Available at https://jtsiskom.undip.ac.id (20 Mar 2020)
DOI:10.14710/jtsiskom.8.2.2020.150-156

Jurnal Teknologi dan Sistem Komputer, 8(2), 2020, 150-156

Unjuk kerja k-nearest neighbor untuk alihaksara citra aksara Nusantara

K-nearest neighbor performance for Nusantara scripts image transliteration

Anastasia Rita Widiarti

Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma
Kampus III Paingan Maguwoharjo Depok Sleman, Yogyakarta, Indonesia 55281

Cara sitasi: A. R. Widiarti, "Unjuk kerja k-nearest neighbor untuk alihaksara citra aksara Nusantara," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 150-156, 2020. doi: 10.14710/jtsiskom.8.2.2020.150-156, [Online].

Abstract* - *The concept of classification using the k-nearest neighbor (KNN) method is simple, easy to understand, and easy to be implemented in the system. The main challenge in classification with KNN is determining the proximity measure of an object and how to make a compact reference class. This paper studied the implementation of the KNN for the automatic transliteration of Javanese, Sundanese, and Bataknese script images into Roman script. The study used the KNN algorithm with the number k set to 1, 3, 5, 7, and 9. Tests used the image dataset of 2520 data. With the 3-fold and 10-fold cross-validation, the results exposed the accuracy differences if the area of the extracted image, the number of neighbors in the classification, and the number of data training were different.

Keywords *- classification; KNN algorithm; transliteration; Nusantara script image*

Abstrak* - *Konsep klasifikasi menggunakan metode k-nearest neighbor (KNN) sederhana, mudah dimengerti, dan diimplementasikan dalam sistem. Tantangan utama dalam klasifikasi dengan KNN adalah menentukan sejauh mana luas perimeter untuk mengukur kedekatan objek dan bagaimana membuat kelas training yang handal. Kajian ini menjelaskan hasil implementasi KNN untuk transliterasi otomatis aksara Jawa, Sunda, dan Batak ke dalam aksara Roman. Kajian menggunakan algoritme KNN dengan nilai k diatur ke 1, 3, 5, 7, dan 9. Pengujian dilakukan pada kumpulan data citra sebanyak 2520 data. Dengan validasi silang 3-fold dan 10-fold, hasil kajian menunjukkan bahwa ada perbedaan akurasi jika area gambar pada saat diekstraksi, jumlah tetangga dalam klasifikasi, dan jumlah data latih berbeda.

Kata kunci *– klasifikasi; algoritme KNN; transliterasi; citra aksara Nusantara*

I. PENDAHULUAN

Sejak ditemukannya konsep analisis diskriminan untuk membedakan objek yang diterapkan oleh Cover dan Hart [1] untuk klasifikasi yang didasarkan pada konsep ketetanggaan terdekat, pemanfaatan *k-nearest neighbour* (KNN) untuk klasifikasi objek menjadi sangat berkembang. KNN juga telah menjadi pendekatan yang populer untuk menyelesaikan persoalan di ranah pengenalan pola, pembelajaran mesin, pengelompokkan teks, penambangan data, dan pengenalan objek.

Konsep KNN tergolong sebagai konsep klasifikasi yang sederhana dan menjadi populer dalam penelitian terkait klasifikasi [2]-[4]. Wahyono dkk. [5] menerapkan KNN untuk klasifikasi teks pada dokumen dengan mengubah representasi kata menjadi vektor. Premavathi dan Thangaraj [6] telah berhasil mengklasifikasikan citra pembuluh darah vena pada tangan, dan mendapatkan akurasi rata-rata klasifikasi pada *k* bernilai 1, 2, 3, 4, dan 5 sebesar 95,10 %. Ong dan Suhartono [7] mendapatkan akurasi 76,9 % untuk klasifikasi citra aksara Roman tulisan tangan. Ajao dkk. [8] juga berhasil menggunakan KNN untuk klasifikasi aksara Yoruba. KNN juga dapat dipergunakan untuk klasifikasi data video, seperti yang digunakan oleh Al-asady dan Al-amery [9] dengan akurasi di atas 75 %.

Unjuk kerja KNN untuk klasifikasi juga disimpulkan lebih baik jika dibandingkan dengan metode SVM dalam [8], [9] dan dengan Bayes dalam [8]. KNN sebagai metode klasifikasi juga dinilai lebih efisien daripada metode Bayes dalam [2], [10]. Selain itu, Alpaydin [11] menyimpulkan bahwa KNN mempunyai performa yang efisien dan untuk beberapa kasus mempunyai akurasi terbesar.

Sifat lain terkait jumlah dan bentuk data yang menjadi bagian dari data latih, KNN menjamin dimungkinkannya kemudahan untuk proses pencegahan pada data latih agar klasifikasi tetap optimal dan tahan terhadap data latih yang mengandung derau [4]. Kim dkk. [12] mengungkapkan bahwa klasifikasi KNN lebih bagus dan stabil untuk memprediksi data cuaca yang hilang dalam penelitian mengenai prediksi PV (*photovoltaic*). Do dkk. [13] menyimpulkan KNN efisien dalam memprediksi data medis yang hilang

*) Penulis korespondensi (Anastasia Rita Widiarti)
Email: rita_widiarti@usd.ac.id

Submitted: 23 August 2019; Revised: 28 February 2020; Accepted: 13 March 2020; Published: 30 April 2020

dibandingkan metode lain dalam kaitannya dengan riset mengenai biomedis.

Pada kajian-kajian penelitian yang memanfaatkan KNN untuk klasifikasi di atas, beberapa kajian menyebutkan nilai variabel *k* yang digunakan seperti dalam [5], [6]. Namun, kajian lain tidak mendeskripsikan secara khusus [7], [9]. Untuk penelitian terkait alihaksara, kajian [7] tidak menyatakan variabel ciri citra yang digunakan secara spesifik untuk klasifikasi. Beberapa kajian juga membandingkan pemakaian rumus jarak untuk mengukur kedekatan seperti dalam [2], [5], [6]. Oleh beberapa peneliti disimpulkan bahwa rumus jarak Euclidean dinilai dapat menghasilkan akurasi yang optimal [2], [5]. Rumus jarak Euclidean menghasilkan pengukuran jarak yang lebih baik dibandingkan dengan rumus jarak lainnya, seperti jarak *cityblock*, *cosine distance*, dan *correlation* [2].

Berdasarkan pada hasil-hasil pemikiran utama mengenai dampak penggunaan variabel-variabel berbeda pada kajian di atas, penelitian ini mengkaji penggunaan metode klasifikasi KNN untuk alihaksara dari citra beraksara Batak, Jawa, dan Sunda ke dalam aksara Roman. Jarak Euclidean digunakan untuk mengukur jarak kedekatan ciri dari objek. Penelitian ini juga mengkaji pengaruh nilai *k* yang berbeda, pengaruh penggunaan ciri citra aksara yang berbeda, dan pengaruh banyaknya data latih untuk klasifikasi yang mempergunakan metode KNN tersebut. Nilai *k* yang digunakan adalah 1, 3, 5, 7, dan 9 yang merupakan kelipatan bilangan ganjil sesuai yang dinyatakan [14]. Penelitian ini juga menyelidiki apakah perbedaan variabel jumlah data latih berdampak pada hasil akurasi seperti yang dilakukan oleh Ong dan Suhartono [7].

Pemilihan citra aksara dari ketiga suku di Indonesia, yaitu dari suku Batak, Jawa, dan Sunda sebagai studi kasus, dilatarbelakangi pemikiran bahwa ketiga suku tersebut relatif dominan banyak penduduknya di Indonesia. Selain itu, di beberapa museum dan perpustakaan daerah di wilayah di Jawa dan Sumatra tersimpan banyak manuskrip yang bertuliskan aksara daerah tersebut di atas.

II. METODE PENELITIAN

Alihaksara citra aksara yang dilakukan pada penelitian ini adalah implementasi dari klasifikasi citra aksara tulisan tangan beraksara Batak, Jawa, dan Sunda menggunakan algoritme KNN. Data yang digunakan dalam penelitian secara garis besar dibagi menjadi data latih dan data uji. Tahap-tahap utama yang dilakukan dalam implementasi klasifikasi terdiri dari tahap pengambilan data (*data capturing*), penyiapan data, dan ekstraksi ciri. Proses pengolahan untuk data latih berakhir di tahap ekstraksi ciri. Pada data uji, ditambahkan satu tahap lagi, yaitu tahap klasifikasi data.

Tahap pengambilan data, baik untuk data uji dan data latih, dilakukan untuk digitalisasi data citra aksara yang digunakan dalam penelitian. Pada tahap ini, alat yang digunakan adalah sebuah *scanner*. Setelah data citra aksara digital diperoleh, data tersebut digunakan sebagai masukan di tahap penyiapan data untuk memproses setiap data agar mempunyai kesamaan properti. Tahap ekstraksi ciri dilakukan untuk memproduksi data baru yang mencirikan keunikan setiap citra aksara dalam kelompok yang sama. Semua tahap tersebut di atas berlaku sama di semua data, baik untuk data uji mapun data latih. Untuk data latih, hasil ektraksi ciri disimpan dalam dataset yang secara pokok mempunyai dua entitas, yaitu ciri dan label dari ciri yang terkait. Pada data uji, tahap terakhir klasifikasi data dilakukan untuk memproduksi secara otomatis label dari ciri citra data uji.

A. Proses penyiapan data uji dan data latih

Proses alihaksara citra aksara Batak, Jawa, dan Sunda dimulai dengan menyiapkan setiap data citra aksara, baik untuk data uji maupun data latih, agar setiap citra mempunyai kesamaan properti dalam warna, ketebalan, dan ukuran. Tahap ini dilakukan dalam empat proses utama. Proses pertama adalah binerisasi citra aksara sehingga properti warna hanya terdiri dari hitam atau putih saja. Proses kedua adalah proses untuk mereduksi sebanyak mungkin derau. Proses selanjutnya dilakukan penipisan citra aksara mempergunakan algoritme penipisan citra Rosenfeld. Proses terakhir adalah tahap untuk menyamakan ukuran citra ke ukuran 50x50 piksel. Ukuran piksel 50x50 dipilih karena dari percobaan dengan menggunakan berbagai ukuran citra aksara yang berbeda, ukuran tersebut menghasilkan akurasi yang paling optimal.

B. Ekstraksi ciri citra aksara

Setelah setiap citra aksara mempunyai kesamaan properti dalam warna, ketebalan, dan ukuran, maka citra-citra aksara tersebut siap untuk diproses lebih lanjut dalam proses ekstraksi ciri. Proses ekstraksi ciri dilakukan untuk memproduksi data baru yang unik dan mewakili dari setiap kelompok data aksara. Salah satu algoritme yang dapat dipakai untuk memproduksi ciri adalah algoritme *Intensity of Character (IoC).*

Prinsip kerja algoritme IoC adalah membagi data citra menjadi area-area tertentu. Dalam setiap area jumlah piksel objeknya dihitung seperti dalam [15]. Dalam penelitian ini, setiap citra awalnya dibagi menjadi 4, 5, 6, 7, dan 8 bagian, baik secara vertikal maupun horisontal. Pada setiap hasil pembagian, dilakukan percobaan untuk klasifikasi dengan *k* mulai dari 1 sampai dengan 9.

Pada setiap area hasil pembagian citra, misalnya dibagi menjadi 5x5, diperoleh unit-unit area sebanyak 25 area, dengan ukuran luas citra sebesar 10x10 piksel. Gambar 1 memperlihatkan representasi pembagian citra masukan dengan 25 area unitnya. Pada setiap unit area, digunakan Persamaan 1 untuk menghitung piksel hitamnya. Pada cuplikan data area unit paling atas kiri, diperoleh data jumlah piksel hitam sebesar 20 satuan piksel. Hal yang sama dilakukan lagi pada unit-unit bagian citra lainnya, sehingga untuk setiap citra diproduksi ciri citra berupa matrik berukuran 5x5.

Persamaan 1 diterapkan untuk menghitung jumlah objek dalam satu area luasan hasil pembagian. Jika `M` adalah citra aksara berukuran 50x50, dan `i,j` = 0, 1, 2, 3, 4, maka diperoleh data baru `C(i,j)` yang menunjukkan jumlah piksel hitam pada area yang bersesuaian.

$$C_{(i+1,\,j+1)} = \sum_{k=10i+1}^{k=10(i+1)} \sum_{p=10j+1}^{p=10(j+1)} M_{(k,\,p)} \qquad (1)$$

C. Klasifikasi dengan KNN

Tahap klasifikasi pada alihaksara dilakukan untuk memberikan label yang sesuai dengan nama kelompok atau label dari ciri data citra aksara yang ditemukan. Jika terdapat sekumpulan objek yang dimasukkan dalam data latih *T* dengan minimal 2 *field* untuk setiap objek data yang berupa data penciri objek (x_i) dan labelnya (y_i), dan data baru yang disebut data test x_0, maka klasifikasi adalah proses untuk memetakan data test x_0 ke data prediksi y_o dengan y_o adalah anggota dari *T*.

Metode klasifikasi yang digunakan dalam penelitian ini adalah KNN. Prinsip klasifikasi yang dipergunakan dengan algoritme KNN adalah menemukan kemiripan ciri citra data uji dengan ciri dari setiap citra data latih. Semakin banyak kemiripan ciri data uji dengan ciri dalam citra data latih di suatu kelas, maka citra aksara tersebut diberi label yang sesuai dengan label dari citra data latih terdekat.

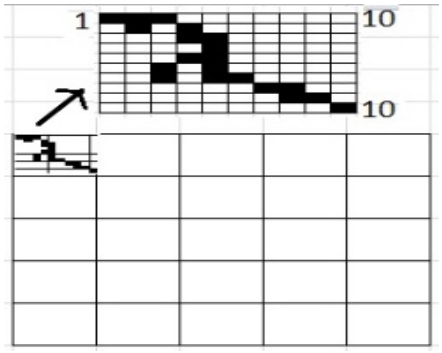
Untuk mengukur kemiripan ciri citra digunakan rumus jarak Euclidean. Setelah jarak dari citra data uji ke setiap ciri dari data latih diperoleh, dipilih label yang dominan muncul, yaitu label citra yang mempunyai jarak terkecil. Nilai *k* dalam konsep KNN menyatakan banyaknya objek yang diambil sebagai penentu dominasi label. Gambar 2 memberikan contoh ilustrasi prinsip kerja klasifikasi 3-NN.

Jika nilai *k* diambil 3, artinya dipilih 3 label dari 3 objek citra yang mempunyai 3 jarak terdekat dari objek baru yang diklasifikasikan. Jika terdapat sebaran objek sebanyak 7, yang terdiri dari 3 objek berlabel aksara jawa *ca*, 3 objek berlabel *pa*, dan sebuah objek baru yang belum diketahui labelnya yang dinyatakan sebagai bintang (★), maka objek bintang akan diberi label *ca*. Label *ca* dipilih karena tiga objek terdekat dari objek bintang didominasi oleh objek berlabel *ca*.

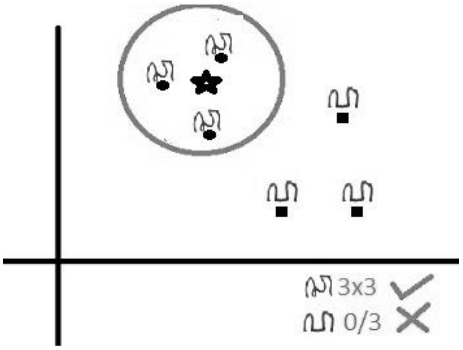
Dalam penelitian ini, nilai *k* yang digunakan mulai dari 1, 3, 5, 7, dan 9. Implementasi klasifikasi mengunakan fungsi *fitcknn* yang sudah tersedia di bahasa pemrograman Matlab. Rumus jarak Euclidean digunakan untuk menghitung jarak antar objek.

D. Evaluasi hasil pengujian

Pada penelitian ini, cara evaluasi yang digunakan untuk mengukur keberhasilan alihasaksara citra beraksara Batak, Jawa, dan Sunda menggunakan konsep 3-*fold* dan 10-*fold cross validation*. Metode k-*fold cross validation* ini digunakan untuk mengevaluasi sebuah sistem dengan membagi jenis data yang tersedia menjadi *k* dataset. Jumlah data dan ragam data di setiap



Gambar 1. Representasi pembagian unit area citra



Gambar 2. Diagram penyebaran data dan hasil klasifikasi dengan 3-NN

dataset sama. Pada saat pengujian, secara bergantian akan digunakan kombinasi dataset menjadi data latih dan data uji. Jumlah dataset yang digunakan sebagai data uji adalah `1/k`.

Dalam penelitian ini, apabila digunakan nilai `k=3`, maka persentase jumlah data latih sebanyak 66,67 % dan data uji sebanyak 33,33 % dari data yang tersedia untuk setiap kelas data yang sama. Pengujian berlangsung sebanyak 3 kali dengan kombinasi data latih dan data uji yang berbeda. Hal yang sama berlaku untuk 10-*fold*, yaitu bahwa persentase data latih sebesar 90 %, dan data uji sebesar 10 %. Pengujian ini dilakukan juga mengetahui pengaruh jumlah data latih terhadap hasil klasifikasi KNN.

III. HASIL DAN PEMBAHASAN

A. Dataset citra aksara

Dataset yang digunakan dalam penelitian ini diperoleh dari hasil digitalisasi tulisan tangan yang dilakukan pada selembar kertas. Sebanyak 90 responden dipilih dan setiap responden menuliskan aksara Jawa, Batak, atau Sunda. Setiap responden menulis aksara dengan melihat contoh ditunjukkan pada responden.

Jumlah data citra yang diperoleh adalah sebanyak 2520 data citra yang digunakan sebagai data latih dan data uji. Dari 2520 data tersebut, jumlah citra aksara Jawa 600 buah, jumlah citra aksara Batak 1170 buah, dan citra aksara Sunda 750 buah. Untuk setiap data citra aksara yang unik, masing-masing digunakan sebanyak 30 data.

B. Hasil pengujian dengan *3-fold cross validation*

Pengujian pertama klasifikasi KNN diterapkan untuk citra aksara Jawa, dengan banyaknya data sebesar 600 citra. Empat ratus data citra menjadi data latihan, dan 200 data citra sisanya menjadi data uji. Pengujian pertama ini dilakukan untuk mengetahui nilai variabel ukuran citra dan variabel *k* yang menghasilkan akurasi paling optimal. Hasil akhir rata-rata persentase akurasi uji 3-*fold* juga digunakan sebagai data pembanding pada hasil uji akurasi 10-*fold* untuk mengetahui pengaruh dari jumlah data latih pada nilai akurasi.

Untuk mencari nilai variabel ukuran citra yang optimal, area pada setiap citra aksara Jawa dibagi menjadi 4x4, 5x5, 6x6, 7x7, dan 8x8. Percobaan ini bertujuan untuk memperoleh kondisi pembagian area citra yang terbaik dalam hal akurasi. Hasil dari percobaan digunakan sebagai patokan ukuran pembagian citra pada data set citra aksara Batak dan Sunda.

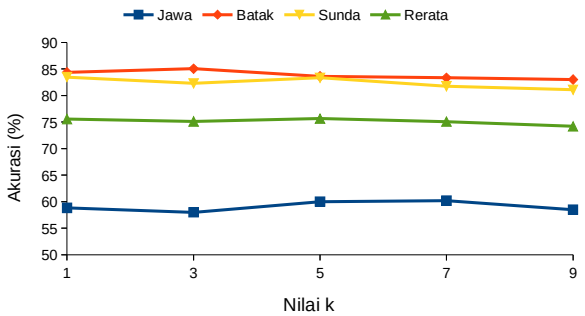
Jumlah tetangga atau nilai *k* diatur mulai dari 1, 3, 5, 7, dan 9 tetangga yang merupakan kelipatan ganjil sesuai [14]. Tujuan dari percobaan tersebut adalah untuk menguji apakah variabel nilai *k* yang diatur berbeda akan mempengaruhi akurasi. Tabel 1 memperlihatkan hasil uji coba KNN untuk alihaksara menggunakan kombinasi variabel nilai *k* dan variabel ukuran pembagian citra.

Data pengujian pada dataset aksara Jawa menunjukkan bahwa pembagian ukuran area, yang dinyatakan dalam kolom *Dm*, mempunyai pengaruh dalam persentase akurasi alihaksara. Pembagian area objek citra sebanyak 7x7 luasan menghasilkan akurasi yang paling tinggi, yaitu sebesar 60,17 %, dengan *k* bernilai 7. Informasi ukuran luasan unit ini menjadi data untuk menguji data tunggal dan penyelidikan dengan 3-*fold* untuk aksara Batak dan Sunda. Hasil pengujian ini dinyatakan dalam Tabel 1 di bagian baris aksara Batak dan aksara Sunda, yaitu dengan ukuran 7x7.

Dari hasil pengujian ini, dapat disimpulkan bahwa tidak terdapat perbedaan besar pada nilai akurasi, jika digunakan nilai *k* berbeda-beda pada setiap dimensi ukuran citra. Perbedaan nilai akurasi rata-rata hanya sebesar 2,25 % di setiap luasan unit dari skala 0-100 %. Namun, jika dihitung rerata akurasi secara total pada setiap nilai *k* di ukuran luasan 7x7 untuk semua dataset, dapat dinyatakan bahwa persentase nilai akurasi terbesar terjadi pada saat diterapkan di *k* bernilai 5 sebesar 75,64 % dan akurasi terendah di *k* bernilai 9 sebesar 74,19 %. Gambar 3 menunjukkan perbedaan persentase akurasi secara keseluruhan pada klasifikasi citra aksara Jawa, Batak, dan Sunda dengan ukuran luasan citra 7x7 untuk setiap nilai *k* berbeda. Perbedaan nilai hanya sebesar 1,45 %. Hal ini sesuai dengan [14] yang menyatakan bahwa akurasi klasifikasi tidak akan sangat berbeda nilainya pada *k* bernilai 1, 3, 5, 7, dan 9.

Tabel 1. Akurasi alihaksara citra aksara Jawa, Batak, dan Sunda dengan 3-*fold cross validation*

Aksara	*Dm*	k	Rerata akurasi (%)			Rerata
			Set 1	Set 2	Set 3	(%)
Jawa	4x4	1	55	57,5	57	56,50
		3	54	55,5	63,5	57,67
		5	54,5	57,5	65	59,00
		7	51,5	60	61	57,50
		9	53	61	62	58,67
	5x5	1	50	57,5	60,5	56,00
		3	48,5	56,5	56	53,67
		5	52	57,5	56	55,17
		7	52,5	55,5	55,5	54,50
		9	51	55	56	54,00
	6x6	1	55	58,5	63	58,83
		3	52,5	59	59	56,83
		5	50,5	61,5	63,5	58,50
		7	55	57	63	58,33
		9	54	55	64,5	57,83
	7x7	1	60,5	60,5	55,5	58,83
		3	53,5	61,5	59	58,00
		5	57	56	67	60,00
		7	55	58,5	67	60,17
		9	55	57,5	63	58,50
	8x8	1	49,5	57	59,5	55,33
		3	45,5	56	56,5	52,67
		5	46	54	57,5	52,50
		7	44,5	56	56,5	52,33
		9	47,5	52	57,5	52,33
Batak	7x7	1	76,9	87,2	89,0	84,36
		3	77,7	86,4	91,0	85,04
		5	74,9	86,2	89,7	83,59
		7	74,6	85,1	90,3	83,33
		9	74,4	86,2	88,5	82,99
Sunda	7x7	1	82,8	84,0	83,6	83,47
		3	80,0	84,0	82,8	82,27
		5	82,4	86,0	81,6	83,33
		7	80,8	82,8	81,6	81,73
		9	80,8	82,8	79,6	81,07



Gambar 3. Akurasi uji 3-*fold cross validation* pada ukuran unit citra 7x7 untuk semua dataset

C. Hasil pengujian dengan *10-fold cross validation*

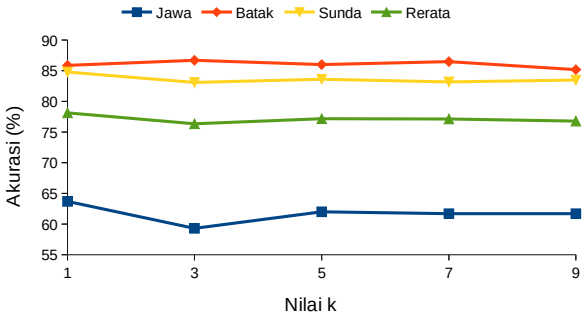
Pengujian *10-fold* dilakukan untuk mengetahui pengaruh jumlah data latih pada nilai akurasi dan pengaruh pemberian nilai *k* yang berbeda dari 1, 3, 5, 7, dan 9. Data citra yang digunakan dalam uji ini dibagi menjadi 10 kelompok dan dibuat 10 set data kombinasi data latih dan data uji. Tabel 2 memperlihatkan nilai persentase akurasi untuk evaluasi dengan 10-*fold cross*

Tabel 2. Akurasi alihaksara citra aksara Jawa, Batak, dan Sunda dengan 10-*fold cross validation*

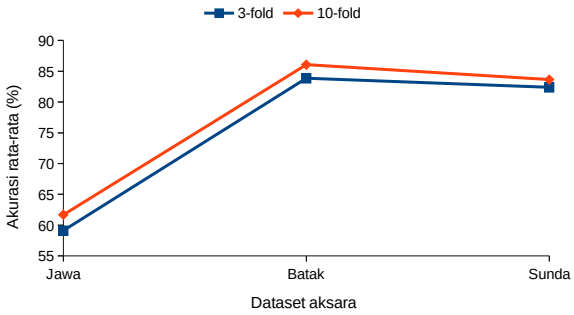
Aksara	*Dm*	k	Akurasi rata-rata di dataset ke- (%)										Rerata (%)
			1	2	3	4	5	6	7	8	9	10	
Jawa	4	1	57	48	58	58	60	57	78	62	65	50	59,3
	x	3	52	52	60	58	63	60	70	58	75	53	60,2
	4	5	55	55	70	60	68	60	73	57	80	58	63,7
		7	58	50	58	60	75	57	72	60	78	52	62,0
		9	57	52	60	57	68	60	73	53	78	53	61,2
	5	1	52	62	62	58	63	63	77	60	63	65	62,5
	x	3	53	58	48	43	58	67	75	50	70	55	57,8
	5	5	55	63	53	55	60	73	75	48	72	63	61,8
		7	57	62	50	53	63	72	75	42	70	60	60,3
		9	55	58	50	55	62	77	75	48	67	57	60,3
	6	1	50	57	67	62	72	70	78	65	73	43	63,7
	x	3	47	55	62	70	70	65	78	62	75	52	63,5
	6	5	38	52	62	65	70	68	78	67	78	52	63,0
		7	47	48	58	63	77	62	80	67	80	58	64,0
		9	50	50	60	65	73	60	80	65	83	53	64,0
	7	1	53	57	68	63	73	68	77	62	65	50	63,7
	x	3	53	57	55	57	68	58	73	57	72	43	59,3
	7	5	57	57	55	57	67	62	75	68	82	42	62,0
		7	58	60	53	58	63	50	82	63	88	40	61,7
		9	57	60	53	57	63	53	78	62	85	48	61,7
	8	1	48	52	55	60	65	63	68	68	68	48	59,7
	x	3	55	52	47	55	63	65	73	58	70	58	59,7
	8	5	52	52	48	57	68	62	70	57	68	53	58,7
		7	50	53	47	55	63	60	73	58	70	50	58,0
		9	45	50	47	62	63	60	67	60	77	52	58,2
Batak	7	1	78	73	80	85	84	97	91	93	88	91	85,9
	x	3	77	79	84	86	82	93	92	91	87	94	86,7
	7	5	77	78	83	82	79	95	91	90	91	94	86,0
		7	78	75	84	80	79	97	92	92	92	95	86,5
		9	76	73	84	77	79	96	92	91	90	94	85,2
Sunda	7	1	83	89	87	87	83	77	83	87	87	87	84,8
	x	3	80	84	80	84	85	84	85	84	84	80	83,1
	7	5	83	85	83	81	85	84	87	83	83	83	83,6
		7	81	84	84	80	83	87	89	83	81	80	83,2
		9	83	84	84	83	84	85	88	81	80	83	83,5

validation pada dataset citra aksara Jawa, Batak, dan Sunda. Hasil tersebut menunjukkan bahwa terdapat 10 nilai akurasi yang berbeda untuk setiap penggunaan nilai *k* berbeda dan setiap ukuran luasan atau dimensi citra (*Dm*), mulai dari 4x4 sampai 8x8 untuk seluruh dataset citra. Nilai persentase rata-rata akurasi di kolom Rerata adalah rata-rata hasil uji coba sistem pada nilai *k* tertentu yang bersesuaian.

Dengan kombinasi cara pembagian area dan kombinasi *k* bernilai 1, 3, 5, 7, dan 9, maka pada setiap dataset berbeda ditemukan 25 hasil uji akurasi. Perbedaan nilai rata-rata akurasi untuk setiap luasan di dataset yang sama sebenarnya juga tidak terlalu besar, seperti disimpulkan pada uji *3-fold*. Namun, ukuran luasan citra, nilai *k*, dan kombinasi data latih dan data uji yang berbeda menentukan akurasi yang diperoleh sesuai dengan [7], [14]. Jika dihitung rerata akurasi untuk seluruh kombinasi nilai *k* untuk semua dataset pada luasan unit 7x7, diperoleh akurasi tertinggi pada *k*



Gambar 4. Akurasi uji 10-*fold cross validation* pada ukuran unit citra 7x7 untuk semua dataset



Gambar 5. Akurasi rata-rata hasil evaluasi 3-fold dan 10-*fold* terhadap aksara

bernilai 1 sebesar 78,13 %, seperti ditunjukkan dalam Gambar 4. Hal ini dapat disimpulkan bahwa jika jumlah dataset lebih banyak, maka cukup diperlukan satu tetangga terdekat sebagai rujukan untuk mengklasifikasi data.

Berdasarkan besarnya nilai akurasi pada luasan 7x7 untuk setiap dataset, dapat disimpulkan bahwa nilai akurasi untuk dataset aksara Jawa paling kecil. Bahkan, perbedaan rentang nilainya sampai lebih besar dari 20, seperti terlihat pada Gambar 5. Perbedaan yang sangat mencolok terjadi antara rerata persentase akurasi hasil alihaksara pada citra aksara Jawa, dan citra aksara Batak atau Sunda. Unjuk kerja sistem untuk dataset citra aksara Batak dan Sunda di atas 80%, sedangkan untuk dataset citra aksara Jawa di bawah 70 %. Perbedaan besar hasil akurasi pada dataset citra aksara Jawa disebabkan karena kualitas citra untuk dataset aksara Jawa kurang baik, seperti ditunjukkan contoh citra pada Gambar 6. Data citra masukan untuk uji sistem menunjukkan 30 citra aksara Jawa *ga* yang berbeda-beda propertinya, baik dalam ukuran, kemiringan maupun ketebalan.

Walaupun sistem alihaksara yang diimplementasikan sudah mengakomodasi proses menyiapkan dataset citra, yaitu agar semua citra dalam dataset mempunyai kesamaan properti dalam warna, ketebalan, dan ukuran, namun masih sulit untuk memproduksi citra yang mempunyai properti yang sama. Hal ini karena bentuk asli citra tulisannya untuk menggambarkan aksara yang sama sudah sangat berbeda. Perbedaan tersebut

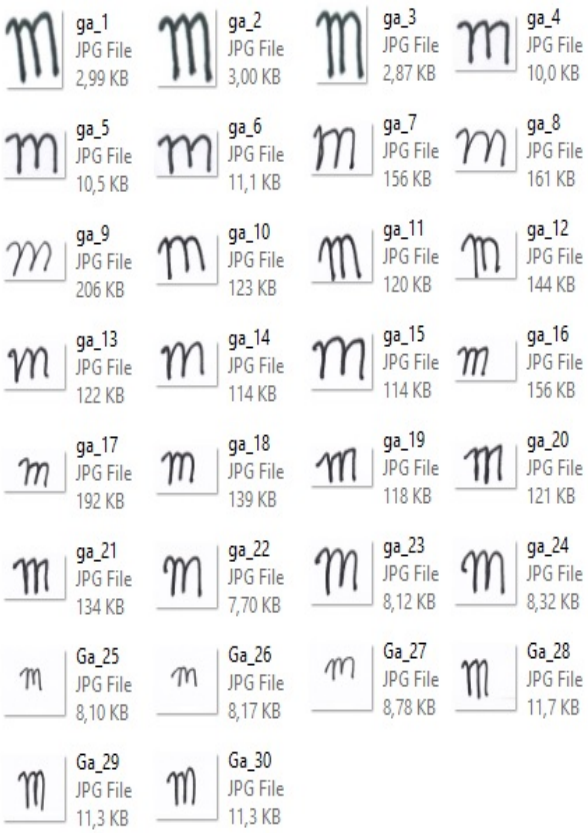
menyebabkan keluaran hasil alihaksara berbeda, karena sifat dari ciri yang diambil tergantung sepenuhnya pada bentuk citra di suatu area.

D. Pembahasan hasil uji data uji dengan *3-fold cross validation* dan *10-fold cross validation*

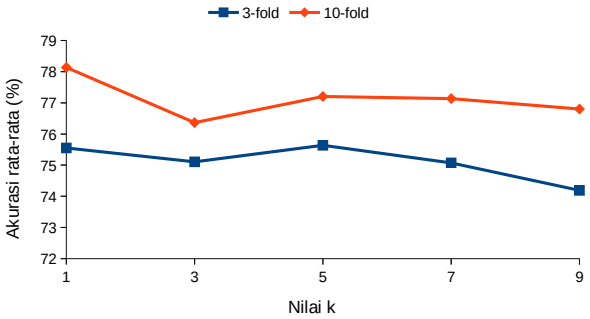
Gambar 7 menunjukkan perbandingan rerata akurasi klasifikasi menggunakan KNN dengan nilai *k* dari 1, 3, 5, 7, dan 9 dari data pada keseluruhan uji akurasi 3-*fold* dan 10-*fold*. Secara keseluruhan akurasi 10-*fold* lebih tinggi dari nilai 3-*fold*. Hal ini berarti bahwa banyaknya data latih yang digunakan mempengaruhi hasil akurasi sesuai dengan [7]. Hasil ini menunjukkan bahwa dengan semakin banyaknya data latih untuk setiap kelompok aksara, cukup dibutuhkan satu model saja untuk mendapatkan informasi kelasnya seperti [11].

Secara umum, dapat dinyatakan bahwa nilai *k* mempunyai dampak di hasil akurasi. Namun, dalam penelitian ini ternyata tidak ada kesimpulan mutlak terkait besarnya nilai *k* yang dapat dijadikan patokan agar akurasi optimal karena nilai akurasi optimal di uji 3-*fold* dan 10-*fold* diperoleh pada kondisi nilai *k* yang berbeda-beda. Akurasi tertinggi di uji 3-*fold* diperoleh saat *k* bernilai 5, sedangkan di uji 10-*fold* pada saat *k* bernilai 1. Secara khusus, kinerja dan efisiensi klasifikasi aksara Nusantara menggunakan KNN dapat dibandingkan dengan metode lain seperti SVM dan Bayes [2], [8]-[11].

Meskipun dengan uji 10-*fold* diperoleh unjuk kerja sistem alihaksara untuk dataset citra Jawa masih di bawah 70 %, namun sistem alihaksara ini dapat dikembangkan sebagai alat bantu pembaca manuskrip beraksara daerah untuk generasi sekarang, seperti halnya [4], [7] untuk manuskrip tulisan tangan aksara Roman dan [8] untuk aksara Yoruba. Hasil penelitian ini diharapkan bisa memberikan kontribusi secara luas mengenai pemanfaatan teknologi pengolahan citra dan pengenalan pola untuk pelestarian dan penyebarluasan kekayaan budaya tulis dalam manuskrip di Indonesia. Selain itu, implementasi teknologi alihaksara yang dikembangkan dalam penelitian ini dapat dimanfaatkan untuk berbagai kepentingan terkait dengan pemanfaatan informasi dalam bentuk tulisan tersebut.



Gambar 6. Dataset citra aksara Jawa *ga*



Gambar 7. Rerata akurasi hasil evaluasi 3-*fold* dan 10-*fold* terhadap terhadap nilai *k*

IV. KESIMPULAN

Perbedaan besar nilai *k* dalam metode klasifikasi KNN mempunyai pengaruh pada hasil unjuk kerja sistem. Dalam penelitian ini, tidak dapat disimpulkan nilai *k* yang paling optimal dari *k* = 1, 3, 5, 7, dan 9 karena hasil uji akurasi dengan 3-*fold* dan 10-*fold* menunjukkan perbedaan. Banyaknya data latih per kelompok dalam dataset latih juga mempunyai dampak dalam unjuk kerja klasifikasi dengan KNN. Semakin banyak dataset pada data latih di tiap kelompok yang sama, maka akan meningkatkan hasil akurasi klasifikasi.

Untuk penerapan klasifikasi KNN pada data citra, dalam hal ini untuk alihaksara citra, perlu studi yang lebih mendalam untuk mengekstrak ciri yang akan diolah. Apabila digunakan ciri berbasis statistika, dalam hal ini dengan menghitung jumlah piksel hitam di suatu luasan citra tertentu, diperlukan proses penyiapan data yang lebih baik. Lebih lagi jika dataset yang digunakan sangat bervariasi, meskipun berada di kelas yang sama. Perlu cara ekstraksi ciri lain yang bisa mewakili gerak tangan dalam menuliskan aksaranya.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh Universitas Sanata Dharma dengan nomor kontrak: 042/LPPM USD/V/2019. Penulis juga mengucapkan terimakasih pada Phalita Nariwastu (TI-99), William Sianturi (TI-15), Jerry Ferdiano (TI-15), dan Osmond Giovany (TI-

15) untuk bantuan dataset yang digunakan sebagai studi kasus dalam penelitian ini dan bantuan untuk pengolahan data di penelitian ini.

DAFTAR PUSTAKA

[1] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification" *IEEE Transactions on Information Theory*, vol. 13, pp. 21-27, 1967. doi: 10.1109/TIT.1967.1053964

[2] A. Kataria and M. D. Singh, "A review of data classification using k-nearest neighbour algorithm," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 6, pp. 354-360, 2013.

[3] N. Bhatia and A. Vandana, "Survey of nearest neighbor techniques," *International Journal of Computer Science and Information Security*, vol. 8, no. 2, pp. 302-305, 2010.

[4] S. Roy, dan M. Saravanan, "Handwritten character recognition using k-nn classification algorithm," *International Journal of Advance Research and Innovative Ideas in Education*, vol 3, no. 5, pp. 1245-1250, 2017.

[5] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 1, pp. 54-48, 2020. doi: 10.14710/jtsiskom.8.1.2020.54-58

[6] C. Premavathi and P. Thangaraj, "Efficient hand-dorsa vein pattern recognition using knn classification with completed histogram cb in tp feature descriptor," *International Journal of Recent Technology and Engineering*, vol. 7, no. 4, pp. 50-55, 2018.

[7] V. Ong and D. Suhartono, "Using k-nearest neighbor in optical character recognition," *ComTech: Computer, Mathematics, and Engineering Applications*, vol. 7, no. 1, pp. 53-65, 2016. doi: 10.21512/comtech.v7i1.2223

[8] J. F. Ajao, D. O. Olawuyi, and O. O. Odejobi, "Yoruba handwritten character recognition using freeman chain code and k-nearest neighbor classifier," *Jurnal Teknologi dan Sistem Komputer*, vol. 6, no. 2, pp. 129-134, Oct. 2018. doi: 10.14710/jtsiskom.6.4.2018.129-134

[9] Z. Al-asady and A. Al-amery, "Human action recognition using a corners and blob detector with different classification methods," *IOP Conference Series: Materials Science and Engineering*, vol. 518, no. 5, pp. 1-9, 2019. doi: 10.1088/1757-899X/518/5/052008

[10] Y. Hamamoto, S. Uchimura and S. Tomita, "A bootstrap technique for nearest neighbor classifier design," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 19, no. 1, pp. 73-79, 1997. doi: 10.1109/34.566814

[11] E. Alpaydin, "Voting over multiple condensed nearest neighbors," *Artificial Intelligence Review*, vol. 11, pp. 115-132, 1997. doi: 10.1007/978-94-017-2053-3_4

[12] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead pv generation forecasting," *Applied Sciences*, vol. 9, no. 1, pp. 1-18, 2019. doi: 10.3390/app9010204

[13] K. T. Do et al., "Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies," *Metabolomics*, vol. 128, pp. 1-18, 2018. doi: 10.1007/s11306-018-1420-2.

[14] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A.A. Alhasanat, "Solving the problem of the k parameter in the knn classifier using an ensemble learning approach," *International Journal of Computer Science and Information Security*, vol. 12, no. 8, pp. 33-39, 2014.

[15] S. Mirah and A. R. Widiarti, “Automatic recognition of the NIK in electronic KTP,” in *1st International Conference on Science and Technology for an Internet of Things*, Yogyakarta, Indonesia, Oct. 2018, pp. 1-11. doi: 10.4108/eai.19-10-2018.2282544