Indonesian Journal of Electronics and Instrumentation Systems (IJEIS)
Vol.x, No.x, April xxxx, pp. 1~5
ISSN (print): 2088-3714, ISSN (online): 2460-7681
DOI: 10.22146/ijeis.xxxx

Analisis Fitur pada Klasifikasi Kualitas Teh

Abstrak

Indonesia merupakan salah satu produsen terbesar teh hitam dunia, maka kualitas dari minuman teh perlu diperhatikan. PT Pagilaran, salah satu perusahaan di Indonesia dengan salah satu komoditas unggulannya berupa teh, khususnya teh hitam melakukan penilaian 10 tingkatan kualitas teh dengan penilaian manual oleh tenaga ahli yang rentan dengan inkonsistensi. Electronic nose adalah salah satu teknologi yang berkembang pesat untuk digunakan penilaian kualitas teh. PT Pagilaran dan Fakultas Teknologi Pertanian Universitas Gadjah Mada melakukan pengembangan solusi yang lebih objektif dan konsisten dalam penilaian kualitas teh dengan electronic nose. Data aroma teh akan diekstrasi dari electronic nose kemudian dianalisis untuk menentukan fitur-fitur yang representatif dalam membedakan 10 kelas kualitas teh. Beberapa penelitian sebelumnya telah menunjukkan potensi penggunaan fitur dari electronic nose untuk mengklasifikasikan kualitas teh, namun sebagian besar hanya mencapai klasifikasi tujuh kelas teh. Penelitian ini melakukan analisis kombinasi fitur yang tepat untuk mengembangkan model klasifikasi kualitas teh menjadi 10 kelas menggunakan data electronic nose. Metode yang digunakan adalah seleksi fitur Recursive Feature Elimination dan Principal Component Analysis serta classifier Support Vecor Machine. Hasil dari penelitian menunjukkan bahwa model klasifikasi dengan nilai akurasi terbaik mencapai 89,85% menggunakan 18 fitur pilihan dari electronic nose.
Kata kunci*—Electronic nose, kualitas teh, fitur, klasifikasi, Support Vector Machine*

Abstract

Indonesia is one of the world's largest producers of black tea, making quality control crucial. PT Pagilaran, a prominent Indonesian tea producer specializing in black tea, currently assesses 10 quality grades through manual evaluation by experts, which is prone to inconsistency.Electronic nose technology has emerged as a promising solution for objective tea quality assessment. PT Pagilaran, in collaboration with the Faculty of Agricultural Technology at Gadjah Mada University, is developing a more consistent quality evaluation system using e-nose. Tea aroma data from the e-nose will be extracted and analyzed to identify representative features for distinguishing the 10 quality grades. While prior studies have demonstrated the potential of e-nose features for tea classification, most were limited to 7 quality grades.This research focuses on optimizing feature selection to develop a 10-grade classification model using e-nose data. The methodology combines Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and Support Vector Machine (SVM) classification. The best-performing model achieved 89.85% accuracy using 18 selected features, demonstrating e-nose's effectiveness as an automated and reliable tea quality assessment tool.
Keywords—*Electronic nose, tea quality, fitur, classification, Support Vector Machine*

1. PENDAHULUAN

Teh adalah salah satu minuman yang paling banyak dikonsumsi di dunia, setelah air [1]. Di Indonesia, teh merupakan salah satu komoditas perkebunan yang memegang peranan cukup penting dalam perekonomian. Indonesia merupakan salah satu produsen terbesar teh hitam dunia, maka kualitas dari minuman teh perlu diperhatikan[2]. Penilaian kualitas teh adalah salah satu poin yang penting karena mempengaruhi harga produk teh dan penerimaan pelanggan[3].

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

Di Indonesia terdapat lembaga bernama Badan Standardisasi Nasional (BSN) yang bertanggung jawab dalam bidang standardisasi yang telah memiliki Standar Nasional Indonesia yang berkaitan dengan produk teh[4]. Hingga tahun 2019 terdapat 9 standar yang berkaitan dengan produk teh. Salah satunya standar mutu kualitas teh hitam[5].

PT Pagilaran adalah salah satu perusahaan di Indonesia yang salah satu komoditas unggulannya berupa teh, khususnya produksi teh hitam, PT Pagilaran menetapkan 10 tingkatan mutu kualitas teh yaitu PF, PF II, PF III, FI, F II, Dust, Dust II, BOPF, BOP, dan BOHEA. Saat ini, penilaian kualitas teh hitam di PT Pagilaran mengandalkan metode tradisional, yaitu inspeksi manual oleh tenaga ahli. Namun, metode ini memiliki beberapa keterbatasan yaitu potensi ketidakkonsistenan akibat dari faktor fisiologis, psikologis, lingkungan, dan kompleksitas senyawa volatil aroma dalam tiap kualitas teh. Dalam penilaian kualitas teh, salah satu teknologi yang mulai banyak digunakan adalah *electronic nose*. Teknologi *electronic nose* adalah instrumen cerdas yang dirancang untuk mendeteksi dan membedakan bau-bau kompleks menggunakan rangkaian sensor. Instrumen ini memberikan teknik pengambilan sampel yang cepat, sederhana, dan tidak invasif untuk mendeteksi dan mengidentifikasi berbagai senyawa volatil [6]. Dalam beberapa dekade terakhir teknologi *electronic nose* berkembang pesat karena aplikasinya luas, biayanya yang relatif murah, aman, dan kecepatan hasil deteksi yang diperoleh[7].

Beberapa penelitian juga telah dilakukan terkait penilaian kualitas teh menggunakan *electronic nose* menggunakan fitur-fitur tertentu. Terdapat penelitian dengan deep learning dan multi sensor fusion yang salah satunya adalah *electronic nose* dalam mengatasi masalah kurangnya teknologi diskriminasi cepat untuk evaluasi komprehensif kualitas teh. Nilai sinyal proses yang diperhalus dari kurva respon sensor *electronic nose* sebagai nilai fitur. Penelitian ini melakukan klasifikasi kualitas tujuh *grade* dari sampel teh hitam *Yunnan Congou* dengan hasil klasifikasi tertinggi adalah model VCPA-IRIV dan SVM dengan skor 97,42% dan model CNN meraih skor tertinggi yaitu 99,14%[8]. Penelitian pengembangan sistem *electronic nose* untuk mengklasifikasikan teh hitam menjadi tiga *grade* dengan metode ekstraksi fitur PCA dan LDA. Metode *line fitting* untuk mengekstraksi fitur dari kurva respon sensor lalu direduksi dimensi fitur dengan PCA dan LDA. Hasil klasifikasi model SVM mencapai tingkat akurasi pelatihan, validasi, dan pengujian terbaik sebesar 99,50%, 95,30%, dan 96,50%[9].

Penelitian lain dalam pengklasifikasian merek teh hijau dengan *electronic nose* dan metode LDA menjadi empat kluster menggunakan fitur berupa kurva *sensing* dan flessing dari sensor *electronic nose* dengan dihitung luasannya menggunakan metode integral trapezium. Nilai total dua fungsi diskriminan pertama sebesar 86,4%[7]. Studi gabungan analisis *Near Infrared Spectroscopy* dan *electronic nose* untuk mengevaluasi kualitas teh hitam menjadi tiga *grade*. Metode *Principal Component Analysis* digunakan untuk mengekstrak *Principal Component* dari *electronic nose*. *Principal Component* yang dioptimalkan dari data NIRS dan *electronic nose* digabung menjadi satu vektor pemodelan. Hasil klasifikasi dengan SVM menunjukkan model gabungan NIRS dan *electronic nose* mencapai akurasi tertinggi 98.13% sedangkan klasifikasi dengan data *electronic nose* saja mencapai akurasi tertinggi 97% [10]. Penelitian pengembangan *electronic nose* portabel yang bisa menilai kualitas teh berdasarkan aroma ampas teh dan menggunakan *Multilayer Perceptron* untuk mengklasifikasikan kualitas aroma teh menjadi 3 kelas. Enam fitur aroma diperoleh untuk setiap pengukuran selama 5 menit dijadikan matriks dengan berukuran 6×600 sebagai data input. Hasil klasifikasi terbaik adalah LP dengan spesifikasi fungsi aktivasi ReLu dengan 3 hidden layer dan 100 hidden node tiap layer dengan akurasi 0.875 standar deviasi 0.0241[4].

Menyadari keterbatasan metode penilaian manual dan penelitian-penelitian tentang penggunaan *electronic nose*, PT Pagilaran berkolaborasi dengan Fakultas Teknologi Pertanian Universitas Gadjah Mada untuk mengembangkan solusi yang lebih objektif dan konsisten dalam penilaian kualitas teh. Kolaborasi ini melibatkan penggunaan teknologi *electronic nose* untuk mengambil data aroma dari setiap tingkatan kualitas teh. Data yang dihasilkan dari *electronic nose* ini kemudian dianalisis lebih lanjut untuk menentukan fitur-fitur yang paling representatif dalam membedakan 10 kelas kualitas teh. Berdasarkan penelitian yang sudah ada, telah

ditunjukkan potensi penggunaan fitur yang diekstraksi dari *electronic nose* untuk mengklasifikasikan kualitas teh seperti nilai sensor, fitur domain waktu sinyal, karakteristik kurva sensing dan flessing *electronic nose*, dan *Principal Component*. Fitur-fitur tersebut bekerja dengan baik secara mandiri maupun dikombinasikan dengan fitur lain untuk melakukan klasifikasi kualitas teh. Namun, sebagian besar penelitian tersebut hanya mengklasifikasikan teh paling banyak menjadi tujuh kelas kualitas teh. Oleh karena itu, diperlukan penelitian lebih lanjut untuk melakukan analisis fitur yang tepat dalam mengembangkan model klasifikasi 10 kualitas teh berdasarkan data aroma ampas teh.

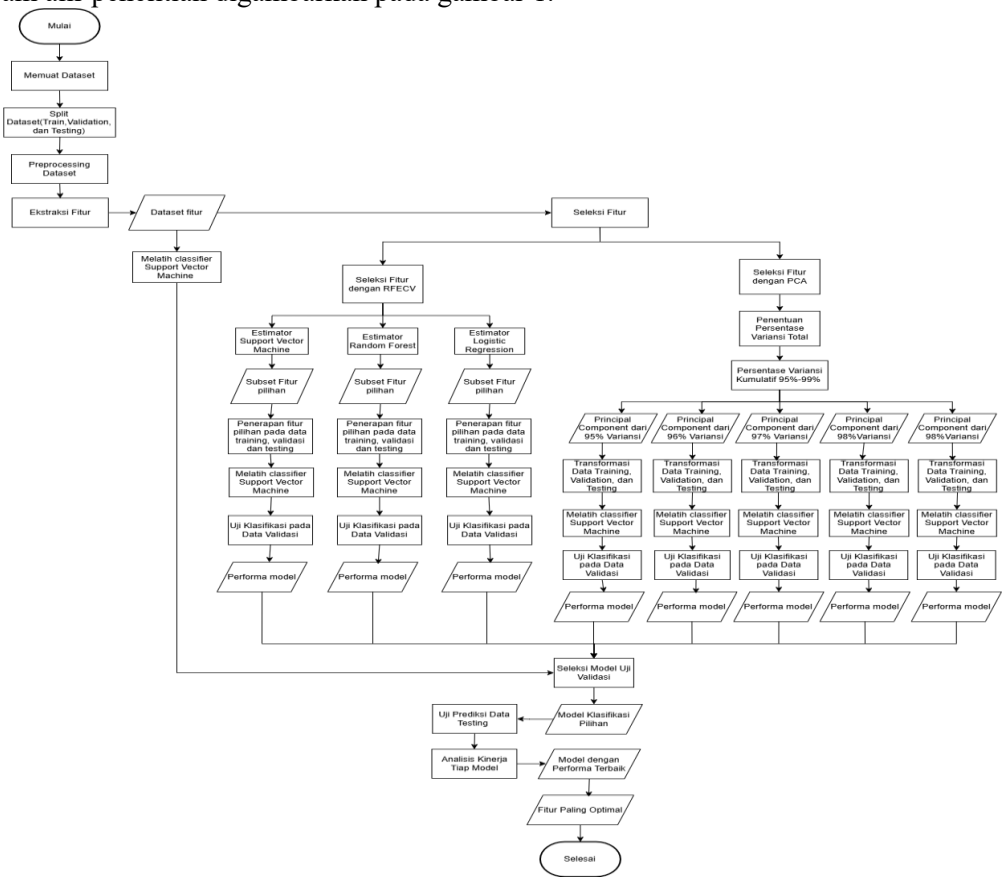
2. METODE PENELITIAN

2.1 Analisis Sistem

Sistem dirancang untuk menganalisis dan menentukan fitur yang tepat untuk mengklasifikasikan 10 kualitas teh berbasis aroma ampas teh berdasarkan performa model klasifikasi terbaik. Data berasal dari enam sensor gas, sensor-sensor ini mendapatkan pola yang mempunyai ciri untuk mempresentasikan aroma yang terdeteksi dari masing-masing sampel aroma teh hitam. Penelitian ini dilakukan melalui tahap akuisisi data, prapemrosesan berupa pembagian dan penanganan *outlier,* kemudian ektraksi fitur, seleksi fitur, dan klasifikasi.

2.2 Perancangan Sistem

Rancangan sistem pada penelitian ini divisualisasikan dengan diagram alir. Diagram alir akan menggambarkan alur atau proses dari sistem pada penelitian ini. Rancangan sistem dari penelitian ini terdiri dari bagian akusisi data, preprocessing data, ekstraksi fitur, seleksi fitur, pengembangan model klasifikasi, uji validasi, pengujian sistem, dan analisis hasil pengujian. Diagram alir penelitian digambarkan pada gambar 1.



Gambar 1 Rancangan Sistem

2.3 Akuisisi Data

Dataset teh hitam yang digunakan pada penelitian ini adalah dataset sekunder yang berupa data aroma ampas teh. Data ini diperoleh dari Departemen Teknologi Pangan dan Hasil Pertanian, Fakultas Teknologi Pertanian, Universitas Gadjah Mada (UGM). Sampel daun teh yang diukur diperoleh dari PT. Pagilaran. Dataset terdiri dari 10 kelas kualitas aroma teh hitam yaitu PF, PF II, PF III, F I, F II, DUST, DUST II, BOP, BOPF, dan BOHEA. Dataset tersebut terdiri dari 60 sampel teh pada tiap kelas yang diukur sebanyak lima kali setiap sampel sehingga pada tiap kelas teh memiliki 300 sampel berformat csv.

2.4 Prapemrosesan Data

Tahap prapemrosesan data diawali dengan pembagian data menjadi tiga komposisi yaitu data *training*, data validasi, dan data *testing*. Data training digunakan untuk melatih dan mengembangkan model klasifikasi. Data *validasi* digunakan untuk melakukan uji klasifikasi pada data yang belum pernah dilihat sebelumnya menggunakan model-model klasifikasi yang telah dikembangkan. Data *testing* untuk melakukan uji prediksi akhir dari model-model klasifikasi terbaik uji validasi. Setelah membagi data menjadi tiga komposisi, selanjutnya adalah melakukan penanganan deteksi *outlier* dengan metode InterQuartile Range(IQR).Deteksi *outlier* menggunakan metode deteksi outlier IQR.

2.4.1 Deteksi Outlier Interquartile Range (IQR)

Langkah deteksi outlier IQR diawali dengan mengurutkan dari nilai terkecil ke nilai terbesar pada nilai tiap kolom sensor tiap kelas teh. Setelah itu, akan dihitung nilai kuartil pertama dan kuartil ketiga nilai tiap sensor pada tiap kelas. IQR dihitung dari nilai selisih antara nilai kuartil ketiga(Q3) dan kuartil pertama(Q1)[11]. Nilai IQR akan dihitung persamaan (1)

$$\text{IQR} = Q_3 - Q_1 \quad (1)$$

Nilai yang dianggap sebagai *outlier* adalah nilai yang berada di luar rentang *threshold* bawah dan *threshold* atas Threshold bawah dihitung dengan cara mengurangkan nilai kuartil pertama dengan hasil perkalian nilai IQR dengan 1,5 seperti persamaan (2) sedangkan *threshold* atas dihitung dengan cara menjumlahkan nilai kuartil ketiga dengan hasil perkalian nilai IQR dengan 1,5 sesuai persamaan (3).

$$\text{Batas bawah} = Q_1 - 1{,}5 * IQR \quad (2)$$

$$\text{Batas atas} = Q_3 + 1{,}5 * IQR \quad (3)$$

Penanganan *outlier* dilakukan dengan menggantinya dengan nilai median data agar lebih representatif. Setelah dinormalisasi, data akan mengikuti distribusi yang lebih normal yang juga berdampak pada peningkatan kualitas data. Setelah proses prapemrosesan, data menjadi memiliki sebaran yang relatif sama serta *outlier* teridentifikasi dan ditangani

2.5 Ekstraksi Fitur

Ekstraksi fitur dilakukan untuk mengekstraksi informasi yang kuat dari respon sensor sebagai fitur atau ciri. Fitur yang diesktraksi adalah fitur statistik. Penelitian ini menggunakan 13 jenis fitur. Berikut ini adalah daftar dan kode penamaan fitur tersebut dalam Tabel 2.

Tabel 2 Kode Penamaan Fitur

No	Fitur	Sensor	Kode Fitur
1.	Mean	MQ 7	MQ 7_Mean
2.	Median	MQ 7	MQ 7_Median
3.	Modus	MQ 7	MQ 7_Modus
4.	Nilai Minimum	MQ 7	MQ 7_Min
5.	Nilai Maksimum	MQ 7	MQ 7_Max
6.	Range	MQ 7	MQ 7_Range
7.	Varians	MQ 7	MQ 7_Varians
8.	Standar Deviasi	MQ 7	MQ 7_StdDev

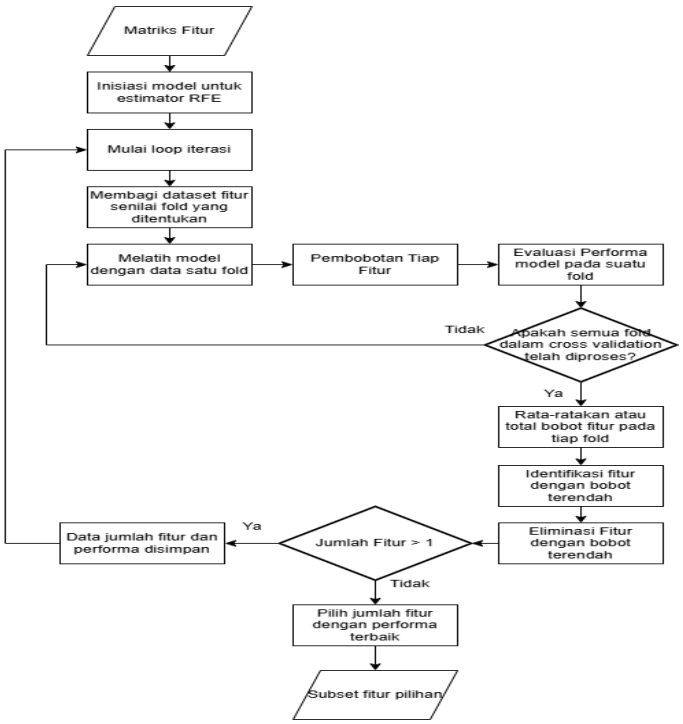
9.	Kuartil_25	MQ 7	MQ 7_Q25
10.	Kuartil_75	MQ7	MQ 7_Q75
11.	Skewness	MQ 7	MQ 7_Skewness
12.	Kurtosis	MQ 7	MQ 7_Kurtosis
13.	Root Mean Square	MQ 7	MQ 7_RMS
….	…….	……..	……
78.	Root Mean Square	TGS 2611	TGS 2611_RMS

2.6 Seleksi Fitur

Proses seleksi fitur akan memilih subset fitur yang relevan dengan dari sekumpulan fitur asli yang lebih besar dengan harapan mendapatkan subset fitur yang lebih sedikit. Pemilihan fitur ini dilakukan berdasarkan kriteria yang telah ditentukan sebelumnya. Pada penelitian ini, seleksi fitur akan dilakukan menggunakan dua metode yaitu *Recursive Feature Elimination with Cross Validation* dan *Principal Component Analysis.*

2.6.1 Recursive Feature Elimination with Cross Validation

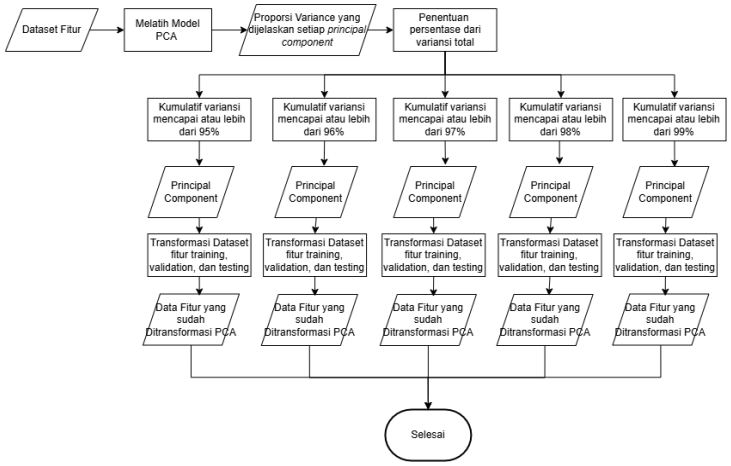
Recursive Feature Elimination with Cross Validation bekerja dengan secara berulang-ulang menghapus fitur yang dianggap kurang penting berdasarkan performa *classifier estimator* sambil menggunakan *cross validation* untuk mendapatkan estimasi bobot fitur berdasarkan performa yang paling baik. Diterapkan tiga model *classifier* sebagai *estimator* yaitu Logistic Regression, Support Vector Machine, dan Random Forest. Pada setiap *estimator* diterapkan dua skenario jumlah pengurangan fitur tiap iterasi yaitu step=1 dan step=3. Cara kerja *Recursive Feature Elimination with Cross Validation* dijelaskan dalam diagram alir pada gambar 2.



Gambar 2 Diagram Alir RFECV

2.6.2 Principal Component Analysis (PCA)

Principal Component Analysis(PCA) bekerja dengan mengidentifikasi arah varians terbesar dalam data dan memproyeksikan data ke ruang dimensi yang lebih rendah sambil mempertahankan sebagian besar informasi. Proses seleksi fitur dengan *Principal Component Analysis* ditunjukkan oleh diagram alir dalam gambar 3.



Gambar 3 Diagram Alir PCA

2.7 Pengembangan Model Klasifikasi

Proses pengembangan model klasifikasi meliputi proses melatih model dengan data *training* untuk mengenali pola dari data yang telah diberi label. Algoritma klasifikasi yang digunakan adalah *Support Vector Machine*. Penelitian ini menggunakan dua skema pendekatan klasifikasi multikelas yaitu pendekatan One vs Rest(OVR) dan pendekatan One vs One(OVO). Demi mengoptimalkan klasifikasi maka dilakukan tuning *hyperparameter* menggunakan metode *grid search*. Penyetelan *hyperparameter* pada *Support Vector Machine* ditunjukkan oleh tabel 3.

Tabel 3 Tuning Hyperparameter Support Vector Machine

Parameter	**Nilai Parameter**
C	0.01
	0.1
	1
	10
	100
Gamma	auto
	scale
	0,01
	0,1
	1
Kernel	rbf
	Linear
	Polynomial
Class_weight	balanced
	None

2.8 Uji Validasi

Tahap uji validasi merupakan tahapan dimana model-model klasifikasi yang telah dikembangkan akan dilakukan uji prediksi pada data validasi. Setelah melakukan uji prediksi data validasi akan dilakukan perhitungan nilai akurasi tiap model klasifikasi tersebut. Dari hasil pengukuran nilai akurasi tiap model tersebut akan dilakukan seleksi model-model klasifikasi yang akan diujikan lebih lanjut pada uji prediksi akhir dengan data *testing.* Seleksi model klasifikasi yang akan diujikan lebih lanjut tersebut berdasarkan nilai ambang batas minimum yang ditentukan dari distribusi nilai akurasi tiap model.

2.9 Pengujian Sistem

Proses pengujian sistem dilakukan dengan menerapkan keseluruhan model klasifikasi yang lolos seleksi uji validasi untuk diujikan lebih lanjut pada prediksi data *testing*. Setelah melakukan uji prediksi data *testing*, maka selanjutnya adalah dilakukan pengukuran performa tiap model klasifikasi menurut metrik akurasi, presisi, recall, dan f-1 score.

2.9.1 Analisis Hasil Pengujian

Dalam analisis hasil pengujian dilakukan penentuan satu model klasifikasi yang memiliki performa paling baik dalam uji prediksi data *testing* berdasarkan *confusion matrix* yang menjelasan metrik akurasi, presisi, recall, dan f-1 score. Fitur yang digunakan model klasifikasi dengan performa paling baik ini dianggap sebagai fitur yang paling tepat untuk digunakan dalam mengklasifikasikan kualitas teh menjadi 10 kelas berdasarkan aroma ampas teh.

3. HASIL DAN PEMBAHASAN

3.1 Akuisisi Data

Pada tahap ini, data yang diakuisisi adalah 3000 data dengan masing-masing 300 sampel tiap kelas teh. Tiap sampel tersebut berformat csv. Dalam tiap berkas csv tersebut berisi 300 baris data yang merepresentasikan hasil pembacaan sensor *electronic nose.*

3.2 Prapemrosesan Data

Tahap pertama dalam prapemrosesan data adalah pembagian data menjadi data *training* 60%, data validasi 20%, dan data *testing 20%.* Detail rinci komposisi data adalah 1800 sampel data *training*, 600 sampel data validasi, dan 600 sampel data *testing*. Setelah dibagi, metode deteksi *outlier* diterapkan pada setiap kolom dalam sampel aroma ampas the lalu mengganti nilai outliernya dengan nilai median dari data tanpa nilai *outlier*. Gambar 6 menunjukkan contoh data sebelum dan sesudah dibersihkan dari *outlier*.

	MQ 7	MQ 9	TGS 822	TGS 2600	TGS 2602	TGS 2611
561	3862	24	0	3867	3706	7629
562	4042	176	0	4071	3464	8034
563	3985	0	0	3603	3893	7811
564	3740	0	0	3582	3777	7622
565	4007	42	0	4034	3697	7449
566	4259	157	0	3725	2955	8142
567	4147	130	0	4511	3706	7804
568	4800	266	0	3440	3910	7785
569	3766	0	0	3812	3788	7597
570	3930	4	0	3716	3737	7436
571	4252	282	0	3705	3356	7713

	MQ 7	MQ 9	TGS 822	TGS 2600	TGS 2602	TGS 2611
561	3862.0	12955.0	0.0	3867.0	3706.0	7629.0
562	4042.0	12955.0	0.0	4071.0	3464.0	8034.0
563	3985.0	0.0	0.0	3603.0	3893.0	7811.0
564	3740.0	0.0	0.0	3582.0	3777.0	7622.0
565	4007.0	12955.0	0.0	4034.0	3697.0	7449.0
566	4259.0	12955.0	0.0	3725.0	2955.0	8142.0
567	4147.0	12955.0	0.0	4511.0	3706.0	7804.0
568	4800.0	12955.0	0.0	3440.0	3910.0	7785.0
569	3766.0	0.0	0.0	3812.0	3788.0	7597.0
570	3930.0	12955.0	0.0	3716.0	3737.0	7436.0
571	4252.0	12955.0	0.0	3705.0	3356.0	7713.0

Gambar 6 Contoh Data Sebelum dan Sesudah Ditangani Outliernya

Pada matriks data yang terdapat di atas itu menunjukkan contoh data sebelum dibersihkan *outliernya* dan matriks data yang ada di bawah adalah data yang sudah ditangani outliernya. Terdapat beberapa nilai data pada kolom sensor MQ 9 tersebut yang terlihat tidak terbaca polanya dan rentang datanya tidak konsisten sehingga beberapa nilainya diganti dengan 12.955 yang menunjukkan bahwa 12.955 merupakan nilai median dari kumpulan data bukan *outlier*. Setelah dilakukan deteksi dan penanganan *outlier* maka distribusi data menjadi lebih stabil dengan pola yang lebih konsisten dan representatif terhadap karakteristik asli data. Nilai ekstrem yang terindikasi sebagai nilai *outlier* digantikan dengan nilai median tanpa nilai *outlier*.

3.3 Ekstraksi Fitur

Hasil dari ekstraksi fitur menghasilkan data fitur tiap komposisi. Data fitur-fitur tersebut diekstraksi dari data tiap sampel pada tiap kelas yang terdapat di tiap komposisi data. Fungsi ekstraksi fitur akan bekerja pada setiap sampel secara berurutan dan ekstraksi fitur tiap sampel tersebut disusun menjadi satu file csv yang dianggap sebagai fitur suatu kelas. Di dalam tiap data fitur kelas tersebut terdapat sebuah dataframe dengan 79 kolom dan baris sejumlah sampel yang diekstraksi. Gambar 7 di bawah ini adalah contoh gambaran data fitur tiap kelas.

File	MQ 7_Mean	MQ 7_Median	MQ 7_Modus	MQ 7_Min	MQ 7_Max	MQ 7_Range	MQ 7_Varians
FI_part_1.csv	7237.878172745893	7232.0	7208.0	6065.0	9414.451823767811	3349.451823767811	87863.34143350595
FI_part_10.csv	8997.906666666666	8992.0	8993.0	7540.0	9920.0	2380.0	69274.39259754738
FI_part_11.csv	9831.08	9823.0	10052.0	8194.0	11958.0	3764.0	102624.9233444816
FI_part_12.csv	9395.596666666666	9402.0	9664.0	8284.0	10272.0	1988.0	59388.95717948717
FI_part_13.csv	8097.3533333333335	8107.0	8191.0	5368.0	9115.0	3747.0	96958.61052396879
FI_part_14.csv	9030.926666666666	9029.0	8865.0	7830.0	10312.0	2482.0	79836.00798216276
FI_part_15.csv	7647.606666666667	7656.0	7124.0	6044.0	9307.0	3263.0	119023.87152731327
FI_part_16.csv	8130.21	8133.5	8191.0	6848.0	9717.0	2869.0	74623.73836120403
FI_part_17.csv	7922.066666666667	7928.5	7844.0	6804.0	9220.0	2416.0	68851.59420289854
FI_part_18.csv	7867.11	7870.5	7890.0	6808.0	9331.0	2523.0	83384.41260869564
FI_part_19.csv	10393.906666666666	10375.0	10352.0	8464.0	13087.0	4623.0	139685.94443701228
FI_part_2.csv	7968.036666666667	7960.0	7801.0	5664.0	9106.0	3442.0	94123.1324303233
FI_part_20.csv	10096.26	10103.5	10050.0	9262.0	12396.0	3134.0	83051.1562541806
FI_part_21.csv	8462.5	8449.0	8191.0	6755.0	9435.0	2680.0	76123.97658862876
FI_part_22.csv	8659.473333333333	8638.5	8492.0	7082.0	10062.0	2980.0	88292.08289855073

Gambar 7 Cuplikan Data Fitur Tiap Kelas

Nama kolom fitur tersebut diberi nama dengan format "Nama Sensor_Jenis Fitur" dengan urutan sensor MQ 7, MQ 9, TGS 822, TGS 2600, TGS 2602, dan TGS 2611. Pada format jenis fiturnya akan berurutan dari mean, median, modus, nilai min, nilai max, range, varians, standar deviasi, kuartil ke-1(Q25), kuartil ke-3(Q75), skewness, kurtosis, dan root mean square. Fitur-fitur statistik deskriptif yang diekstraksi dari tiap sensor cukup merepresentasikan karakteristik distribusi nilai pengukuran dari setiap sensor untuk setiap kelas teh.

3.4 Seleksi Fitur

3.4.1 Seleksi Fitur dengan Principal Component Analysis

Teknik *Principal Component Analysis* akan mereduksi dimensi linear yang mentransformasi data ke sistem koordinat baru dimana varians terbesar pada data terletak pada *principal component* pertama, *principal component* kedua, dan seterusnya. Struktur varians dalam data akan dianalisis lalu direpresentasikan oleh matriks berisi proporsi varians dari setiap *principal component*(PC) berjumlah 78 PC. Seleksi fitur dengan PCA dilakukan dengan pemilihan jumlah *PC* yang ditentukan jumlahnya dengan beberapa skenario *threshold* varians yang mempertahankan persentase varians kumulatif yang ditentukan seperti pada tabel 4. Setelah mendapatkan jumlah PC yang sesuai, maka selanjutnya dilakukan transformasi data asli untuk diproyeksikan ke ruang dimensi lebih rendah menggunakan PC yang dipilih

Tabel 4 Pemilihan Jumlah PC Berdasarkan Persentase Varians Kumulatif

Persentase Varians Kumulatif	Jumlah PC yang Dipertahankan
95%	20
96%	21
97%	23
98%	25
99%	29

3.4.2 Seleksi Fitur dengan Recursive Feature Elimination with Cross Validation (RFECV)

Metode *Recursive Feature Elimination with Cross Validation* akan bekerja secara berulang-ulang menghapus fitur yang dianggap kurang penting berdasarkan performa *estimator* sambil menggunakan *cross validation* untuk mendapatkan estimasi bobot fitur berdasarkan performa yang paling baik. Hasil seleksi fitur berdasarkan jenis estimatornya seperti ditunjukkan tabel 5.

Tabel 5 Seleksi Fitur Recursive Feature Elimination with Cross Validation

Estimator	Akurasi Estimator	Step	Jumlah Fitur Pilihan

	Terbaik (%)		
Support Vector Machine	79,89	1	48 fitur
Support Vector Machine	79,89	3	48 fitur
Logistic Regression	78,39	1	33 fitur
Logistic Regression	78,39	3	27 fitur
Random Forest	85,06	1	18 fitur
Random Forest	85,06	3	18 fitur

3.5 Pengembangan Model Klasifikasi

Pada tahap ini akan disajikan hasil *tuning hyperparameter* model klasifikasi SVM terbaik pada tiap skenario dengan *GridSearch Cross Validation* pada model-model. Hasil tuning *hyperparameter* model *Support Vector Machine* ditunjukkan oleh tabel 6.

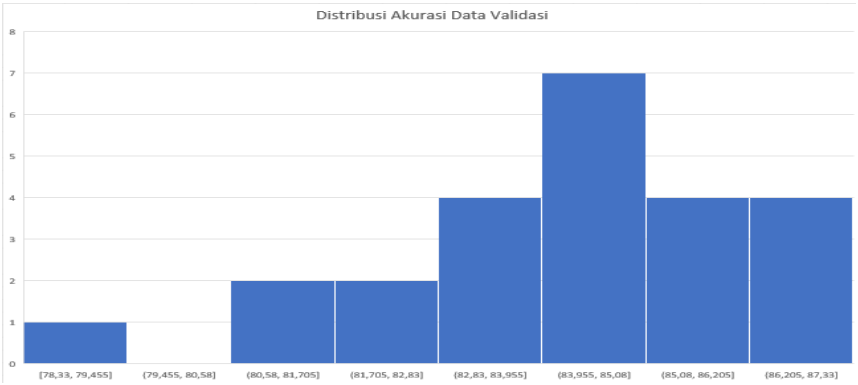
Tabel 6 Hasil Tuning Parameter

Fitur Data	Kombinasi Parameter Terbaik	Akurasi Terbaik (%)
20 Principal Component Threshold PCA 95%	C=10,class_weight= None, gamma=scale, kernel=rbf	79,33
21 Principal Component Threshold PCA 96%	C=10,gamma=0.01, kernel='rbf',class_weight='balanced'	79,56
23 Principal Component Threshold PCA 97%	C=10, gamma=0.01, kernel='rbf',class_weight='balanced'	80,06
25 Principal Component Threshold PCA 98%	C=10, gamma='scale', kernel='rbf',class_weight='balanced'	80,22
29 Principal Component Threshold PCA 99%	C=10, gamma=0.01, kernel='rbf',class_weight='balanced'	79,83
48 fitur RFECV estimator SVM step 1	C=0.1, gamma='auto', kernel='linear', class_weight='balanced'	80,89
48 fitur RFECV estimator SVM step 3	C=10, gamma=0.01, kernel='rbf', class_weight=None	80,89
33 fitur RFECV estimator Logistic Regression step 1	C=10, gamma=0.01, kernel='rbf', class_weight=None	82,17
27 fitur RFECV estimator Logistic Regression step 3	C=10, gamma=0.01, kernel='rbf', class_weight='balanced'	82,11
18 fitur RFECV estimator Random Forest step 1	C=1, gamma=0.1, kernel='rbf', class_weight='balanced'	82
18 fitur RFECV estimator Random Forest step 3	C=10, gamma='scale', kernel='rbf', class_weight=None	81,78
78 fitur tanpa seleksi fitur	C=10, gamma=0.01, kernel='rbf', class_weight=None	79,99

Evaluasi kinerja model dilakukan sebanyak sebanyak tiga kali *fold* perulangan dalam proses *grid search* dari tiap parameter. Kombinasi nilai parameter terbaik akan dijadikan parameter untuk pengembangan model klasifikasi SVM. Model klasifikasi SVM yang sudah dikembangkan berdasarkan parameter terbaik pada masing-masing skenario penanganan fitur akan disimpan pada sebuah variabel kemudian diterapkan pada *classifier* SVM *multiclass* OVO dan OVR. Model klasifikasi OVO dan OVR akan menjadikan variabel sebelumnya sebagai *classifier* dasar untuk pemisahan kelas. Satu variabel *classifier* dasar akan digunakan pada kedua *classifier* OVO dan OVR sehingga menjadi 24 model *classifier*. Seluruh 24 model *classifier* tersebut akan dilatih menggunakan data *training*.

3.6 Uji Validasi

Pada tahap ini akan dilakukan uji prediksi pada data validasi. Hasil distribusi nilai akurasi model klasifikasi pada uji prediksi data validasi ditunjukkan oleh histogram pada gambar 10.



Gambar 10 Histogram Distribusi Akurasi Model pada Uji Validasi

Pada histrogram di atas terlihat bahwa distribusi nilai akurasi model pada uji validasi tersebar dari rentang nilai 78,33% hingga 87,33%. Nilai akurasi model pada uji validasi paling banyak terletak pada rentang nilai 83,95% hingga 85,08%. Nilai tengah rentang nilai yaitu senilai 84,5% akan digunakan sebagai nilai ambang batas akurasi minimum model klasifikasi. Terdapat 11 model klasifikasi yang selanjutnya diujikan menggunakan data *testing.*

3.7 Pengujian Sistem

Bagian ini menyajikan hasil evaluasi model-model klasifikasi yang diujikan dengan prediksi data *testing.* Hasil pengujian sistem dievaluasi menggunakan *Confusion Matrix* yang menghasilkan nilai tolak ukur berupa akurasi, presisi, recall, dan f-1 score pada masing-masing model. Tabel komparasi hasil uji prediksi testing masing-masing model ditunjukkan oleh tabel 8.

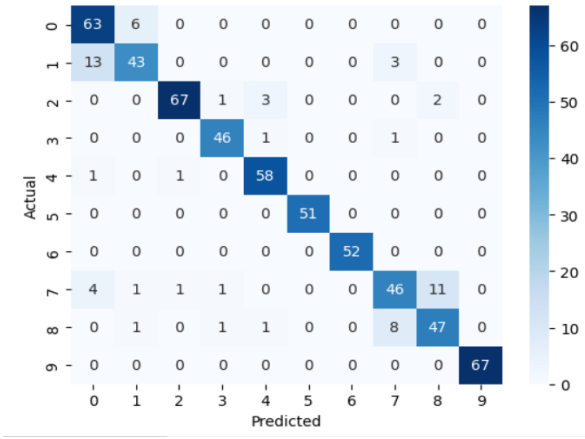
Tabel 8 Hasil Uji Prediksi Data Testing

Spesifikasi Model Klasifikasi	Fitur yang Digunakan	Akurasi	Presisi	Recall	F-1 Score
SVM OVO dengan PCA 95% Varians Kumulatif	20 Principal Component	86,86	87.28	86.86	86,83
SVM OVR dengan PCA 95% Varians Kumulatif	20 Principal Component	87,85	87.93	87.85	87.74
SVM OVO dengan PCA 96% Varians Kumulatif	21 Principal Component	86,52	86,82	86,52	86,48
SVM OVO dengan PCA 97% Varians Kumulatif	23 Principal Component	87,69	88,08	87,69	87,68
SVM OVO dengan PCA 98% Varians Kumulatif	25 Principal Component	88,19	88,69	88,19	88,18
SVM OVO dengan PCA 99% Varians Kumulatif	29 Principal Component	89,02	89,45	89,02	88,96
SVM OVO dengan RFECV Estimator SVM Step 3	48 fitur	88,52	88,91	88,52	88,42
SVM OVO dengan RFECV estimator Random Forest Step 3	18 fitur	89,85	89,98	89,85	89,78
SVM OVR dengan RFECV estimator Logistic Regression Step 1	33 fitur	86,52	86,99	86,52	86,18
SVM OVO dengan RFECV Estimator Logistic Regression	33 fitur	87,19	87,46	87,19	86,94

Step 1					
SVM OVO tanpa seleksi fitur	78 fitur	89,02	89,37	89,02	88,98

3.8 Analisis Hasil Pengujian

Berdasarkan tabel 8, model yang memiliki performa terbaik secara keseluruhan adalah model klasifikasi SVM OVO dengan 18 fitur hasil seleksi fitur RFECV dengan estimator Random Forest dan Step=3. Pada analisis model ini akan digunakan tabel *Confusion Matrix* untuk menghitung nilai performanya seperti ditunjukkan gambar 11 .



	precision	recall	f1-score	support
BOP	0.78	0.91	0.84	69
BOPF	0.84	0.73	0.78	59
Bohea	0.97	0.92	0.94	73
Dust	0.94	0.96	0.95	48
DustII	0.92	0.97	0.94	60
FI	1.00	1.00	1.00	51
FII	1.00	1.00	1.00	52
PF	0.79	0.72	0.75	64
PFII	0.78	0.81	0.80	58
PFIII	1.00	1.00	1.00	67
accuracy			0.90	600
macro avg	0.90	0.90	0.90	600
weighted avg	0.90	0.90	0.90	600

Gambar 11 Confusion Matrix Model Klasifikasi dengan Performa Terbaik

Berdasarkan gambar 11, sampel kelas F I, F II, dan PF III mampu mengklasifikasikan dengan benar seluruh data sebagai kelas asli alias nilai akurasinya 100%. Kesalahan klasifikasi signifikan terjadi pada kelas BOPF sebanyak 13 sampel yang salah diklasifikasikan sebagai kelas BOP dan terdapat 11 sampel kelas PF yang salah diklasifikasikan sebagai kelas PF II. Nilai presisi terendah bernilai 78% pada kelas BOP dan kelas PF II. Nilai recall dan F-1 Score terendah bernilai 72% dan 75% pada kelas PF. Model dengan performa terbaik memiliki nilai akurasi 89,85%, presisi 89,98, recall 89,85%, dan F-1 Score bernilai 89,78%. Pada model dengan performa terbaik ini menggunakan 18 fitur dalam mengklasifikasikan 10 kualitas teh. Tiap fitur ini memiliki skor kepentingan masing-masing yang mengukur seberapa berkontribusinya fitur ini pada prediksi ditunjukkan pada Tabel 9

Tabel 9 Tabel Skor Kepentingan Fitur

Fitur	**Skor Kepentingan**
TGS 2600_Max	0,257
TGS 2600_Range	0.106
MQ 9_Max	0.093
MQ 7_Max	0,092
TGS 2600_Min	0,081
MQ 7_Min	0,066
TGS 2611_Max	0,063
MQ 9_Modus	0,047
MQ 7_Modus	0,028
TGS 2600_Modus	0,026
TGS 2611_Min	0,024
TGS 2602_Max	0,022
TGS 822_Max	0,021
TGS 2611_Modus	0,018
TGS 2602_Modus	0,015
TGS 2600_Q75	0,015
TGS 2602_Min	0,014
TGS 822_Range	0.013

Pada tabel 9 ditunjukkan bahwa fitur TGS 2600_Max itu merupakan fitur yang memiliki skor kepentingan tertinggi yaitu 0,257. Selain itu, fitur-fitur yang cukup berkontribusi besar pada prediksi model adalah fitur TGS 2600_Range dan MQ 9_Max. Fitur dari sensor TGS 2600 juga mendominasi dari keseluruhan fitur, ini berarti nilai sensor TGS 2600 memberikan representasi paling besar pada tiap kelas. Dari tabel tersebut juga diketahui bahwa jenis fitur statisik nilai maksimum adalah nilai paling informatif untuk klasifikasi kualitas teh.

4. KESIMPULAN

Berdasarkan penelitian ini, disimpulkan bahwa fitur yang tepat untuk mengklasifikasikan kualitas teh menjadi 10 kelas berdasarkan aroma ampas teh adalah 18 fitur yang terdiri dari nilai maksimum TGS 2600, nilai range TGS 2600, nilai maksimum MQ 9, nilai maksimum MQ 7, nilai minimum TGS 2600, nilai minimum MQ 7, nilai maksimum TGS 2611, nilai modus MQ 9, nilai modus MQ 7, nilai modus TGS 2600, nilai minimum TGS 2611, nilai maksimum TGS 2602, nilai maksimum TGS 822, nilai modus TGS 2611, nilai modus TGS 2602, nilai kuartil ketiga (Q75) TGS 2600, nilai minimum TGS 2602, dan nilai range TGS 822.

5. SARAN

Saran yang bisa diterapkan dari penulis untuk penelitian selanjutnya adalah mencoba menggunakan jenis fitur-fitur yang berbeda dari penelitian ini. Selain itu, bisa dilakukan metode seleksi fitur yang berbeda dari penelitian ini dan menambah jumlah sampel dataset sehingga akan menambah variasi dalam penelitian.

6. DAFTAR PUSTAKA

[1] F. A. A. o. T. U. Nations, "Market and Trade," [Online]. Available: https://www.fao.org/markets-and-trade/commodities-overview/beverages/tea/en. [Accessed 4 October 2024].

[2] A. N. Komariyah, B. Rohmatulloh, Y. Hendrawan, S. M. Sutan, D. F. Al Riza and M. B. Hermanto, "Klasifikasi Kualitas Teh Hitam Menggunakan Metode Convolutional Neural Network (CNN) Berbasis Citra Digital," *Jurnal Ilmiah Rekayasa Pertanian dan Biosistem,* vol. 11, no. 2, pp. 221-231, September 2023.

[3] A. D. Guritno, A. Harjoko, M. R. Tanuputri, D. U. K. Diyah Utami Kusumaning Putri and N. A. S. Putro, "Development of a portable electronic nose for the classification of tea quality based on tea dregs aroma," *International Journal on Smart Sensing and Intelligent System,* vol. 17, no. 1, 2024.

[4] M. I. P.-. Atmaja, H. Maulana, S. G. P. Riski, A. Fauziah and S. Harianto, "Evaluation of the conformity of the quality of tea products with the requirements of the," *Jurnal Standardisasi ,* vol. 23, no. 1, pp. 43-52, 2021.

[5] B. S. Nasional, "Standar Nasional Indonesia "Teh Hitam"," Badan Standardisasi Nasional, Jakarta, 2016.

[6] K. M. I. Tazi and M. F. Falah, "Klasifikasi Pola Aroma Teh Hijau Menggunakan Hidung Elektronik (E-Nose) Berbasis Linear Diskriminan Analisis(LDA)," *Jurnal Pendidikan MIPA,* vol. 12, no. 3, 2022.

[7] G. Ren, R. Wu, L. Yin, Z. Zhang and J. Ning, "Description of tea quality using deep learning and multi-sensor," *Journal of Food Composition and Analysis ,* vol. 126, 2024.

[8] O. K. Kombo , N. Ihsan, T. S. Syahputra, S. N. Hidayat, M. Puspita, W. R. Roto and K. Triyana, "Enhancing classification rate of electronic nose system and," *Scientific African,* vol. 24, 2024.

[9] H. Xia, W. Chen, D. Hu, A. Miao, X. Qiao, G. Qiu, J. Liang, W. Guo and C. Ma, "Rapid discrimination of quality grade of black tea based on near-infrared spectroscopy (NIRS), electronic nose (E-nose) and data fusion," *Food Chemistry,* vol. 440, May 2024.

[10] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein and S. M. F. El-Salhi, " The Effect of Preprocessing Techniques, Applied to Numeric," *Data,* vol. 6, no. 2, 2021.

[11] . C. S. . K. Dash, A. K. Behera, . S. Dehuri and . A. Ghosh, "Anoutliers detection andelimination frameworkinclassification task of data," *Decision Analytics Journal,* vol. 6, 2023.