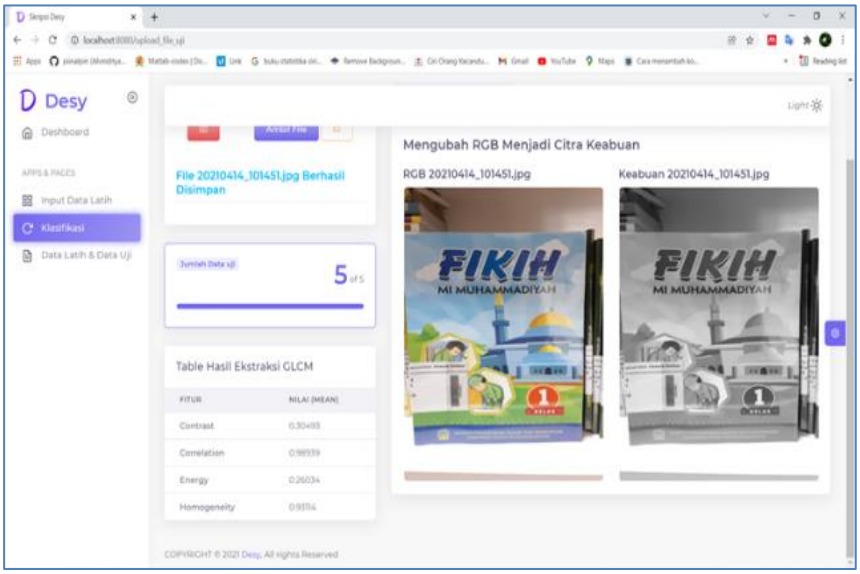
P = matriks GLCM normalisasi
i = indeks baris matriks
P j = indeks kolom matriks P

Energy menggunakan persamaan: $H=\sum_i^m \sum_j^n \frac{p(i,j)}{1+(i,j)^2}$ ……..…….(9)

Masalah klasifikasi dapat dijabarkan dengan usaha menemukan hyperplane yang memisahkan kelima kelompok tersebut. *Hyperplane* pemisah yang terbaik diantara kelima kelas ditemukan dengan cara mengukur margin hyperplane dan mencari titik maksimalnya. Margin adalah jarak antara *hyperplane* dengan pola terdekat dari setiap kelas. Pola yang paling dekat ini disebut sebagai *support vector*.

Klasifikasi dengan Model SVM yang telah diperoleh selanjutnya diujikan pada 80 data uji. Proses pengujian ini dilakukan untuk mengetahui akurasi model klasifikasi dengan metode SVM, seberapa akurat metode SVM mampu mengklasifikasi penduduk yang berhak menerima bantuan atau tidak berhak menerima bantuan. Dalam proses pelatihan, terdapat beberapa variabel yang mempengaruhi formula seperti nilai alpha, bias, dan weight. Nilai ini memiliki keterkaitan yang sangat erat karena untuk menentukan nilai bias dibutuhkan nilai *weight* nilai alpha sedangkan nilai alpha sangat berkaitan dengan nilai yang terkandung dalam data. Dari proses ini tercipta formula yang memiliki tingkat klasifikasi yang cukup baik untuk digunakan dalam pengujian data. Hasil pengujian dari 80 data uji tidak dapat mencapai 100%. Hal ini terjadi karena salah satu parameter yang digunakan memiliki nilai yang tinggi. Proses *input* citra sampul buku setelah diperoleh hasil klasifikasi dengan SVM selanjutnya dilakukan uji performa dengan metode *confusion matrix.*

Proses awal dengan *pre-prosessing*, yaitu sebuah proses pengubahan bentuk data yang terstruktur sembarang menjadi data yang terstruktur sesuai dengan kebutuhan untuk proses penghitungan kemiripan teks. Pada penelitian ini tahap *preprocessing* yang akan digunakan terdiri dari *case folding, tokenizing, filtering*. Pada hasil implementasi system pemrosesan deteksi teks dapat dilakukan ekatraksi teks dari citra sampul buku yang diinputkan untuk mendapatkan teks sebanyak 70 % dokumen. Gambar 5 merupakan tampilan antarmuka aplikasi proses proses ekstraksi gambar dengan menggunakan algoritma GLCM.

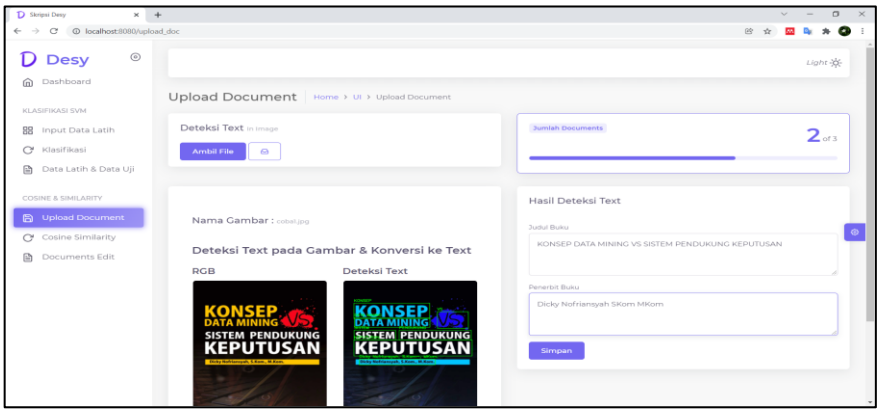


Gambar 5. Proses ektraksi citra dengan GLCM

Untuk pengelompokan buku berdasarkan judul teks buku menggunakan metode *Cosine Similarity* dan TF IDF untuk mencari kemiripan teks pada sampul buku.
Kemiripan data dari text citra *cover* buku dalam penelitian ini dikategorikan sebagai berikut:

a. Tidak Ada Kemiripan : 0%
b. Kemiripan Ringan : 1% - 29%
c. Kemiripan Sedang : 30% - 69%
d. Kemiripan Tinggi : >= 70%

Pada Gambar 6 menunjukkan proses data buku yang diinputkan ke dalam sistem, kemudian dilakukan deteksi *text* pada judul buku untuk selanjutnya disimpan sebagai *document*. Hasil deteksi kemiripan *text* citra sampul buku juga dilakukan ekstraksi *text* pada sampul buku untuk mendapatkan *text* dari citra sampul buku yang diinputkan dan dilakukan dekteksi kemiripan *tex*t dengan *document* yang telah disimpan dalam basis data menggunaan metode tf-idf dan *Cosine Similarity*.



Gambar 6. Proses *input* data buku untuk deteksi kemiripan jenis buku berdasarkan teks

Pada Gambar 6 buku yang telah disimpan sebagai document dilakukan deteksi kemiripan dengan TF IDF Dan *Cosine Similarity* menggunakan *source code* dalam pemrograman Python seperti pada Gambar 7.

```
vectorizer = CountVectorizer(stop_words='english')
pada data corpus
response = vectorizer.fit_transform(data_corpus)
words = vectorizer.get_feature_names()
df = pd.DataFrame(response.todense().T,
index = vectorizer.get_feature_names(),columns=[f'D{i+1}'for i in
range(len(data_corpus))])
total_d_corpus = len(data_corpus)
queryTFIDF = TfidfVectorizer().fit(words)
queryTFIDF = queryTFIDF.transform([detect_txt])
cosine_similarities=cosine_similarity(queryTFIDF,
response).flatten()
```

Gambar 7. *Source code* penghitungan kemiripan

d. Pengujian Akurasi

Pengujian citra sampul buku ke data latih citra sampul buku dengan banyaknya citra uji sampul buku yaitu 15 citra sampul buku uji. Data uji yang akan dilakukan klasifikasi dengan data latih menggunakan metode SVM dimana pada proses klasifikasi menggunakan kernel linear, polynomial, sigmoid dan RBF yang akan menghasilkan label atau kategori dan akurasi data. Data uji menggunakan sempel sebanyak 15 data citra sampul buku yang akan diujikan kedalam *system* dapat menghasilkan label atau kategori berupa Teknik, Kesehatan, Pendidikan, Sastra, dan Agama dengan menggunakan *Confusion matrix*. Sebagai salah matu metode uji akurasi, *Confusion matrix* biasanya digunakan untuk melakukan perhitungan akurasi yang digambarkan dengan tabel yang menyatakan jumlah data uji yang mengandung *True* dan jumlah data uji yang mengandung *false* setelah data diklasifikasikan (Rahman, Darmawidjadja, & Alamsah, 2017). Tabel 2 menunjukkan nilai *Confusion matrix*.

Tabel 2.Tabel *confusion matrix*

Correct Classification	**Classified as**	
	Predicted "+"	**Predicted "-"**
Actual "+"	True Positives	False Negatives
Actual "-"	False Positives	True Negatives

Berdasarkan Tabel 2 dijelaskan sebagai berikut:

- **True Positives** (TP) adalah jumlah record datapositif yang diklasifikasikan sebagai nilai positif .
- **False Positives** (FP) adalah jumlah record data negatif yang diklasifikasikan sebagai nilai positif.
- **False Negatives** (FN) adalah jumlah record data positif yang diklasifikasikan sebagai nilai positif
- **True Negatives** (TN) adalah jumlah record data negatif yang diklasifikasikan sebagai nilai negative.

Nilai yang dihasilkan melalui metode *Confusion Matrix* adalah berupa evaluasi sebagai berikut :

a. *Accuracy*, presentase jumlah record data yang diklasifikasikan (prediksi) secara benar oleh algoritma.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad \text{.....(10)}$$

$$Hasil\ accurasi\ yang\ didapat: \frac{13+0}{13+0+2+0} * 100\%$$

$$Accuracy = \frac{10}{15} * 100\% = 86.67\%$$

Dengan mengacu pada Tabel 2, didapat hasil kebenaran data dari sistem dari 15 data yang diuji sebanyak 13 data bernilai True, dan bernilai False sebanyak 2 data. Berdasarkan Tabel 2 *Confusion matrix*, maka untuk klasifikasi buku ini, dihasilkan nilai performansinya seperti ditunjukkan pada Tabel 3.

Tabel 3. Hasil performansi

Kernel	Linear
True Positive	13
True Negative	0
False Positive	2
False Negative	0

b. *Misclassification* (Error) *Rate*, presentase jumlah record data yang diklasifikasikan (prediksi secara salah oleh algoritma.

(FP + FN) / Total data = *Misclassification Rate* ……..……....(11)

$$Mis\ classification\ Rate = \frac{2+0}{13+0+2+0} = \frac{2}{15} = 0{,}13\ x\ 100\% = 13.3\%$$

Untuk pengujian akurasi dengan metode SVM menggunakan data citra uji Sampul buku sebanyak 15 dataset disajikan datanya pada Tabel 4 dengan menggunakan persamaan 1,2,3 dan 4.

Tabel 4. Hasil Klasifikasi dari Citra sampul Buku

No.	Label cover buku asli	*Contrast*	*Correlation*	*Energy*	*Homogeneity*	Hasil System
						Kernel Linear
1	Agama	0.17462	0.98310	0.35501	0.93713	4
2	Agama	0.16097	0.98682	0.31365	0.94817	3
3	Agama	0.15001	0.99069	0.32740	0.95804	3
4	Kesehatan	0.15947	0.99047	0.38832	0.92869	4
5	Kesehatan	0.17293	0.98183	0.41887	0.93411	4
6	Kesehatan	0.23966	0.99269	0.27490	0.90462	3
7	Pendidikan	0.41318	0.97505	0.30238	0.88294	5
8	Pendidikan	0.32569	0.98382	0.25144	0.88877	5
9	Pendidikan	0.34177	0.98455	0.23912	0.87610	5
10	Sastra	0.09262	0.99408	0.33637	0.95800	3
11	Sastra	0.15508	0.99501	0.26044	0.94719	3
12	Sastra	0.27492	0.98851	0.33607	0.93612	3
13	Teknik	0.30448	0.98454	0.26977	0.90849	1
14	Teknik	0.28948	0.99057	0.26819	0.91916	1
15	Teknik	0.23681	0.99377	0.25103	0.92518	3

Setelah mendapatkan hasil klasifikasi berupa kategori baru yang didapatkan dari hasil *system* selanjutnya data baru hasil identifikasi kategori *system* dilakukan perbandingan dengan citra *cover* buku asli untuk melihat tingkat akurasi dari *system*. Selanjutnya dilakukan pencocokan hasil klasifikasi berdasarkan *system* dengan citra sampul buku asli untuk mendapatkan akurasi *system.* Dengan menggunakan persamaan 10 dan 11 maka didapat hasil akurasi pada Tabel 4, sedangkan hasil identifikasi dapat

mengidentifikasi buku dan sesuai 13 data uji dari total data uji yang digunakan. Detil hasil identifikasi disajikan pada Tabel 5.

Tabel 5. Hasil Identifikasi System dengan kernel linear

No	Label cover buku asli	Hasil system
		Kernel linear
1	Agama	4
2	Agama	4
3	Agama	4
4	Kesehatan	4
5	Kesehatan	4
6	Kesehatan	3
7	Pendidikan	5
8	Pendidikan	5
9	Pendidikan	5
10	Sastra	3
11	Sastra	3
12	Sastra	3
13	Teknik	1
14	Teknik	1
15	Teknik	1

Pengujian akurasi untuk klasifilasi berdasarkan judul teks menggunakan *Cosine Similairty* dengan menggunakan persamaan 10 dan 11. Pengujian menggunakan dataset 15 sampul buku data uji dengan kategori sampul buku teknik, Kesehatan, sastra, agama dan Pendidikan. Sedangkan tingkat kemiripan dikategorikan dengan kemiripan tinggi, sedang, ringan, tidak mirip. Hasil Uji performa dengan Cosine Simmilairity disajikan pada Tabel 7 dan Tabel 6.

Tabel 6. Hasil performa *cosine similarity*

Kernel	Cosine Similarity
True Positive	14
rue Negative	0
False Positive	1
False Negative	0

Pada akurasi *Cosine similarity system* yang telah dilakukan dengan uji coba 15 sampul buku perhitungan dari citra asli sampul buku dengan hasil yang didapat pada sistem dapat dilihat pada Tabel 7 dimana data bernilai *True* 14 data buku sedangkan bernilah *False* 1 data buku. Jadi, uji coba 15 sampul buku didapatkan hasil perhitungan akurasi dengan metode *Cosine similarity* pada sistem yang telah dibuat yaitu dengan tingkat akurasi 93.3%. Dengan menggunakan persamaan 10 dan 11 diperoleh nilai akurasinya sebagai berikut:

$$Hasil\ accurasi\ yang\ didapat: \frac{14+0}{14+0+1+0} * 100\%$$

$$Accuracy = \frac{14}{15} * 100\% = 93.3\%$$

KESIMPULAN

Berdasarkan pengujian yang dilakukan, maka dapat disimpulkan bahwa identifikasi buku dapat dilakukan dengan menggunakan 2 parameter, yaitu gambar dan judul teks pada sampul buku. Untuk identifikasi sampul buku berdasarkan gambar menggunakan

Algoritma *Support Vector Model* (SVM) memperoleh akurasi 86.67% dan identifikasi jenis buku berdasarkan teks dengan Algoritma *Cosine Similarity* memperoleh akurasi 93.3%. Sistem dapat mengklasifikasi kemiripan dokumen *text* dengan mengelompokan kategori berdasarkan ciri tekstur menggunakan metode *gray level co-occurrence matrix* dan SVM dengan nilai akurasi yang berbeda antara metode SVM dan *Cosine similarity*. Hasil pengujian dengan metode *Cosine similarity* memperoleh hasil akurasi lebih tinggi dibandingkan metode SVM. Hal ini disebabkan karena Algoritma SVM lebih optimal untuk klasifikasi dengan 2 kelas, sedangkan dalam penelitian ini menggunakan 5 kelas dengan jumlah data yang sedikit. Untuk Algoritma Cosine Similarity yang dipakai untuk mengidentifikasi kemiripan buku memang pada umumnya memiliki hasil akurasi yang tinggi. Hal ini karena *Cosine similarity* lebih efektif untuk klasifikasi dokumen teks adalah menggunakan pengukuran sudut cosinus dan kesamaan dokumen tidak terlalu dipengaruhi oleh panjang dokumen. Dengan demikian dapat disimpulkan bahwa presentase yang didapat dari perbandingan citra sampul buku asli dengan hasil citra sampul buku yang dapat dari system berhasil melakukan identifikasi dengan baik.

UCAPAN TERIMA KASIH

Terima kasih kepada Lembaga Penelitian dan Pengabdian Universitas Ahmad Dahlan yang telah memfasilitasi terselanggaranya penelitian ini. Terima kasih juga kepada Perpustakaan Universitas Ahmad Dahlan yang telah membantu memberikan data dan informasi terkait ijin akses koleksi buku yang tersedia secara mudah. Dukungan kepada semua tim peneliti yang membantu hingga selesainya penelitian dengan capaian luaran seperti yang dijanjikan.

DAFTAR PUSTAKA

A. A. Puspitasari, E. Santoso, & Indriati. (2018). Klasifikasi Dokumen Tumbuhan Obat Menggunakan Metode Improved kNearest Neighbor. *Jurnal Pengembangan Teknologi Informasi. dan Ilmu Komputer, vol. 2, no. 2* , pp. 486–492.

Bakti, L. D., Imran, B., Wahyudi, E., Arwidiyarti, D., Suryadi, E., Multazam, M., & Maspaeni. (2021). Data extraction of the gray level Co-occurrence matrix (GLCM) Feature on the fingerprints of parents and children in Lombok Island, Indonesia. *Data in Brief-elsavier, 36*.

Basuki, S. (1991). *Pengantar Ilmu Perpustakaan.* Jakarta: PT Gramedia.

Hall-Beyer, M. (2017). GLCM Texture: a tutorial. *17th International Symposium on Ballistics, 2(March), 267–274.*

Kadir , A., & Susanto, A. (2013). *Teori dan Aplikasi Pengolahan Citra (D. Hardjono, Ed.).* . Yogyakarta: CV.ANDI OFFSET.

M. Habibi, & P. W. Cahyo. (2020). Journal Classification Based on Abstract Using Cosine Similarity and Support Vector. *JISKA (Jurnal Inform. Sunan Kalijaga), vol. 4, no. 3*, pp. 48, 2020.

Musrini B, M., Andriana, & Hidayat, A. S. (2017). Implementasi Algoritma GLCM Dan MED pada Aplikasi Pendeteksi Kolesterol Melalui Iris Mata. *MIND Journal*, 23 - 41.

Nuraedah, M. Bakri, , & A. A. Kasim. (2018). Quadratic support vector machine for the bomba traditional textile motif classification. *Indones. J. Electr. Eng. Comput*, 1004–1014.

P. dwi Nurfadila, A. P. Wibawa, & I. A. E. Zaeni. (2019). Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stopword Removal. *International Journal Artificial Intelligence Res. , vol. 3, No2*.

R. A. Surya, A. Fadlil, & A. Yudhana. (2017). Ekstraksi Ciri Metode Gray Level Co-Occurrence Matrix (GLCM) dan Filter. *Jurnal Informasi Pengemb. IT (JPIT), Vol. 02, No. 02*, 23-26.

Rahman, M., Darmawidjadja, M., & Alamsah, D. (2017). Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). *JURNAL INFORMATIKA*, 36-45.

Sifeng Jing, Xiwei Liu, Xiaoyan Gong, Ying Tang, Gang Xionga, Sheng Liu, . . . Rongshan. Bi. (2022). Correlation analysis and text classification of chemical accident cases. *Process Safety and Environmental Protection, 158*, 698-710.

Susilowati, E., Sabariah, M. K., & Gozali, A. A. (2015). Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter. *Proceeding of Engineering, 2(1), 1478–1484.*

Suyanto. (2018). *Machine Learning: Tingkat Dasar dan Lanjut.* Bandung: Informatika Bandung.

Sya'bani , M., & R. Umilasari,. (2018). Penerapan Metode Cosine Similarity dan Pembobotan TF / IDF pada Sistem Klasifikasi Sinopsis Buku di Perpustakaan Kejaksaan Negeri Jember,. *Jastindo (Jurnal Sistem Teknologi Indonesia), vol. 3, no. 1*, 31-42.

Widyaningsih. (2017). Identifikasi Kematangan Buah Apel Dengan Gray Level Co-Occurrence Matrix (GLCM). . *Jurnal SAINTEKOM, , 6*.

Winiarti, S., Andi Saputro, M. Y., & Sunardi. (2021). Deep Learning dalam Mengindetifikasi Jenis Bangunan Heritage. *JURNAL MEDIA INFORMATIKA BUDIDARMA, Volume 5, Nomor 3*, 831-837.

Winiarti, S., Pramono, H., & Pranolo, A. (2022). *JARINA - Journal of Artificial Intelligence in Architecture, 1 No 1*, 20-29.

Zhao , R., Ying, X., Ju Shuaichen, & LuYonghe. (2021). Patent text modeling strategy and its classification based on structural features. *World Patent Information, 67*.