Efektifitas Algoritma K-NN dan Random Forest Dalam Mengenali Gender Berdasarkan Suara

Berlin Pratama[a1], I Ketut Gede Suhartana[a2]

[a]Informatics Departement, Udayana University
Badung, Bali, Indonesia
[1]berlinprtm1@gmail.com
[2]ikg.suhartana@unud.ac.id

Abstract

Gender is the grammatical classification of words and other words related to it that broadly relates to the existence of two sexuality or neutrality[1]*. Human beings have the ability to recognize one's gender through hearing and vision. Within the many differences that in each individual, there are some similarities between men and women that can be directly observed. Women that can be observed directly. One of them is their voice, although the voice of each individual human being is different, but there are similarities between the voices of one and another woman and also man's voice. If listen carefully, woman's voice tends to be higher when compared to men's voices. Through these differences, it's possible to identify human voice through computer media by paying attention to comparison of the frequency of the voice, so that it can be classified into male or female voices*[2]*. In computer science this is called sound analysis, but often human sounds differ from the original after processing by computer. In this case, we try to differentiate human voices by gender using the K-Nearest Neighbor and Random Forest algorithms. The K-Nearest Neighbor algorithm has an accuracy of 76%, while Random Forest has an accuracy of 97%.*

Keywords: *Random Forest, K-Nearest Neighbor, Voice Recognition, Female, Male*

1. Pendahuluan

Suara manusia memiliki berbagai bentuk yang berbeda, mereka dapat dilihat dari persepsi fisik manusia tentang suara diantaranya, bentuk, jenis suara, nada, timbre, dan volume. Persepsi fisik ini dapat didengar dengan jelas sewaktu diucapkan oleh pria atau Wanita. Tentunya akan lebih mudah lagi jika seseorang dapat mendengar langsung lawan bicaranya[2]. Semakin berkembangnya teknologi pada era sekarang, menjadikan pengidentifikasian individu lebih mudah berkat bantuan komputer, salah satu bentuk identitas diri adalah biometrik. Suara yang dulunya mudah dikenali oleh manusia melalui pendengaran dan penglihatannya, sekarang akan dibantu oleh komputer[2]. Metode *random forest* bersifat relatif *robust* kepada *outliers* dan noise, hal itu menjadikannya unggul. Ketika membandingkan metode *random forest* dengan metode klasifikasi pohon keputusan lainnya seperti, *ADTree, LADTree, C4.5, CART, Random Tree, REPTree* dan *BFTree* metode *random forest* memiliki tingkat keakuratan sebesar 96,65% dalam hal pengklasifikasian *file audio*[3]. Selain itu, penelitian ini juga menggunakan metode *K-Nearest Neighbor* sebagai faktor pembanding tingkat akurasi di akhir. Metode K-NN memiliki kelebihan yaitu tahan terhadap data latih dengan banyak *noise* dan keefektifan apabila data latihnya berjumlah banyak. Selain itu proses klasifikasi metode K-NN mudah direpresentasikan dibandingkan dengan metode klasifikasi lain[4].

2. Metode Penelitian

2.1 Dataset

Penelitian ini menggunakan data sekunder untuk proses pengklasifikasian. Dataset sejumlah 3.168 *sample* rekaman suara, yang dikumpulkan dari suara laki-laki dan perempuan. Suara dilakukan *pre-processing* menggunakan analisis akustik dalam R menggunakan *seewave* dan *tuner packages.* Frekuensi yang dianalisis berada di antara 0Hz-280Hz. Diperoleh dalam bentuk *spreadsheet* yang berisi data akustik dari hasil ekstraksi *file audio* yang berbentuk .wav menjadi

bentuk .csv. Dataset dibuat oleh Kory Backer yang dimuat dalam situs opensource Kaggle pada tautan:
https://www.kaggle.com/datasets/primaryobjecta/voicegender.

2.2 Properti Akustik

Tabel 1. Properti Akustik Suara

No.	*Properti Akustik*	
1.	*Q75*	*third quantile (in kHz)*
2.	*IQR*	*interquantile range (in kHz)*
3.	*Skew*	*skewness*
3.	*Kurt*	*kurtosis*
4.	*sp.ent*	*spectral entropy*
5.	*sfm*	*spectral flatness*
6.	*mode*	*mode frequency*
7.	*centroid*	*frequency centroid*
8.	*peakf*	*peak frequency (frequency with highest energy)*
9.	*meanfun*	*average of fundamental frequency measured across acoustic signal*
10.	*minfun*	*minimum fundamental frequency measured across acoustic signal*
11.	*maxfun*	*maximum fundamental frequency measured across acoustic signal*
12.	*meandom*	*average of dominant frequency measured across acoustic signal*
13.	*mindom*	*minimum of dominant frequency measured across acoustic signal*
14.	*maxdom*	*maximum of dominant frequency measured across acoustic signal*
15.	*dfrange*	*range of dominant frequency measured across acoustic signal*
16.	*modindx*	*modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range*
17.	*label*	*male or female*
18.	*meanfreq*	*mean frequency (in kHz)*
19.	*sd*	*standard deviation of frequency*
20.	*median*	*median frequency (in kHz)*
21.	Q25	*First quantile*

2.3 *K-Nearest Neighbor (K-NN)*

K-Nearest Neighbor (K-NN) merupakan model klasifikasi yang menggunakan kelas terbanyak dengan jarak terdekat dalam grup data yang telah di-*training* untuk menentukan kategori kelas. Algoritma *K-Nearest Neighbor* (K-NN) menjalankan perintah klasifikasi berdasarkan pada jarak terdekat dari sampel uji ke sampel terlatih untuk menetapkan K-NNnya[6]. Sebagian besar dari K-NN diambil untuk menjadi prediksi dari sampel ujinya Jarak objek terdekat biasanya dihitung berdasarkan jarak *Eucledian*. Berikut merupakan metode penghitungan K-NN:

$$\text{di}= \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})2} \qquad (1)$$

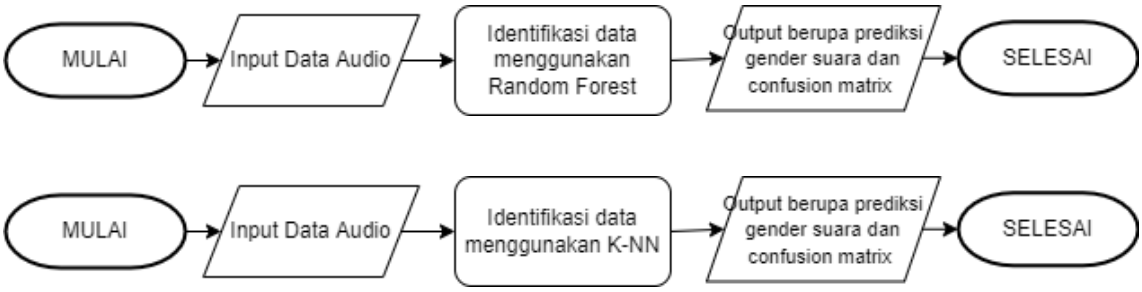
Keterangan:
x = Sampel dari data
x_2 = Data yang akan diujikan
I = Parameter dari data
d = Selisih atau Jarak

p = Luas dari data
Langkah-langkah dari teknik K-Nearest Neighbor yaitu mulai dari input data terlatih, kemudian label data terlatih, nilai k, dan juga data pengujian

2.4 *Random Forest*

Random Forest Algorithm merupakan suatu algoritma pembelajaran mesin yang berguna dalam klasifikasi. *Random forest* mempunyai bagian yaitu pohon keputusan, yang berfungsi untuk membedakan antara data satu dengan yang lainnya[5]. *Random Forest Algorithm* dapat juga dianggap sebagai bentuk perkembangan dari metode klasifikasi *Decision Tree* yang didasarkan dari pemilihan atribut random disetiap node untuk membentuk klasifikasi. Proses klasifikasinya berdasar pada jumlah suara terbanyak dari *decision tree* yang didapat. Berdasarkan teknis yang diterapkan oleh *Random Forest*, menjadikan efisiensi hasil prediksi. Beberapa kelebihan algoritma *Random Forest,* antara lain:
a. Memberi solusi untuk *overfitting*
b. Tingkat sensitivitas terhadap *data outlier* rendah
c. Memiliki parameter yang fleksibel (dapat diubah).
Random forest memiliki karakteristik yang dapat meminimumkan korelasi yang dapat menurunkan hasil kesalahan prediksi random forest. Pada random forest pemilihan pemilah hanya melibatkan beberapa variabel prediktor yang terambil secara acak. Algoritma random forest dijelaskan sebagai berikut[7].
a. Mengambil n data sampel dari dataset awal dengan menggunakan teknik *resampling bootstrap* dengan pengambilan.
b. Menata pohon klasifikasi dari setiap dataset hasil *resampling bootstrap*, dengan penentuan pemilah terbaik didasarkan pada variabel prediktor yang diambil secara acak. Jumlah variabel yang diambil secara acak dapat ditentukan melalui perhitungan $log_2 (Z + 1)$ dimana z adalah banyaknya variabel prediktor atau $\sqrt{Z}$[8].

2.5 Alur Kerja Sistem



Gambar 1. Alur Kerja Sistem

Mengacu pada Gambar 1. Dijelaskan bahwa alur kerja sistem untuk pengujian ini diawali dari menginput data audio berupa *file* berformat .csv yang berisi properti akustik lalu kemudian diidentifikasi menggunakan algoritma K-NN dan *Random Forest* secara terpisah, kemudian akan menghasilkan output berupa prediksi gender suara.

2.6 Evaluasi

Dalam evaluasi, menggunakan bantuan *confusion matrix* untuk menghitung *precision, recall, f1-score,* dan akurasi seperti yang tertera pada Tabel 2*. Confusion Matrix* menampilkan performa klasifikasi dari sebuah *classifier* yang sehubungan dengan data uji.

Tabel 2. *Confusion Matrix*

	Male	*Female*
Male	TP	FP
Female	FN	TN

Keterangan :

TP = *True Positive* (total prediksi benar dari data positif)

FN = *False Negative* (total prediksi negatif tetapi data positif)

TN = *True Negative* (total prediksi benar dari data negatif)

FP = *False Positive* (total prediksi positif tetapi data negatif)

Adapun rumus untuk menghitung *precision, recall, F1-Score,* dan akurasi adalah sebagai berikut:

$$Precision = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (2)$$

3. Hasil dan Pembahasan

Data sejumlah 3.168 kemudian dibagi menjadi dua jenis dengan presentase 80% data latih (*training*) dan 20% data uji (*testing*). Kemudian, lanjut pada dua metode atau algoritma dilakukan proses klasifikasi, didapatkan hasil *precision, recall, f1-score* dan akurasi seperti yang tertera pada Tabel 2.

Tabel 2. Hasil *Precision, Recall, F1-Score K-Nearest Neighbor*

	Precision	*Recall*	*F1-Score*
Female	0.75	0.80	0.77
Male	0.82	0.78	0.80
Accuracy			0.79

Pengimplementasian algoritma K-Nearest Neighbor menghasilkan tingkat akurasi sebesar 78.54%. Hasil *precision* ada di angka 0.75 untuk kelas wanita dan 0.82 untuk kelas pria. Hasil recall ada di angka 0.80 untuk wanita dan 0.78 untuk pria. Hasil F1-Score adalah 0.77 untuk wanita, 0.80 untuk pria, dan 0.79 untuk akurasi. Data hasil klasifikasi juga dimasukkan ke dalam tabel *confusion matrix* yang terdapat pada Tabel 3.

Tabel 3. Confusion Matrix K-Nearest Neighbor

Matrix	*female*	*male*	*all*
Female	231	59	290
Male	77	267	344
all	308	326	634

Tabel 4. Hasil *Precision, Recall, F1-Score Random Forest*

	Precision	*Recall*	*F1-Score*
Female	0.99	0.97	0.98
Male	0.97	0.99	0.98
Accuracy			0.85

Mengacu pada Tabel 4. Pengimplementasian algoritma *Random Forest* menghasilkan tingkat akurasi tertinggi, yaitu sebesar 97.94%. Hasil *precision* ada di angka 0.99 untuk kelas wanita dan 0.97 untuk kelas pria. Hasil *recall* ada di angka 0.97 untuk wanita dan 0.99 untuk pria. Hasil *F1-Score* adalah 0.98 untuk wanita, 0.898 untuk pria, dan 0.85 untuk akurasi. Data hasil klasifikasi juga dimasukkan ke dalam tabel *confusion matrix* seperti pada Tabel 5.

Tabel 5. *Confusion Matrix Random Forest*

Matrix	*Female*	*Male*	*All*
Female	305	10	315
Male	3	316	319
all	308	326	634

4. Kesimpulan

Hasil dari pengujian menampilkan bahwa metode *Random Forest* menghasilkan performa yang lebih baik dibandingkan dengan *K-Nearest Neighbor.* Metode *Random Forest* menghasilkan tingkat akurasi sejumlah 97.94% dengan rata-rata *precision, recall,* dan *f1-score* sebesar 98% dan tingkat akurasi *f1-score* sebesar 85%, dibandingkan dengan *K-Nearest Neighbor* yang memiliki tingkat akurasi sejumlah 78.54% dengan rata-rata *precision, recall,* dan *f1-score* sebesar 78%. Hasil tersebut dapat menyatakan bahwa *Random Forest* dapat membedakan suara berdasarkan gender lebih baik dibandingkan *K-Nearest Neighbor.*

References

[1] C. Kurniawan, H. Irsyad, “Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Untuk Klasifikasi Gender Berdasarkan Mata” *Jurnal Algoritme*, vol. 2, no.2, pp. 82-91, 2022.
[2] I.S. Pratama, F. I. Kurniadiname, "Klasifikasi Jenis Kelamin Berdasarkan Pitch Suara Menggunakan Metode Pitch Detection Algorithm" *Jurnal Sistem Komputer dan Kecerdasan Buatan,* vol. 2, no. 1, p. 1-4, 2018.
[3] A. Sarofi, Irhamah, A. Mukarromah, “Identifikasi Genre Musik dengan Menggunakan Metode Random Forest” *Jurnal Sains dan Seni ITS*, vol. 9, no. 1.
[4] Mursyidah, Jamilah, and Zayya, “Pengenalan Karakter Suara Laki-Laki Aceh Menggunakan Metode FFT (Fast Fourier Transform)” J. Infomedia, vol. 2, no. 1, pp. 20–24, 2017.
[5] S. V. Thambi, K. T. Sreekumar, C. Santhosh Kumar, and P. C. Reghu Raj, “Random Forest Algorithm for Improving the Performance of Speech/non-speech Detection,” 2014 1st Int. Conf. Comput. Syst. Commun. ICCSC 2014, no. December, pp. 28–32, 2003, doi: 10.1109/COMPSC.2014.7032615.
[6] R. Yessivirna, Marji, and D. E. Ratnawati, “Klasifikasi Suara Berdasarkan Gender (Jenis Kelamin) Dengan Metode KNearest Neighbor (KNN)” *Jurnal Ilmu Komputer*., vol. 1, pp.1–9, 2011.
[7] S. Shankar, Raghaveni, J. Rudraraju, P., and V. Sravya. (2020). “Classification of gender by voice recognition using machine learning algorithms”. *Journal of Critical Reviews*, *7*(9), 1217–1229. https://doi.org/10.31838/jcr.07.09.222.
[8] M. Azhar, H. F. Pardede, “Klasifikasi Dialek Pengujar Bahasa Inggris Menggunakan Random Forest” *Jurnal Media Informatika Budidarma,* vol. 5, no. 2, pp. 439-446, 2021.