PENGGUNAAN SELEKSI FITUR UNTUK KLASIFIKASI BENIH PADI RAWA KALIMANTAN SELATAN BERDASARKAN CIRI FISIK

Muhammad Syahid Pebriadi[1)], Nahdi Saubari[2)]

[1,2]Informatika, Universitas Muhammadiyah Banjarmasin
Jl. Gubernur Sarkawi, Alalak, Barito Kuala, Banjarmasin
Email: [1]m.syahid.f@umbjm.ac.id, [2]nahdisaubari@umbjm.ac.id

Abstrak

Indonesia memiliki ragam padi (Oryza sativa L.) sebanyak 4000 ragam lebih yang tersimpan di bank gen Balai Besar Biogen. Penelitian ini bertujuan mengklasifikasikan benih padi rawa Kalimantan Selatan menggunakan algoritma klasifikasi KNN tanpa dan dengan seleksi fitur serta membandingkan akurasinya. Metode penelitian meliputi Pengumpulan Data, Pembuatan Data Sintetis dengan nilai statistik, Seleksi Fitur dengan algoritma ReliefF, Pembagian Data Latih dan Data Uji dengan Stratified k-fold cross validation, pemodelan menggunakan algoritma KNN serta perhitungan akurasi. Hasil penelitian berupa tiga fitur yang optimal hasil seleksi fitur meliputi AspectRatio, MajorAxis, dan Feret dengan nilai bobot berturut-turut 0,489; 0,485; dan 0,456. Rata-rata akurasi paling tinggi didapatkan pada saat nilai k=3 sebesar 71% untuk Data Sintetis tanpa Seleksi Fitur sedangkan akurasi algoritma KNN terhadap Data Sintetis dengan Seleksi Fitur untuk nilai k=3 sebesar 74%, nilai k=5 sebesar 75% dan k=7 sebesar 76%. Jika dibandingkan dengan Data Sintetis tanpa Seleksi Fitur, maka penambahan akurasi algoritma KNN untuk nilai k=3 sebesar 3%, serta untuk nilai k=5 dan k=7 sebesar 6%. Hal ini menunjukkan bahwa penggunaan seleksi Fitur dapat meningkatkan akurasi algoritma KNN untuk mengklasifikasikan data benih padi rawa Kalimantan Selatan berdasarkan ciri fisiknya.

Kata Kunci: *Akurasi, klasifikasi, KNN, seleksi fitur.*

1. Pendahuluan

Indonesia merupakan salah satu negara yang menjadikan nasi sebagai makanan pokoknya, hal ini berkaitan dengan jumlah produksi padi Indonesia yang menempati urutan ke-3 setelah China dan India [1]. Selain itu, ragam padi (Oryza sativa L.) tercatat sebanyak 4000 ribu ragam lebih disimpan sebagai koleksi plasma nutfah yang disimpan di bank gen Balai Besar Biogen [2]. Untuk memprediksi hasil dan kualitas padi, maka diperlukan identifikasi kelas padi untuk pemulia tanaman.

Para ahli yang berpengalaman, menentukan kelas padi melalui karakteristik visualnya. Cara ini memiliki kelemahan dikarenakan faktor subjektifitas dan kesalahan dari masing-masing ahli. Para ahli pertanian mengklasifikasikan padi berdasarkan uji perbedaan, keseragaman dan stabilitas yang menggunakan begitu banyak mesin. Hal ini berpengaruh terhadap biaya, sumber daya, tenaga dan waktu yang digunakan. Oleh karena itu diperlukan pendekatan objektif untuk mengklasifikasikan benih padi.

Beberapa peneliti melakukan pendekatan objektif menggunakan teknik pengolahan citra dan teknik klasifikasi menggunakan algoritma seperti *k-nearest neighbor* (KNN), *Support Vector Machine* (SVM), *Random Forest* (RF) dan *Artificial Neural Network* (ANN) [3] [4] [5]. Teknik klasifikasi memerlukan data yang pada umumnya terdiri atas data latih dan data uji. Setiap data memiliki atribut atau fitur di tambah label kelas. Seringkali atribut atau fitur yang ada tidak berpengaruh signifikan terhadap hasil klasifikasi sehingga diperlukan proses seleksi fitur.

Proses seleksi fitur merupakan tahapan penting yang biasanya dilakukan pada proses klasifikasi dengan data berdimensi tinggi. Seleksi fitur merupakan suatu permasalahan untuk menemukan optimal atau sub optimal irisan dari sebagian fitur terhadap keseluruhan fitur. Seleksi fitur sangat penting untuk mengeluarkan fitur yang tidak berhubungan dan redundan. Hal tersebut membuat kompleksitas sistem berkurang dan meningkatkan akurasi [6].

Penelitian ini bertujuan mengklasifikasikan benih padi rawa Kalimantan Selatan menggunakan proses seleksi fitur dan algoritma KNN. Data benih padi yang digunakan merupakan benih padi rawa Kalimantan Selatan [7]. Selanjutnya, data sintetis di buat menggunakan data tersebut. Model klasifikasi KNN menggunakan data sintetis tanpa seleksi fitur dan dengan seleksi fitur. Hasil akurasi kedua model tersebut nantinya dibandingkan untuk menentukan mana model yang terbaik.

2. Metode

Pengumpulan Data

Data yang digunakan pada penelitian ini diperoleh dari Penelitian Soesanto, dkk [7]. Data ini berupa data benih padi rawa lokal Kalimantan Selatan sebanyak 72 jumlah data yang terdiri atas 8 varietas yaitu Bayar Papuyu, Bayar Putih, Benih Kuning, Benih Putih, Ketan, Siam Gadis, Siam Unus dan Karan Dukuh. Data ini memiliki 7 fitur meliputi Area, Perimeter, MajorAxis, MinorAxis, Circularity, AspectRatio, Roundness, dan Feret.

Pembuatan Data Sintetis

Data sintetis dibuat dengan cara memanfaatkan nilai statistik dari data asli. Pertama-tama, setiap fitur dicari nilai minimum dan nilai maksimalnya. Kemudian nilai 0

atau 1 di proses secara acak. Jika nilai 0 maka tambahkan nilai minimum dengan hasil perkalian antara bilangan acak dari 0-1 dengan selisih nilai minimum dan nilai maksimum. Jika nilai 1 maka kurangkan nilai maksimum dengan hasil perkalian antara bilangan acak dari 0-1 dengan selisih nilai minimum dan nilai maksimum. Sehingga data sintetis yang dihasilkan memiliki dimensi matriks 100x8 untuk fitur data dan dimensi matriks 100x1 untuk kelas data. Jumlah data untuk masing-masing kelas berjumlah 100 jumlah data.

Seleksi Fitur

Seleksi fitur merupakan proses yang melibatkan subset dari kumpulan fitur yang menghasilkan keluaran seperti keseluruhan kumpulan fitur. Seleksi fitur biasanya digunakan untuk memilih fitur yang optimal, mereduksi dimensi, meningkatkan akurasi algoritma klasifier, dan menghapus fitur yang tidak relevan [8]. Salah satu algoritma seleksi fitur yang bisa digunakan adalah algoritma ReliefF. Algoritma ini memanfaatkan teknik pembobotan untuk mengukur signifikansi fitur dalam konteks klasifikasi. Fitur yang dipilih adalah fitur yang memiliki bobot paling besar [9]. Pseudocode algoritma ReliefF lengkap [10] dapat dilihat pada Gambar 1.

Algorithm ReliefF
Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimationas of the qualities of attributes
1. *set all weights* $W[A] := 0.0$;
2. *for* $i := 1$ *to m do begin*
3. *randomly select an instance* R_i;
4. *find k nearest hits* H_j;
5. *for each class* $C \neq class\ (R_i)$ *do*
6. *from class C find k nearest misses* $M_j(C)$;
7. *for* $A := 1$ to a do
8. $W[A] := W[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)/(m.k) +$
9. $\sum_{C \neq class(R_i)} \left[\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} diff\left(A, R_i, M_j(C)\right)\right]/(m.k)$;
10. *end;*

Gambar 1. Pseudocode algoritma ReliefF secara umum

Pembagian Data Latih dan Data Uji

Teknik pembagian data latih dan data uji adalah dengan stratified k-fold cross validation [11], dengan k yang digunakan adalah 5. Teknik ini akan membagi data menjadi 5 sub sample terpisah, dengan 4 sub sample akan digunakan sebagai data latih dan 1 sub sample akan digunakan sebagai data uji. Gambar 2 menunjukkan ilustrasi stratified k-fold cross validation.

Percobaan	Banyaknya Data				
1	S1	S2	S3	S4	S5
2	S1	S2	S3	S4	S5
3	S1	S2	S3	S4	S5
4	S1	S2	S3	S4	S5
5	S1	S2	S3	S4	S5

Gambar 2. Ilustrasi Stratified k-fold cross validation

Algoritma K-Nearest Neighbor

K-Nearest Neighbor (k-NN atau KNN) dapat dikategorikan sebagai algoritma instance-based learning yaitu algoritma yang melakukan pembelajaran berdasarkan data. Jika suatu data x tanpa label diinputkan, maka KNN akan menghitung jarak semua data yang sudah ada terhadap data x tersebut. Kemudian sebanyak k data yang memiliki jarak paling dekat dengan data x dipilih untuk menentukan label data x. Sebanyak k data yang dipilih tersebut disebut dengan *Nearest Neighbor* [12].

Untuk mendefinisikan jarak antara dua titik yaitu titik pada data training (x) dan titik pada data testing (y) maka digunakan rumus pengukuran jarak. Salah satu persamaan jarak yang sering digunakan adalah *Euclidean Distance*.

$$D\ x, y\ = \sqrt{\sum_{k-1}^{n} (x_k - y_k)^2}$$

Dengan D adalah jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi, dimana x=x1,x2,…,xi dan y=y1,y2,…,yi dan I merepresentasikan nilai atribut serta n merupakan dimensi atribut.

Pada fase training, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi data training sample. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk testing data (yang klasifikasinya tidak diketahui). Jarak dari vektor baru yang ini terhadap seluruh vektor training sample dihitung dan sejumlah k buah yang paling dekat diambil.

Langkah-langkah untuk menghitung metode Algoritma *k-Nearest Neighbour* adalah sebagai berikut:

a. Menentukan Parameter k (Jumlah tetangga paling dekat)
b. Menghitung jarak masing-masing objek terhadap data sampel yang diberikan menggunakan *Euclidean Distance*.
c. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak terkecil.
d. Mengumpulkan kategori Y (Klasifikasi *Nearest Neighbour*).
e. Dengan menggunakan kategori *Nearest Neighbour* yang paling mayoritas maka dapat diprediksi nilai label yang telah dihitung.

Perhitungan Akurasi

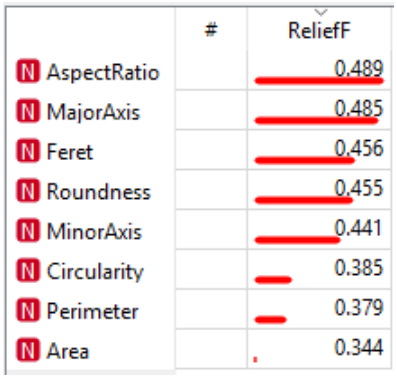
Perhitungan akurasi dilakukan dengan cara menjumlahkan data uji yang berhasil diidentifikasi dengan benar dibagi total data yang digunakan untuk pengujian.

$$Akurasi = \frac{\Sigma\ data\ pengujian\ benar\ klasifikasi}{\Sigma\ total\ data\ pengujian}\ x\ 100\%$$

3. Hasil dan Pembahasan

Hasil Seleksi Fitur

Seleksi fitur menggunakan algoritma ReliefF dilakukan dengan bantuan tool Orange Data Mining. Algoritma ini mengurutkan fitur dari yang paling berpengaruh (optimal) sampai sedikit berpengaruh. Tiga fitur yang optimal hasil algoritma ini meliputi AspectRatio, MajorAxis, dan Feret dengan nilai bobot berturut-turut 0,489; 0,485; dan 0,456. Hasil penilaian bobot dari setiap fitur secara lengkap dapat dilihat pada Gambar 3.



Gambar 3. Hasil Seleksi Fitur

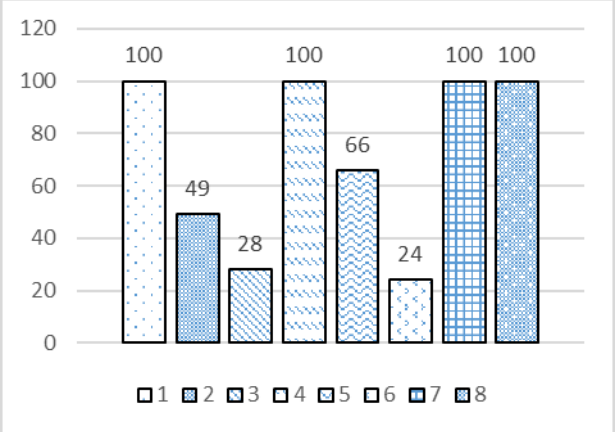
Pengujian KNN Tanpa Seleksi Fitur

Hasil pengujian menggunakan jumlah k = 3 menunjukkan bahwa algoritma KNN dapat mengklasifikasikan secara benar data benih padi dengan varietas Bayar Papuyu, Benih Putih, Siam Gadis dan Siam Unus. Data benih padi dengan varietas Bayar Putih masih salah klasifikasi di kelas Benih Kuning (28 data) dan Ketan (23 data). Data Benih Kuning dan Ketan lebih 50% salah klasifikasi di kelas lain. Sedangkan untuk Data Karan Dukuh salah klasifikasi di kelas Bayar Putih (19 data) dan kelas Benih Kuning (10 data).

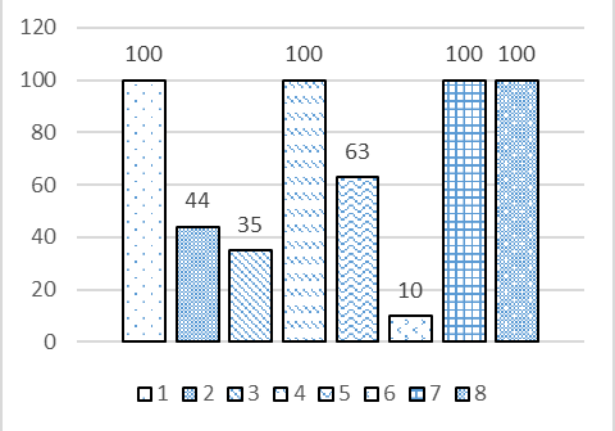
Akurasi yang didapat algoritma KNN dengan jumlah k=5 sangat tinggi di beberapa kelas. Akurasi pada kelas Bayar Papuyu, Benih Putih, Siam Gadis dan Siam Unus adalah 100%. Akurasi pada kelas Bayar Putih hanya sebesar 44% dan selebihnya salah klasifikasi di kelas Benih Kuning (30%) dan Ketan (26%). Akurasi pada kelas Benih Kuning hanya sebesar 35% dan salah klasifikasi di kelas Bayar Putih (45%) dan Ketan (20%). Akurasi pada kelas Karan Dukuh sebesar 63% dan salah klasifikasi di kelas Bayar Putih (26%) dan Benih Kuning (8%). Sedangkan akurasi paling rendah didapat pada kelas Ketan yaitu hanya 10%.

Sama dengan hasil akurasi algoritma menggunakan k = 3 dan k = 5, akurasi menggunakan k = 7 mencapai 100% untuk kelas Bayar Papuyu, Benih Putih, Siam Gadis dan Siam Unus. Akurasi paling rendah di dapat pada kelas Ketan dengan akurasi sebesar 25%.

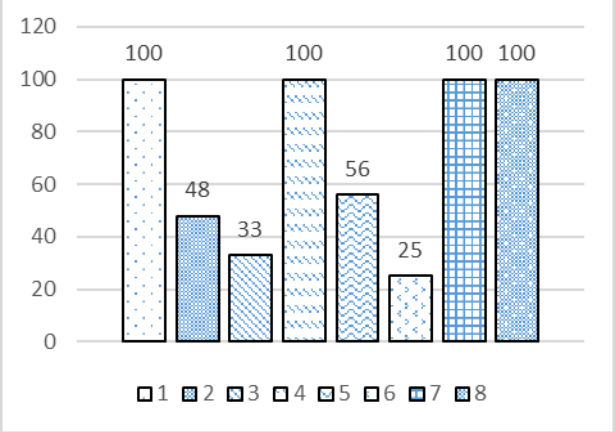
Gambar 4, 5 dan 6 menunjukkan akurasi algoritma KNN pada Data Sintetis tanpa Seleksi Fitur. Akurasi yang di dapat pada kelas Bayar Papuyu, Benih Putih, Siam Gadis dan Siam Unus untuk semua variasi nilai k sebesar 100%. Akurasi pada kelas Bayar Putih paling tinggi didapat pada nilai k=3 dan yang terendah pada nilai k=5. Akurasi pada kelas Benih Kuning paling tinggi di dapat pada nilai k=5 dan yang terendah pada nilai k=3. Akurasi pada kelas Karan Dukuh paling tinggi di dapat pada nilai k=3 dan terendah pada nilai k=7. Akurasi pada kelas Ketan paling rendah di dapat pada nilai k=5. Jika dilihat rata-rata akurasi untuk semua kelas pada masing-masing nilai k maka didapatkan akurasi paling besar pada nilai k=3 dengan akurasi sebesar 70,88%. Hal ini menunjukkan bahwa nilai k=3 merupakan yang terbaik untuk data sintetis tanpa seleksi fitur. Kode 1-9 merepresentasikan kelas padi berturut-turut yaitu Bayar Papuyu, Bayar Putih, Benih Kuning, Benih Putih, Ketan, Siam Gadis, Siam Unus dan Karan Dukuh.



Gambar 4. Pengujian KNN dengan nilai k=3



Gambar 5. Pengujian KNN dengan nilai k=5

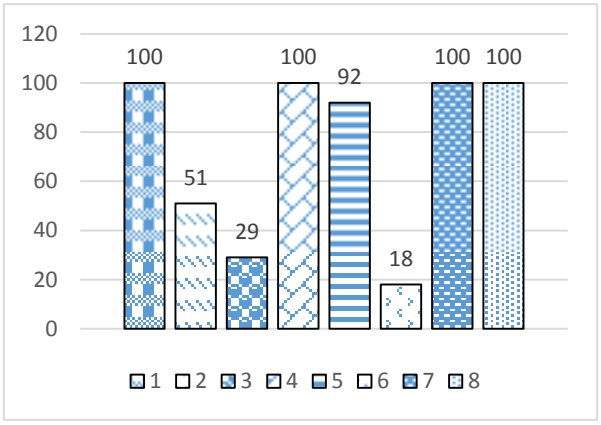


Gambar 6. Pengujian KNN dengan nilai k=7

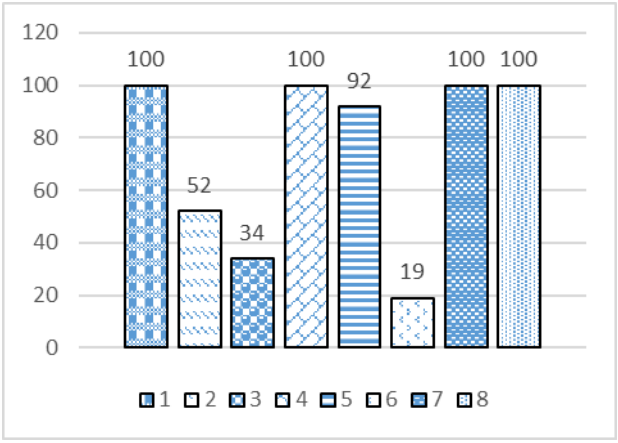
Pengujian KNN dengan Seleksi Fitur

Akurasi yang didapatkan algoritma KNN dengan k=7 mencapai lebih dari 90% pada beberapa kelas dengan rincian kelas Bayar Papuyu (100%), Benih Putih (100%), Karan Dukuh (92%), Siam Gadis (100%) dan Siam Unus (100%). Sedangkan akurasi pada kelas lain berturut-turut kelas Bayar Putih (48%), Benih Kuning (41%), dan Ketan (29%).

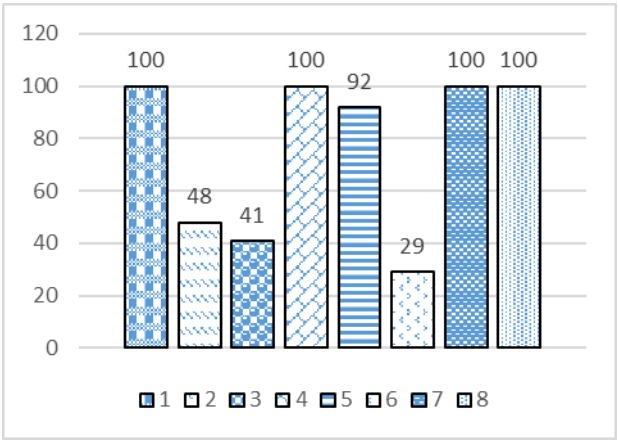
Gambar 7, 8 dan 9 menunjukkan perbandingan akurasi algoritma KNN pada Data Sintetis dengan Seleksi Fitur. Akurasi yang di dapat pada kelas Bayar Papuyu, Benih Putih, Siam Gadis dan Siam Unus untuk semua variasi nilai k sebesar 100%. Akurasi pada kelas Karan Dukuh sebesar 92% untuk semua nilai k. Akurasi pada kelas Bayar Putih paling tinggi didapat pada nilai k=5 dan yang terendah pada nilai k=7. Akurasi pada kelas Benih Kuning paling tinggi di dapat pada nilai k=7 dan yang terendah pada nilai k=3. Akurasi pada kelas Ketan paling tinggi di dapat pada nilai k=7 dan paling rendah pada nilai k=3. Jika dilihat rata-rata akurasi untuk semua kelas pada masing-masing nilai k maka didapatkan akurasi paling besar pada nilai k=7 dengan akurasi sebesar 76,25%. Hal ini menunjukkan bahwa untuk data sintetis dengan seleksi fitur nilai k=7 merupakan yang terbaik.



Gambar 7. Pengujian KNN dengan nilai k=3



Gambar 8. Pengujian KNN dengan nilai k=5



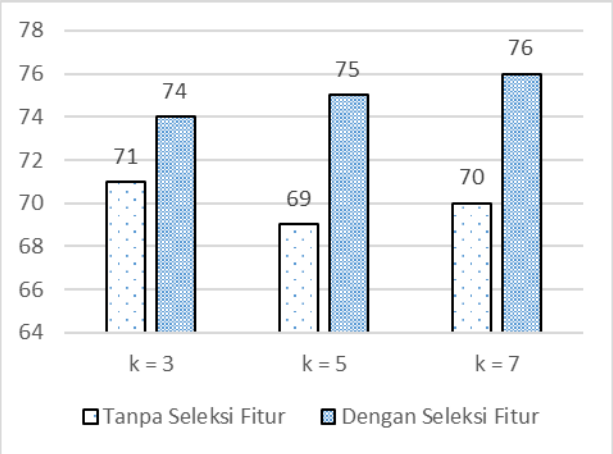
Gambar 9. Pengujian KNN dengan nilai k=7

Akurasi yang didapatkan algoritma KNN dengan k=3 mencapai lebih dari 90% pada beberapa kelas dengan rincian kelas Bayar Papuyu (100%), Benih Putih (100%), Karan Dukuh (92%), Siam Gadis (100%) dan Siam Unus (100%). Sedangkan akurasi pada kelas lain berturut-turut kelas Bayar Putih (51%), Benih Kuning (29%), dan Ketan (18%).

Akurasi yang didapatkan algoritma KNN dengan k=5 mencapai lebih dari 90% pada beberapa kelas dengan rincian kelas Bayar Papuyu (100%), Benih Putih (100%), Karan Dukuh (92%), Siam Gadis (100%) dan Siam Unus (100%). Sedangkan akurasi pada kelas lain berturut-turut kelas Bayar Putih (52%), Benih Kuning (34%), dan Ketan (19%).

Perbandingan Akurasi

Gambar 10 menunjukkan perbandingan akurasi algoritma KNN pada data sintetis tanpa seleksi fitur dengan data sintetis dengan seleksi fitur. Pada nilai k=3, k=5 dan k=7 akurasi algoritma KNN setelah menggunakan seleksi fitur mampu meungguli akurasi algoritma KNN tanpa seleksi fitur. Secara keseluruhan, akurasi paling tinggi di dapat pada saat nilai k=7 dan sudah melalui proses seleksi fitur. Hal ini menunjukkan bahwa penggunaan seleksi fitur dapat mempengaruhi hasil akurasi algoritma KNN.



Gambar 10. Grafik perbandingan akurasi algoritma KNN untuk setiap nilai k antara Data Sintetis tanpa Seleksi Fitur dan dengan Seleksi Fitur

4. Kesimpulan

Kesimpulan yang diperoleh dari penelitian ini yaitu data sintetis dapat digunakan untuk menguji akurasi algoritma KNN. Rata-rata akurasi paling tinggi didapatkan pada saat nilai k=3 sebesar 71% untuk Data Sintetis tanpa Seleksi Fitur. Selanjutnya, untuk meningkatkan akurasi ditambahkan proses seleksi fitur dengan hasil 3 fitur terbaik dari dataset yaitu AspectRatio, MajorAxis dan Feret. Akurasi algoritma KNN terhadap Data Sintetis dengan Seleksi Fitur untuk nilai k=3 sebesar 74%, nilai k=5 sebesar 75% dan k=7 sebesar 76%. Jika dibandingkan dengan Data Sintetis tanpa Seleksi Fitur, maka penambahan akurasi algoritma KNN untuk nilai k=3 sebesar 3%, dan untuk nilai k=5 dan k=7 sebesar 6%. Sehingga dapat disimpulkan, bahwa penggunaan proses Seleksi Fitur dapat meningkatkan akurasi algoritma KNN untuk mengklasifikasikan data benih padi rawa Kalimantan Selatan berdasarkan ciri fisiknya.
Selain menggunakan algoritma KNN, kasus klasifikasi juga bisa menggunakan algoritma seperti Jaringan Saraf Tiruan, Random Forest, Decision Tree dan SVM. Sehingga saran dari penelitian ini adalah mencoba membandingkan akurasi dari beberapa metode. Selain itu, juga bisa digunakan Seleksi Fitur lain seperti ANOVA, Information Gain Ratio dan Gini Decrease.

Daftar Pustaka

[1] United States Department of Agriculture, "Milled Rice Domestic Consumption by Country in 1000 MT – Country Rankings," [Online]. Available: http://www.indexmundi.com/agriculture/?commodity=milled-rice&graph=domestic-consumption. [Accessed 25 September 2017].

[2] P. Lestari, "Metode PCR (Polymerase Chain Reaction) Cara Mengidentifikasi Padi Bermutu Rasa Tinggi," *SinarTani*, pp. 13-16, 2011.

[3] P. T. T. Hong, T. T. T. Hai, L. T. Lan, V. T. Hoang, V. Hai and T. T. Nguyen, "Comparative study on vision based rice seed varieties identification," in *2015 Seventh International Conference on Knowledge and Systems Engineering*, 2015.

[4] S. S. Telang and S. Buradkar, "Review Paper on Analysis and Grading of Food, Grains Using Image Processing and SVM," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 2, pp. 050-054, 2015.

[5] H. Kaur and B. Singh, "Classification and Grading Rice Using Multi-Class SVM," *International Journal of Scientific and Research Publications*, vol. 3, no. 4, pp. 1-5, 2013.

[6] R. I. Srinivas, "Feature Subset Selection using Rough Sets for High Dimensional Data," *International Research Journal of Engineering and Technology (IRJET)*, vol. 2, no. 5, pp. 8-12, 2015.

[7] O. Soesanto, A. Yusuf, D. H. Mursyidin and M. S. Pebriadi, "Jaringan Saraf Radial Basis Probabilistic untuk Identifikasi Morfologi Benih Padi Rawa Kalimantan Selatan," *Jurnal Ilmu Komputer Agri-Informatika*, vol. 4, no. 1, pp. 14-21, 2015.

[8] S. Kumbhar and S. Mulla, "Literature Review on Feature Subset Selection Techniques," *International Journal of Application or Innovation in Engineering and Management*, vol. 3, no. 9, pp. 231-233, 2014.

[9] E. R. M. Saleh, E. Noor, T. Djatna and Irzaman, "Seleksi Parameter Dielektrik Penentuan Masa Kadaluwarsa Biskuit (wafer) dengan Pendekatan Regresi Linier, Feature Selection (ReliefF) dan Artificial Neural Network," *Jurnal Teknologi Industri Pertanian*, vol. 23, no. 2, pp. 164-173, 2013.

[10] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RreliefF," *Machine Learning*, vol. 53, pp. 23-69, 2003.

[11] D. L. Olson and D. Delen, "Advanced Data Mining Techniques", *Springer Science & Business Media*, 2008.

[12] Z. Deng, X. Zhu, D. Cheng, M. Zong and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, pp. 143-148, 2016.