Program Studi : Teknik Informatika

LAPORAN PENELITIAN

PENGARUH OVERSAMPLING DAN TRANSFORMASI SPEKTRAL PADA PEMODELAN KLASIFIKASI SPEKTROSKOPI NEAR INFRARED
(Studi Kasus: Prediksi Kultivar Mangga)

Oleh:

Ali Khumaidi (0324108103)

Ridwan Raafi'udin (0002058804)

Setyo Budiono (1870231197)

Panji Kusmanto (1770231084)

FAKULTAS TEKNIK

UNIVERSITAS KRISNADWIPAYANA

JAKARTA

JUNI 2021

HALAMAN PENGESAHAN

1. Judul Penelitian : Pengaruh Oversampling Dan Transformasi Spektral Pada Pemodelan Klasifikasi Spektroskopi Near Infrared (Studi Kasus: Prediksi Kultivar Mangga)
2. Bidang Penelitian : Machine Learning
3. Ketua Peneliti
   a. Nama Lengkap dan Gelar : Ali Khumaidi, S.Kom, M.Kom
   b. Jenis Kelamin : Laki-Laki
   c. Golongan/Pangkat/NIDN : 0324108103
   d. Jabatan Fungsional : Lektor
   e. Program Studi : Teknik Informatika
4. Alamat Ketua Peneliti
   a. Alamat Rumah : Jl. Batu Tumbuh, Jaticempaka, Pondok Gede, Kota Bekasi
   b. E-mail : alikhumaidi@unkris.ac.id
   c. Telp/HP : 087770047269
5. Jumlah Anggota Peneliti 3 orang
   a. Nama Anggota 1 : Ridwan Raafi'udin
   b. Nama Anggota 2 : Setyo Budiono
   c. Nama Anggota 3 : Panji Kusmanto
6. Lokasi Penelitian : Kampus Unkris
7. Kerjasama dengan Institusi Lain
   a. Nama Istitusi(jika ada) :
   b. Alamat :
   c. Telepon/ E-mail :
8. Lama Penelitian bulan : 6 Bulan
9. Biaya yang diperlukan
   a. Sumber dari FT UNKRIS : Rp 10.000.000,00
   b. Sumber Lain, sebutkan : Rp
   Total : Rp 10.000.000,00

Bekasi, 30 Juni 2021

Menyetujui,

Ketua Pengusul Dekan

Dr Harjono Padmono Putro, ST, MKom
NIDN : 0329067102

Nuke L Chusna, S.Si, M.Kom
NIDN : 0315066706

Mengetahui
Ketua P2M FT UNKRIS

Ir Sutaryo, MSi
NIDN :0321126001

RINGKASAN

Near-in spectroscopy (NIR) adalah teknik analisis non-destruktif yang mampu memberikan informasi kimia dan struktur pada sampel dalam waktu yang sangat cepat dan akurat. Namun, pita absorbansi spektrum NIR seringkali luas, tidak spesifik, dan tumpang tindih. Analisis spektrum NIR memerlukan metode multivariat yang sangat subjektif terhadap noise yang timbul dari instrumentasi. Sampai saat ini belum ada protokol standar dalam pemodelan untuk klasifikasi dan prediksi menggunakan spektrum NIR, beberapa model telah dikembangkan dengan dan tanpa teknik preprocessing. Teknik SMOTE dapat memperbaiki model sehingga dapat memprediksi semua respon kelas secara akurat. Penelitian ini berkontribusi pada pengembangan model klasifikasi multikelas untuk pengelompokan kultivar mangga dengan menemukan teknik preprocessing terbaik dan menggunakan oversampling SMOTE. Hasil dari 4 skenario pengujian terhadap performansi model yang dibangun menggunakan Support Vector Machine (SVM) bahwa model terbaik diperoleh dengan menggunakan transformasi spektral dengan operasi LSNV dan CLIP dengan nilai akurasi, presisi dan recall 100%, sedangkan pada Decision Tree (DT) hasil performansi model 100% diperoleh dengan menggunakan transformasi spektral dengan operasi LSNV, CLIP dan SAVGOL dengan parameter {'deriv_order': 0,1, 2, 'filter_win': 11, 13, 'poly_order': 3} . Akurasi 92% pada SVM dan 94% pada DT. Sedangkan kombinasi SMOTE dan Spectral Transformation memberikan hasil klasifikasi untuk SVM dan DT dengan akurasi yang sama yaitu 96%.

DAFTAR ISI

HALAMAN SAMPUL	i
HALAMAN PENGESAHAN	ii
RINGKASAN	iii
DAFTAR ISI	iv
BAB 1. PENDAHULUAN	1
BAB 2. TINJAUAN PUSTAKA	3
BAB 3. METODE PENELITIAN	7
BAB 4. HASIL DAN PEMBAHASAN	9
BAB 5. KESIMPULAN DAN SARAN	19
DAFTAR PUSTAKA	20
LAMPIRAN	

BAB 1. PENDAHULUAN

Kemajuan teknologi telah mendorong inovasi dalam pengembangan teknologi untuk menentukan karakteristik dan kualitas buah secara non-destruktif [1]. Teknologi ini dapat mengukur kualitas buah berdasarkan klasifikasi faktor eksternal dan internal yang tidak dapat dideteksi oleh indera manusia. Pada umumnya metode penentuan kualitas saat ini lebih banyak dilakukan secara destruktif yang membutuhkan waktu, tenaga, biaya dan terdapat faktor bias karena subjektifitas manusia [2]. Sehingga metode pengukuran kualitas dan deteksi destruktif tidak cocok diterapkan di industri. Metode pengukuran kualitas non-destruktif lebih efektif berdasarkan korelasi antara sifat fisik buah yang dikaitkan dengan faktor kualitas buah. Penggunaan peralatan non-destruktif menghasilkan hasil yang lebih konsisten dibandingkan tenaga manusia, sehingga meminimalkan kemungkinan kesalahan dalam menentukan kualitas buah [3].

Penerapan Near Infrared Spectroscopy (NIR) dengan memanfaatkan sinar infra merah bukanlah hal baru, spektroskopi NIR dikembangkan pada tahun 1950 di bidang industri yang berfokus pada analisis kandungan kimia bahan [4]. Spektroskopi NIR mulai banyak digunakan untuk menganalisis kadar air, protein dan lemak pada produk pertanian dan pangan. Spektroskopi NIR merupakan teknik analisis non-destruktif yang mampu memberikan informasi kimia dan struktur pada sampel tertentu dalam waktu yang sangat cepat (kurang dari 1 menit). NIR memiliki panjang gelombang 750-2500 nm, sampel target disinari dengan cahaya dan cahaya yang dipantulkan atau backscatter diukur dengan spektrometer. Produk hortikultura juga dapat memanfaatkan metode NIR ini dalam proses grading, sortasi, internal quality, dan penentuan waktu panen. Penentuan kualitas produk hortikultura dapat dilakukan secara non-destruktif dengan menggunakan spektroskopi NIR yang telah diterapkan pada semangka dan melon [6][7], menggunakan spektral untuk deteksi Kandungan Lycopene pada Tomat [8], deteksi kualitas mangga [9][10].

Tidak ada protokol standar dalam pemodelan untuk klasifikasi dan prediksi menggunakan spektrum NIR. Beberapa model telah dikembangkan dengan dan tanpa teknik preprocessing. Support Vector Machine (SVM) merupakan algoritma yang paling banyak digunakan dalam model prediksi baik klasifikasi maupun regresi untuk pendeteksian kualitas buah dan memiliki akurasi yang cukup baik. Deteksi kualitas teh hitam menggunakan transformasi spektral standard normal variate (SNV), dan kombinasi Savitzky Golay dengan turunan pertama dan algoritma SVM [11]. Pengukuran kekerasan buah pir menggunakan transformasi spektral SNV dan turunan pertama dengan SVM [12].

Teknik transformasi spektral dapat meningkatkan kinerja model [13]. Teknik-teknik tersebut antara lain: Smoothing, Scatter Correction, Trimming, Clipping,

Resampling dan Derivatives. Urutan operasi preprocessing yang diterapkan dapat mempengaruhi kinerja model [14]. Smoothing bertujuan untuk menghaluskan spektral dan membantu menghilangkan noise. Koreksi hamburan bertujuan untuk melawan efek ukuran partikel. Pemangkasan memungkinkan ekstraksi daerah panjang gelombang kontinu dan non-kontinyu dari data spektral penuh. Kliping bertujuan untuk menghapus atau mengganti titik data dengan nilai yang melebihi ambang batas yang ditentukan pengguna. Resampling memproses resolusi spektral baru menggunakan metode Fourier yang dapat menggabungkan spektral yang diperoleh dengan beberapa perangkat yang memiliki resolusi spektral berbeda.

Kelas keseimbangan adalah suatu kondisi distribusi yang tidak seimbang antar kelas dalam suatu dataset, dimana satu kelas memiliki jumlah data yang sangat besar (kelas mayoritas) dibandingkan dengan kelas lainnya (kelas minoritas) [15]. Perbedaan jumlah data yang besar antar kelas dapat mengakibatkan model klasifikasi seringkali tidak dapat memprediksi kelas minoritas dengan tepat sehingga banyak data uji yang seharusnya berada pada kelas minoritas diprediksi salah oleh model klasifikasi [16] . Untuk mengatasi masalah ketidakseimbangan kelas, salah satu metode yang digunakan adalah sampling. Metode sampling memodifikasi distribusi data antara kelas mayoritas dan kelas minoritas dalam dataset pelatihan untuk menyeimbangkan jumlah data untuk setiap kelas [17]. Salah satu metode pengambilan sampel yang sering digunakan adalah Syntetic Minority Oversampling Technique (SMOTE).

Tujuan dari penelitian ini adalah (1) untuk mengetahui pengaruh kinerja metode transformasi spektral dan operasi yang paling optimal pada dataset; (2) untuk mengetahui pengaruh keseimbangan data terhadap model; (3) mengeksplorasi model klasifikasi machine learning yang optimal berdasarkan penerapan transformasi spektral dan keseimbangan data.

BAB 2. TINJAUAN PUSTAKA

A. Near Infrared

Infrared adalah gelombang elektromagnetik yang memiliki rentang panjang gelombang diantara panjang gelombang sinar tampak dan gelombang mikro. Berdasarkan panjang gelombangnya, infrared dibagi menjadi near infrared, mid infrared dan far infrared. Menurut Naripati (2017) , Near Infrared Spectroscopy (NIRS) merupakan sebuah teknologi yang dapat menduga kandungan kimia suatu bahan dari spektrum absorbansi yang dihasilkan. Teknologi ini dapat digunakan pada objek pengamatan dengan cara menghancurkan (destruktif) maupun tanpa menghancurkan (nondestruktif) objek pengamatan tersebut. Near Infrared dapat digunakan untuk menentukan kandungan kimia suatu bahan organik, dikarenakan ikatan molekul bahan organik sangat peka terhadap panjang gelombang ini . Sebagian besar bahan organik terdiri atas beberapa atom-atom seperti karbon, oksigen, hidrogen, nitrogen, fosfor dan sulfur yang membentuk suatu molekul dan bergerak konstan dengan adanya ikatan kovalen dan elektrokovalen. Molekul-molekul tersebut akan bervibrasi apabila berinteraksi dengan gelombang NIR dikarenakan adanya pergerakan atom yang mendekati atau menjauhi atom yang lainnya. NIRS menggunakan gelombang elektromagnetik dengan panjang gelombang sebesar 780 nm – 2500 nm atau jumlah gelombang per cm 12.800 cm-1 hingga 4000 cm-1.

Menurut Munawar (2008), ketika ada sebuah sinar yang berasal dari suatu sumber jatuh mengenai obyek tertentu, maka yang akan terjadi adalah sebuah interaksi antara obyek dengan sinar tersebut. Obyek yang dikenai sinar tersebut akan memberi respon yang berupa pantulan, serapan dan terusan. Respon pantulan (reflectance) dapat berupa pantulan langsung (specular reflectance) yang mana sinar sepenuhnya dipantulkan kembali oleh obyek, dan pantulan semu (diffuse reflectance) yang mana sinar diserap terlebih dahulu oleh obyek, kemudian dipantulkan kembali. Respon serapan (absorbance) terjadi dimana seluruh sinar pada panjang gelombang tertentu sepenuhnya diserap oleh bahan, sedangkan respon terusan (transmittance) merupakan sebuah respon dimana sinar yang mengenai obyek pada panjang gelombang tertentu akan diteruskan menembus obyek tersebut.

Kelebihan dari penggunaan teknologi Near Infrared Spectroscopy (NIRS) untuk pengujian kandungan kimia suatu bahan pertanian adalah tidak memerlukan bahan-bahan kimia dalam pengujiannya, serta proses yang dilakukan dapat lebih cepat dan akurat. Pada daerah panjang gelombang NIRS memiliki dua keunggulan utama yaitu memiliki kecepatan yang tinggi untuk akuisisi spektral yang memfasilitasi pengumpulan data deskriptif sehingga prasyarat untuk sistem

kontrol dapat bekerja secara real time dan spektrum dapat direkam dari berbagai macam bahan.

B. Prapemrosesan

Tujuan dari preprocessing data adalah untuk mereduksi sifat fisis pada spektrum, sehingga dapat mengurangi variabilitas yang disebabkan oleh hamburan cahaya, nonlinier, dan memperbaiki model yang akan digunakan [19]. Transformasi spektral yang digunakan adalah resampling, clipping, smoothing, derivatif dan koreksi pencar. Dalam smoothing menggunakan savitzky golay filtering dengan parameter : filter windows (7, 11, 13), orde polinomial = 3, dan orde turunan (0, 1, 2). Koreksi hamburan menggunakan beberapa operasi, yaitu koreksi pencar berganda (MSC), variat normal standar (SNV), variat normal kuat (RNV), normalisasi, baseline, detrend, versi lokal SNV (LSNV) dan Versi diperpanjang MSC (EMSC) .

Metode MSC adalah menggambar titik sampel spektral untuk mendekati spektrum referensi dengan memanfaatkan hasil estimasi parameter regresi linier sederhana dan dapat menghilangkan variasi antar spektrum dengan mengoreksi posisi hamburan setiap nilai intensitas setiap replikasi ke posisi hamburan rata-rata. intensitas seluruh replikasi [20]. Metode SNV menghilangkan efek hamburan dari spektrum dengan memusatkan dan mengatur skala masing-masing spektrum [21]. Metode RNV lebih cocok untuk data dengan banyak noise dengan menggunakan konsep koreksi berdasarkan nilai median dan interval antar kuartil[22]. Metode normalisasi spektral menggunakan rentang nilai tertentu dan biasanya menggunakan Euclidean. Metode baseline pada prinsipnya menggunakan rata-rata nilai sentral dari spektral. Konsep metode LSNV mirip dengan SNV dengan prinsip operasi pembagian pada jendela spektral. Metode EMSC pada prinsipnya hampir sama dengan MSC tetapi dalam EMCS memperhitungkan koreksi linier dan kuadratik.

Semua metode dan operasi transformasi spektral ini akan dibandingkan dengan model untuk mendapatkan akurasi yang paling optimal. Penggunaan teknik transformasi spektral sebagai salah satu faktor dalam meningkatkan kinerja model.

C. SMOTE

Pada penelitian ini diidentifikasi bahwa dataset yang digunakan memiliki masalah ketidakseimbangan kelas sehingga diperlukan metode over-sampling untuk mengatasi masalah ketidakseimbangan kelas tersebut. Metode yang dapat digunakan adalah SMOTE. SMOTE adalah metode over-sampling di mana data di kelas minoritas direproduksi menggunakan data sintetis yang berasal dari replikasi

data di kelas minoritas. Over-sampling di SMOTE mengambil instance dari kelas minoritas dan kemudian mencari k-nearest tetangga dari setiap instance, kemudian menghasilkan instance sintetis alih-alih mereplikasi instance kelas minoritas; oleh karena itu, dapat menghindari masalah overfitting yang berlebihan [23]. Algoritma yang bekerja pada SMOTE pertama akan mengambil selisih antara vektor-vektor fitur pada kelas minoritas dengan nilai tetangga terdekat dari kelas minoritas kemudian mengalikan nilai tersebut dengan bilangan acak antara 0 sampai 1. Selanjutnya hasil perhitungan ditambahkan ke vektor fitur sehingga diperoleh hasil nilai vektor. yang baru [24].

Untuk memvalidasi keefektifan model yang diusulkan, dilakukan dua skenario eksperimen yaitu menggunakan pendekatan algoritma SVM dan Decision Tree dan masing-masing digunakan untuk pemodelan tanpa mempertimbangkan ketidakseimbangan kelas, dan kedua, dilakukan oversampling SMOTE untuk menambah jumlah dataset. untuk mencapai dataset yang seimbang.

D. Pemodelan

Pemodelan yang dikembangkan dalam penelitian ini adalah klasifikasi. Model yang dikembangkan akan dikelompokkan berdasarkan spektral NIR pada 4 kelas kultivar mangga. Klasifikasi adalah teknik multivariat untuk memisahkan set objek yang berbeda dan mengalokasikan objek baru ke dalam kelompok yang telah ditentukan. Metode klasifikasi yang baik akan menghasilkan lebih sedikit kesalahan klasifikasi. Untuk melakukan klasifikasi secara akurat, diperlukan metode yang tepat. Support Vector Machine (SVM) merupakan salah satu metode yang dapat melakukan klasifikasi. SVM adalah teknik untuk menemukan hyperplane yang dapat memisahkan dua set data dari dua kelas yang berbeda [25]. SVM memiliki kelebihan diantaranya dalam menentukan jarak menggunakan support vector sehingga proses komputasi menjadi cepat. Proses pembelajaran dalam SVM bertujuan untuk mendapatkan hipotesis berupa bidang pembagi terbaik yang tidak hanya meminimalkan risiko empiris yaitu rata-rata error pada data latih, tetapi juga memberikan generalisasi yang baik. Generalisasi adalah kemampuan suatu hipotesis untuk dapat mengklasifikasikan data yang tidak terdapat dalam data latih dengan benar. Prinsip SVM sebenarnya adalah pengklasifikasi linier kemudian dikembangkan kembali sehingga dapat bekerja pada masalah non-linier menggunakan metode kernel trick yaitu mencari hyperplane dengan mentransformasikan dataset menjadi ruang vektor dengan dimensi yang lebih besar (feature space) menggunakan fungsi kernel yang kemudian akan diklasifikasikan. dan dilakukan pada ruang fitur. Penentuan fungsi kernel yang digunakan akan sangat mempengaruhi hasil klasifikasi [26].

Pohon keputusan merupakan salah satu metode klasifikasi yang paling populer, karena mudah diinterpretasikan oleh manusia. Pohon keputusan merupakan model prediktif dengan menggunakan struktur pohon atau struktur hierarkis [27]. Konsep pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan keputusan. Decision Tree digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan hubungan variabel atribut x dan variabel target y dalam bentuk pohon. Pohon keputusan menyerupai flowchart dimana setiap simpul internal (simpul yang bukan merupakan daun atau simpul terluar) merupakan pengujian dari variabel atribut, setiap cabang merupakan hasil pengujian, sedangkan simpul terluar yaitu daun merupakan . Manfaat utama penggunaan pohon keputusan adalah kemampuannya untuk memecah proses pengambilan keputusan yang kompleks menjadi proses yang lebih sederhana, sehingga pengambil keputusan akan lebih baik dalam menginterpretasikan solusi dari masalah [28].

Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah variabel input potensial dan variabel target. Pohon keputusan menggabungkan eksplorasi dan pemodelan data, sehingga sangat bagus sebagai langkah pertama dalam proses pemodelan bahkan ketika digunakan sebagai model akhir dari beberapa teknik lain. Keuntungan lain dari metode ini adalah dapat menghilangkan perhitungan atau data yang tidak perlu. Hal ini dikarenakan sampel yang ada biasanya hanya diuji berdasarkan kriteria atau kelas tertentu [29].

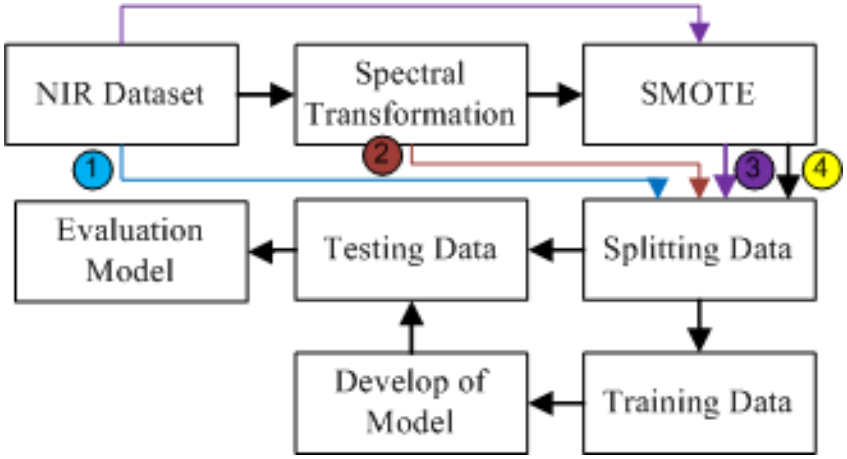
E. Evaluasi model

Dalam mengukur kinerja model klasifikasi menggunakan matriks konfusi. Confusion matrix, juga dikenal sebagai error matrix, memberikan informasi tentang perbandingan hasil klasifikasi yang dilakukan oleh model dengan hasil klasifikasi yang sebenarnya [30]. Terdapat 4 istilah yang mewakili hasil proses klasifikasi pada confusion matrix, yaitu True Positive (TP), True Negative (TN), False Positive (FP) dan False Negative (FN). TP adalah data positif yang diprediksi benar, TN adalah data negatif yang diprediksi benar, FP adalah data negatif tetapi diprediksi data positif dan FN adalah data positif tetapi diprediksi data negatif.

Confusion matrix dapat menghitung berbagai metrik kinerja untuk mengukur kinerja model yang telah dibuat, beberapa di antaranya sering digunakan yaitu akurasi, presisi, dan recall. Akurasi menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Precision menggambarkan tingkat akurasi antara data yang diminta dan hasil prediksi yang diberikan oleh model. Recall menggambarkan keberhasilan model dalam mengambil informasi.

BAB 3. METODE PENELITIAN

Tahapan penelitian meliputi persiapan dataset, preprocessing atau transformasi spektral untuk mendapatkan data yang paling optimal untuk mendukung model, SMOTE untuk menangani ketidakseimbangan kelas, memecah data menjadi data pelatihan dan data pengujian, mengembangkan model dan mengevaluasi model untuk mengetahui yang terbaik. model. Pemodelan akan membandingkan antara support vector machine (SVM) dan algoritma pohon keputusan. Untuk mengetahui performansi model klasifikasi dengan mengevaluasi model dengan 4 skenario yaitu (1) tanpa perlakuan apapun; (2) menerapkan SMOTE; (3) menerapkan transformasi spektral; (4) menerapkan SMOTE dan transformasi spektral. Hubungan antar tahapan dapat dilihat pada Gambar 1.



Gambar 1. Metodologi Penelitian

A. Dataset

Dataset yang digunakan berasal dari hasil penelitian [18]. Sebanyak 186 sampel mangga utuh dari 4 kultivar yang berbeda (Cengkir, Kweni, Kent dan Palmer) diambil menggunakan spektral inframerah dekat yang diperoleh berupa absorbansi dengan panjang gelombang 1000 sampai 2500 nm. Jumlah sampel kultivar Cengkir 18, Kweni 29, Kent 85 dan Palmer 54. Dataset ini dapat diakses https://data.mendeley.com/datasets/b9d6s7hr33/1.

B. Transformasi Spektral

Semua metode dan operasi transformasi spektral ini akan dibandingkan dengan model untuk mendapatkan akurasi yang paling optimal. Penggunaan teknik transformasi spektral sebagai salah satu faktor dalam meningkatkan kinerja model.

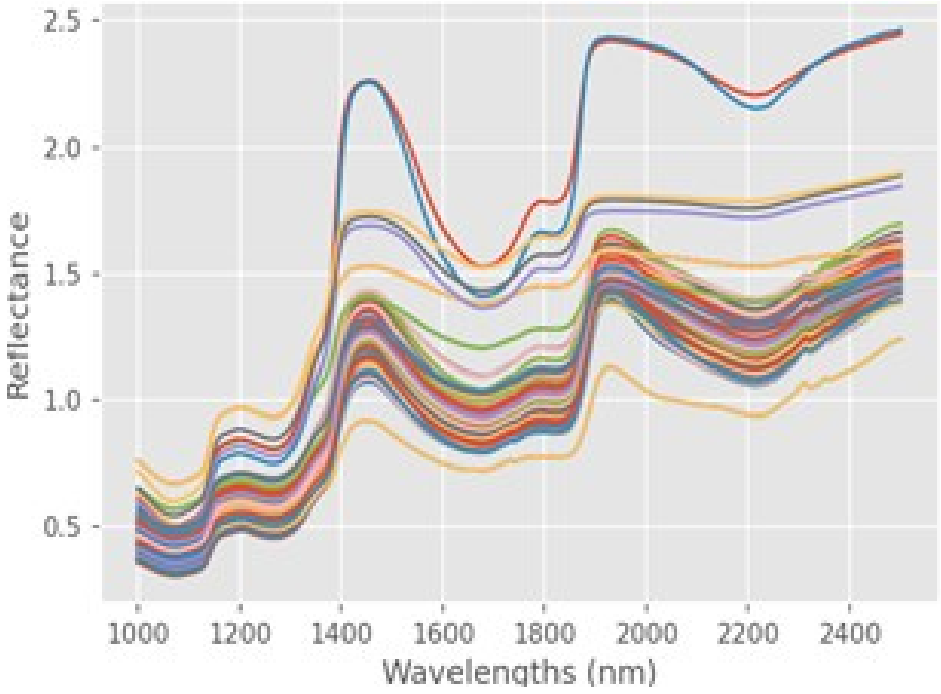
Penggunaan metode transformasi spektral secara lebih rinci dapat dilihat pada Tabel 1.

Tabel 1. Metode Transformasi Spektral

Method	**Operation**	**Parameter**	**Value**
Resampling	RESAMPLE	Rasio	0.8
Clipping	CLIP	Threshold	1e4
		substitute	None
Smoothing	SAVGOL	filter_win	7, 11, 13
		poly_order	3
		deriv_order	0, 1, 2
Scatter Correction	MSC		
	SNV		
	RNV	iqr	75-25, 90-10
	NORML		
	BASELINE		
	LSNV		
	EMSC		

BAB 4. HASIL DAN PEMBAHASAN

Dataset spektral NIR dari 186 sampel dengan panjang gelombang 1000 hingga 2500 nm dapat dilihat pada Gambar 2. Terdapat beberapa puncak serapan yang dapat ditemukan dari spektrum transmitansi aslinya. Dataset mangga terdiri dari 4 kultivar sehingga metode klasifikasi multikelas digunakan dalam penelitian ini. Berikut hasil pengukuran performansi model dari 4 skenario dengan sebaran data latih dan data uji 70 dan 30.



Gambar 2. Spektral Orisinal NIR

A. Pemodelan Tanpa Transformasi Spektral

Hasil klasifikasi pengolahan data spektrum NIR menggunakan SVM tanpa preprocessing menghasilkan akurasi yang cukup baik yaitu 90%, kesalahan klasifikasi kelas 6 kultivar Kent yang sebenarnya adalah kultivar Palmer. Hasil klasifikasi menggunakan DT tanpa preprocessing menghasilkan akurasi yang lebih tinggi dari SVM, yaitu 94%, dengan salah menebak 4 kultivar Cengkir yang seharusnya kultivar Palmer. Hasil dari confusion matrix secara lebih rinci dapat dilihat pada Tabel 2.

Tabel 2. Hasil Confusion matrix tanpa transformasi spektral

Algorithm		Confusion Matrix Prediction				
		Class	Cengkir	Kent	Kweni	Palmer	
SVM	Actual	Cengkir	6	0	0	0	
		Kent	0	28	0	0	
		Kweni	0	0	10	0	
		Palmer	0	6	0	12	
	Accuracy						**0.90**
	Precision		1	0.82	1	1	**0.96**
	Recall		1	1	1	0.67	**0.92**
DT	Actual	Cengkir	6	0	0	0	
		Kent	0	28	0	0	
		Kweni	0	0	10	0	
		Palmer	4	0	0	14	
	Accuracy						**0.94**
	Precision		0.6	1	1	1	**0.9**
	Recall		1	1	1	0.78	**0.94**

B. Pemodelan dengan oversampling SMOTE

Dengan penerapan oversampling SMOTE, jumlah data antar kelas seimbang dengan jumlah buah mangga sebanyak 85 buah pada setiap kelas. Kultivar Cengkir yang semula memiliki 18 data, Kweni yang memiliki 29 data dan Palmer yang semula memiliki 54 data disamakan dengan total data Kent yaitu 85 data. Hasil klasifikasi pengolahan data spektrum NIR menggunakan SVM dengan penerapan oversampling SMOTE meningkatkan akurasi klasifikasi sebesar 92% dibandingkan tanpa preprocessing, kesalahan pengelompokan kelas 1 kultivar Palmer yang sebenarnya merupakan kultivar Cengkir, 1 kultivar Cengkir yang sebenarnya adalah kultivar Kent dan kultivar Palmer yang seharusnya diprediksi menjadi 2 kultivar Cup dan 5 Kent.

Hasil klasifikasi menggunakan DT dengan penerapan oversampling SMOTE meningkatkan akurasi klasifikasi sebesar 94% dibandingkan tanpa preprocessing, kesalahan klasifikasi kelas adalah 1 kultivar Palmer dan 1 kultivar Kent yang sebenarnya kultivar Cengkir, 3 kultivar Palmer yang sebenarnya Kent kultivar dan 1 kultivar Cengkir yang seharusnya Palmer. Hasil dari confusion matrix secara lebih rinci dapat dilihat pada Tabel 3.

Tabel 3. Hasil Confusion matrix hasil menggunakan oversampling smote

Algorithm		Confusion Matrix Prediction				
		Class	**Cengkir**	**Kent**	**Kweni**	**Palmer**	
SVM	Actual	Cengkir	27	0	0	1	
		Kent	1	27	0	0	
		Kweni	0	0	29	0	
		Palmer	2	5	0	21	
	Accuracy						**0.92**
	Precision		0.90	0.84	1	0.95	**0.92**
	Recall		0.96	0.96	1	0.75	**0.92**
DT	Actual	Cengkir	26	1	0	1	
		Kent	0	25	0	3	
		Kweni	0	0	29	0	
		Palmer	1	0	0	22	
	Accuracy						**0.94**
	Precision		0.96	0.96	1	0.85	**0.94**
	Recall		0.93	0.89	1	0.96	**0.94**

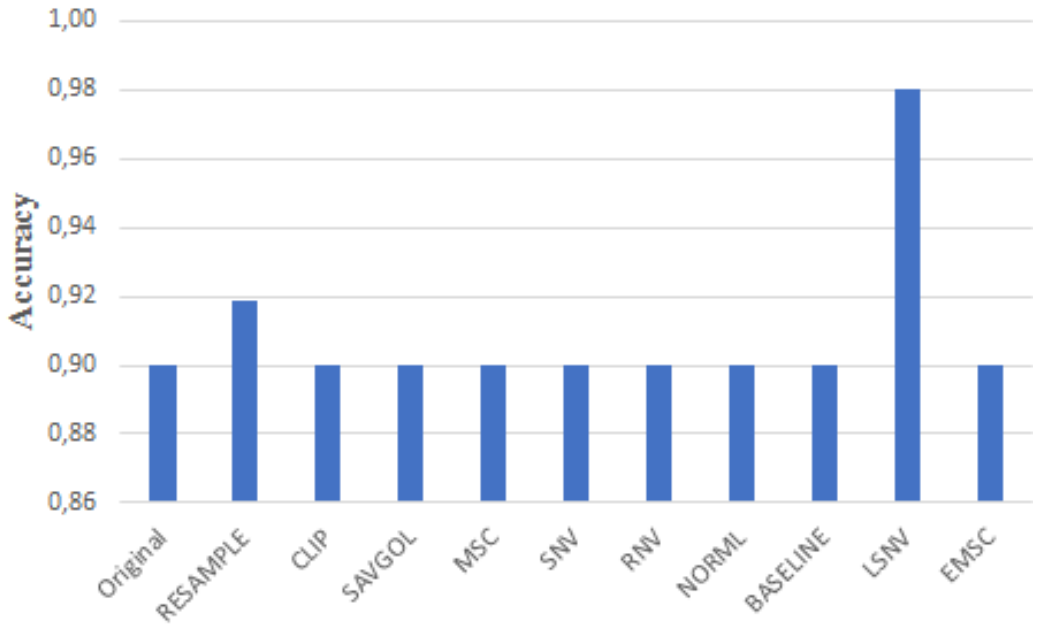
C. Pemodelan dengan transformasi spektral

Hasil klasifikasi pengolahan data spektrum NIR menggunakan SVM dengan penerapan metode transformasi spektral yaitu Smoothing, Scatter Correction, Clipping, Resampling dan Derivatives beserta kombinasi, operasi dan parameter metode tersebut dapat dilihat pada Tabel 1. Menggunakan Nippy library dengan Python, didapatkan hasil klasifikasi dengan nilai akurasi 100%, dimana penggunaan metode Clipping and Scatter Correction dengan operasi LSNV memberikan hasil yang paling optimal untuk SVM. Hasil klasifikasi DT dengan transformasi spektral menggunakan metode Clipping and Scatter Correction dengan operasi LSNV juga menghasilkan akurasi 100% tanpa kesalahan prediksi kelas menggunakan metode Clipping, Scatter Correction dan Smoothing. Hasil dari confusion matrix secara lebih rinci dapat dilihat pada Tabel 4.

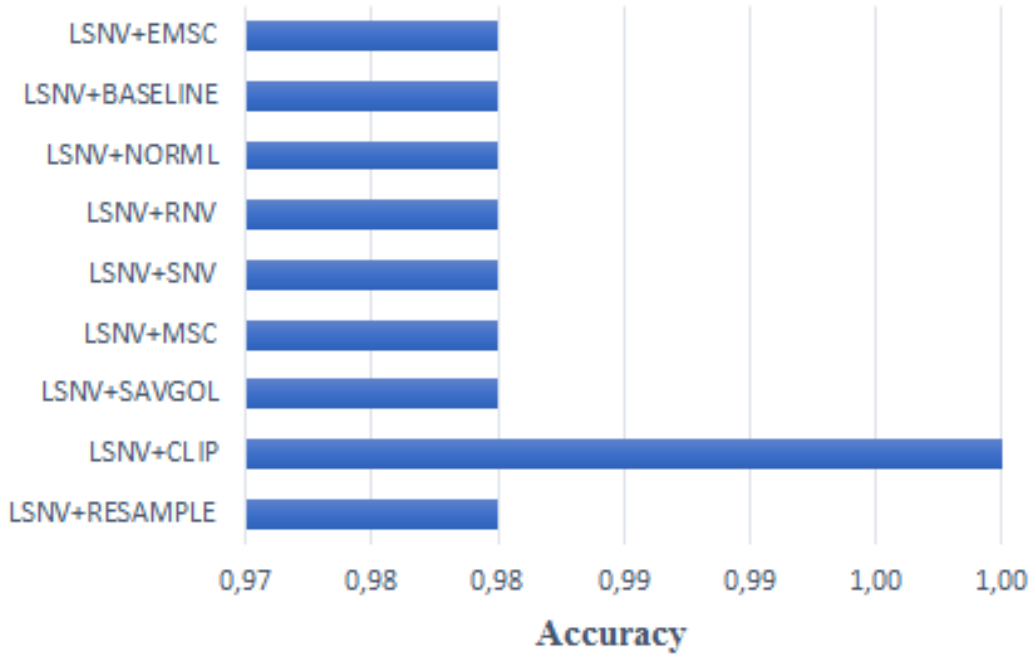
Selama proses perhitungan akurasi klasifikasi menggunakan SVM dan DT dengan kombinasi 5 metode transformasi spektral memberikan hasil yang bervariasi, bahkan ada yang memiliki akurasi lebih buruk dengan nilai akurasi tanpa preprocessing. Oleh karena itu, sangat penting untuk mengadaptasi metode dan operasi transformasi spektral pada data. Untuk mendapatkan hasil akurasi yang paling optimal dalam klasifikasi menggunakan SVM, dilakukan eksperimen pada

semua operasi transformasi spektral. Pertama, pengujian dilakukan pada setiap operasi dan parameternya. Hasil operasi transformasi spektral dengan akurasi terbaik kemudian digabungkan dengan operasi lainnya. Hasil dari 2 kombinasi operasi transformasi spektral tersebut kemudian digabungkan dengan operasi lainnya. Begitu seterusnya hingga diperoleh nilai akurasi yang paling optimal.

Gambar 3 menjelaskan hasil penerapan 1 operasi transformasi spektral pada SVM. Operasi transformasi spektral yang menghasilkan nilai akurasi lebih tinggi dibandingkan tanpa preprocessing adalah RESAMPLE dengan akurasi 92% dan LSNV dengan akurasi 98%. Untuk operasi lainnya, nilai akurasinya 90% sama dengan tanpa preprocessing. Operasi LSNV sebagai operasi transformasi spektral dengan nilai tertinggi kemudian digabungkan dengan operasi lainnya. Hasil akurasi dari 2 kombinasi operasi transformasi spektral pada SVM telah mencapai akurasi 100% dengan menggunakan operasi LSNV dan CLIP dengan threshold=10000. Sedangkan kombinasi lainnya dengan nilai akurasi maksimal 98%. Dengan diperolehnya operasi transformasi spektral yang paling optimal, maka kombinasi selanjutnya akan mencapai optimal lagi, setidaknya dengan menggunakan 2 kombinasi operasi transformasi spektral yaitu LSNV dan CLIP. Untuk hasil akurasi 2 kombinasi operasi transformasi spektral pada SVM dapat dilihat pada Gambar 4.



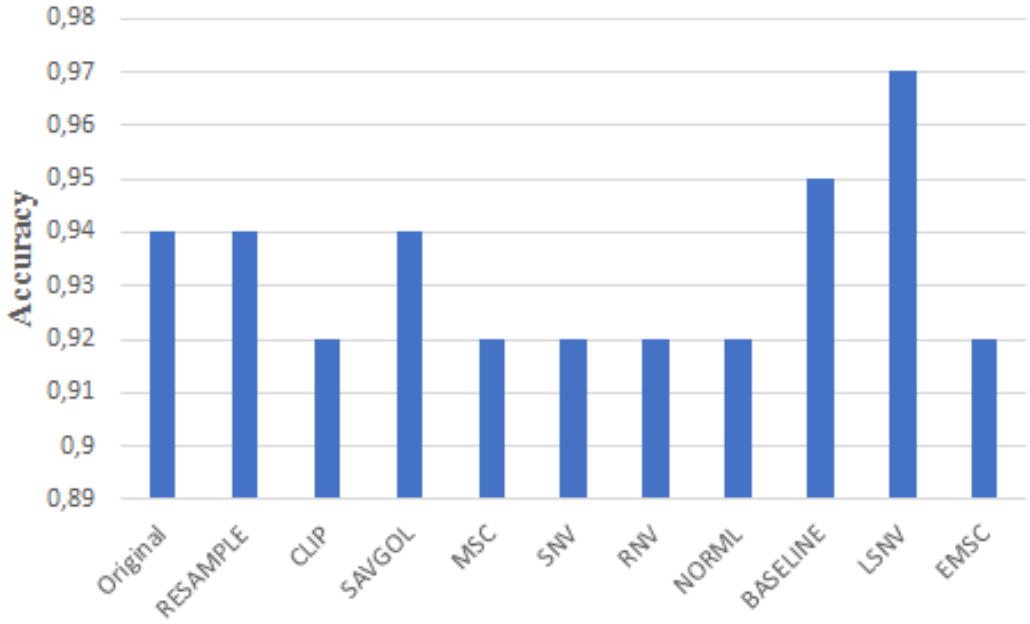
Gambar 3. Hasil akurasi menggunakan operasi transformasi spektral pada SVM



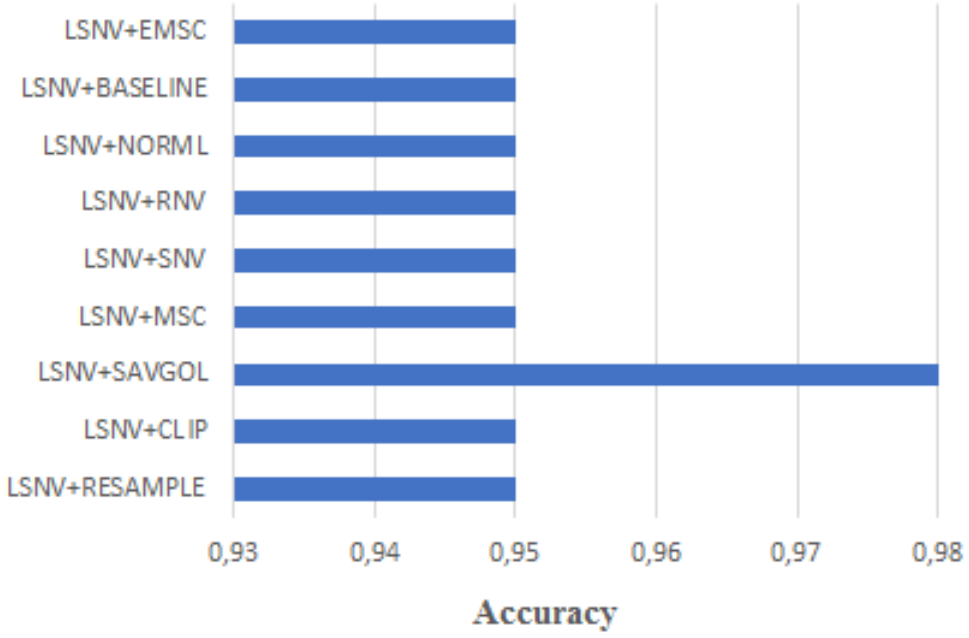
Gambar 4. Hasil akurasi penggunaan 2 kombinasi operasi transformasi spektral pada SVM

Gambar 5 menjelaskan hasil penerapan 1 operasi transformasi spektral ke DT. Operasi transformasi spektral yang menghasilkan nilai akurasi lebih tinggi dibandingkan tanpa preprocessing adalah BASELINE dengan akurasi 95% dan LSNV dengan akurasi 97%. Untuk operasi transformasi spektral lainnya, nilai akurasinya lebih rendah dari nilai akurasi tanpa preprocessing, yaitu 92%. Operasi LSNV sebagai operasi transformasi spektral dengan nilai tertinggi kemudian digabungkan dengan operasi lainnya. Hasil akurasi dari 2 kombinasi operasi transformasi spektral pada DT yang mengalami peningkatan yaitu operasi LSNV dan SAVGOL dengan parameter {'deriv_order': 2, 'filter_win': 11, 'poly_order': 3} dengan nilai akurasi sebesar 98%. Sedangkan kombinasi LSNV dengan operasi transformasi spektral lainnya, nilai akurasi klasifikasi maksimum adalah 95%. Hasil akurasi dengan 2 kombinasi operasi transformasi spektral pada DT dapat dilihat pada Gambar 6. Selanjutnya melakukan 3 kombinasi operasi transformasi spektral menggunakan LSNV dan SAVGOL sebagai operasi dengan akurasi terbaik pada 2 kombinasi operasi transformasi spektral yang dikombinasikan dengan transformasi spektral lainnya operasi. Kombinasi 3 operasi transformasi spektral telah mencapai akurasi 100% dengan menggunakan LSNV, CLIP dengan threshold=10000 dan SAVGOL dengan parameter {'deriv_order': 2, 'filter_win': 11, 'poly_order': 3}, {'deriv_order' : 1, 'filter_win': 11, 'poly_order': 3}, {'deriv_order': 2, 'filter_win': 13, 'poly_order': 3}, dan {'deriv_order': 0, 'filter_win': 13, ' poly_order': 3}. 3 kombinasi transformasi spektral lainnya dengan nilai akurasi maksimum 98%. Dengan diperolehnya operasi transformasi spektral yang paling optimal, maka kombinasi

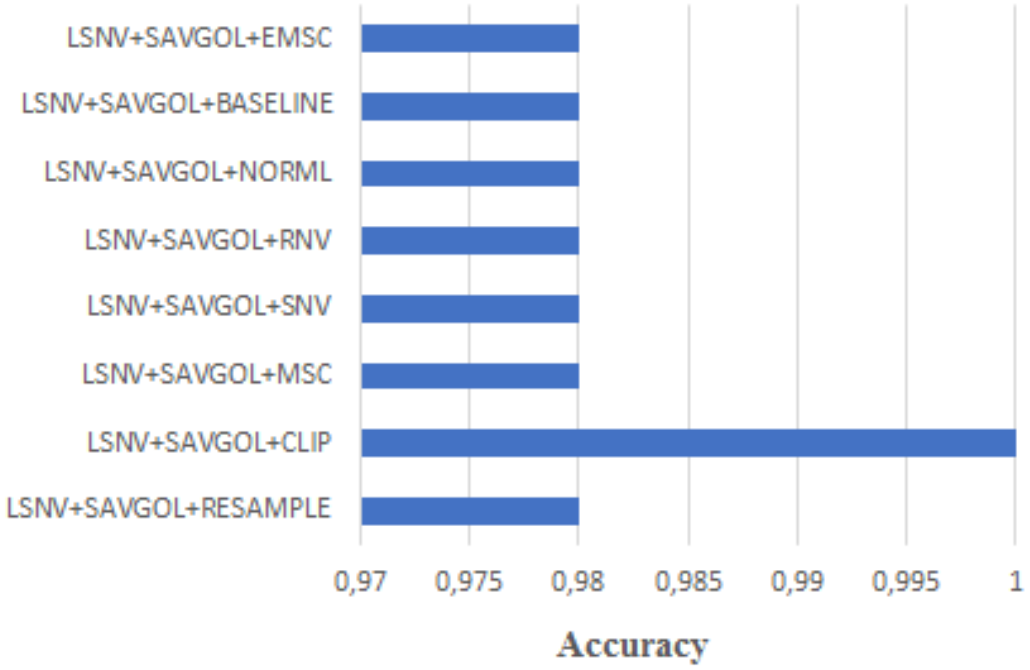
selanjutnya akan mencapai optimal lagi, setidaknya dengan menggunakan 3 kombinasi operasi transformasi spektral yaitu LSNV, CLIP dan SAVGOL. Untuk hasil akurasi 3 kombinasi operasi transformasi spektral pada DT dapat dilihat pada Gambar 7.



Gambar 5. Hasil akurasi menggunakan operasi transformasi spektral pada DT



Gambar 6. Hasil akurasi penggunaan 2 kombinasi operasi transformasi spektral pada DT



Gambar 7. Hasil akurasi penggunaan 3 kombinasi operasi transformasi spektral pada DT

Tabel 4. Hasil kebingungan matriks menggunakan transformasi spectral

Algorithm		Confusion Matrix Prediction				
	Class	Cengkir	Kent	Kweni	Palmer	
SVM	Actual Cengkir	5	0	0	0	
	Actual Kent	0	26	0	0	
	Actual Kweni	0	0	13	0	
	Actual Palmer	0	0	0	18	
	Accuracy					1
	Precision	1	1	1	1	1
	Recall	1	1	1	1	1
DT	Actual Cengkir	5	0	0	0	
	Actual Kent	0	26	0	0	
	Actual Kweni	0	0	13	0	
	Actual Palmer	0	0	0	18	
	Accuracy					1
	Precision	1	1	1	1	1
	Recall	1	1	1	1	1

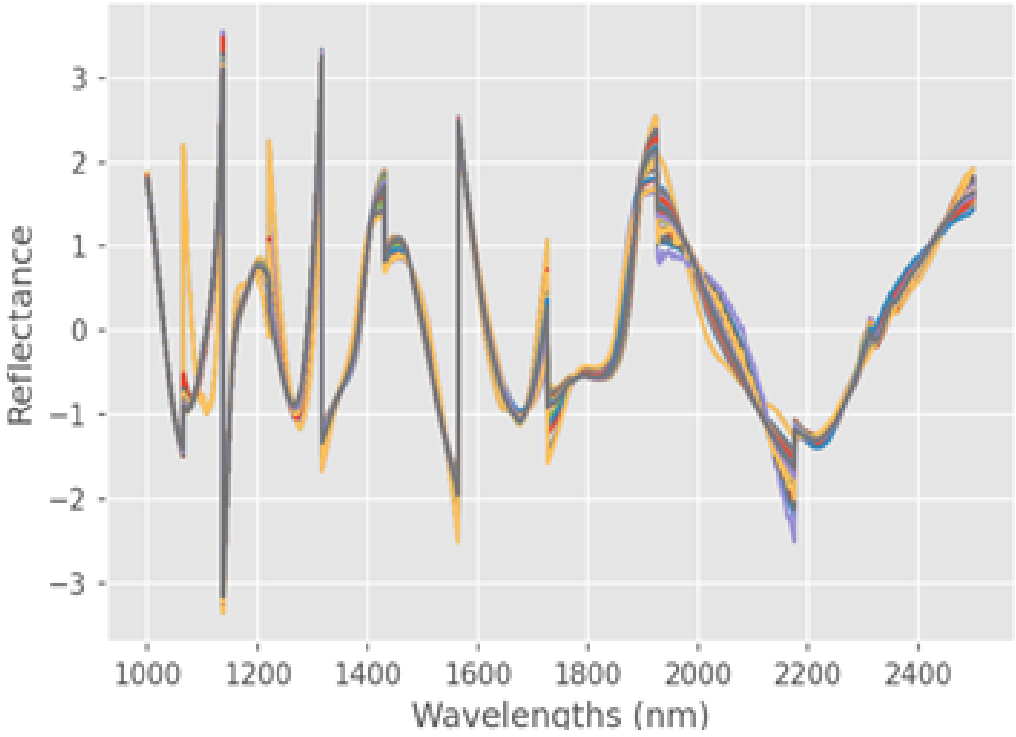
D. Pemodelan dengan oversampling SMOTE dan transformasi spektral

Hasil klasifikasi pengolahan data spektrum NIR menggunakan SVM dengan penerapan SMOTE oversampling dan transformasi spektral, dimana terdapat keseimbangan jumlah data dalam kelas dan penggunaan metode Clipping and Scatter Correction dengan operasi LSNV menghasilkan klasifikasi nilai akurasi 96%. Kesalahan klasifikasi kelas adalah 3 kultivar Kent yang sebenarnya adalah kultivar Cengkir dan 1 kultivar Kent yang sebenarnya adalah kultivar Palmer. Hasil klasifikasi menggunakan DT juga menghasilkan nilai akurasi klasifikasi sebesar 96%. Kesalahan klasifikasi kelas adalah 1 kultivar Palmer yang sebenarnya kultivar Cengkir, 1 kultivar Palmer yang sebenarnya kultivar Kent dan 3 kultivar Kent yang seharusnya Palmer. Hasil dari confusion matrix secara lebih rinci dapat dilihat pada Tabel 5. Kinerja akurasi klasifikasi dengan SVM dan DT memiliki nilai yang sama dan keduanya kurang lebih lebih baik daripada menggunakan transformasi spektral saja. Namun nilai akurasinya lebih tinggi dibandingkan tanpa preprocessing dan penggunaan oversampling SMOTE.

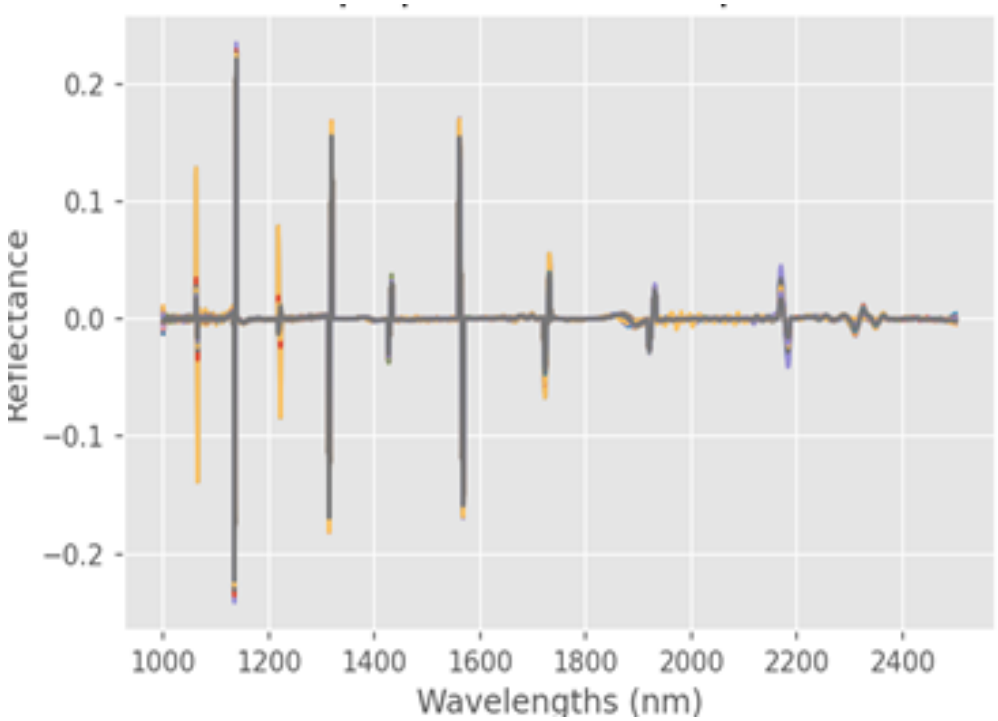
Tabel 5. Hasil confusion matrix menggunakan oversampling smote dan spectral transformation

Algorithm		Confusion Matrix Prediction					
		Class	**Cengkir**	**Kent**	**Kweni**	**Palmer**	
SVM	Actual	Cengkir	25	3	0	0	
		Kent	0	28	0	0	
		Kweni	0	0	29	0	
		Palmer	0	1	0	27	
	Accuracy						**0.96**
	Precision		1	0.88	1	1	**0.97**
	Recall		0.89	1	1	0.96	**0.96**
DT	Actual	Cengkir	27	0	0	1	
		Kent	0	27	0	1	
		Kweni	0	0	29	0	
		Palmer	0	3	0	25	
	Accuracy						**0.96**
	Precision		1	0.90	1	0.93	**0.96**
	Recall		0.96	0.96	1	0.89	**0.96**

Transformasi spektral memiliki pengaruh yang sangat penting dalam meningkatkan akurasi model klasifikasi. Pada keempat skenario pengujian, hasil klasifikasi dengan penerapan metode dan operasi transformasi spektral yang tepat mampu mencapai akurasi 100%. Hasil transformasi spektral terbaik dengan metode Clipping and Scatter Correction dengan operasi LSNV untuk pemodelan dengan SVM dapat dilihat pada Gambar 8 dan untuk pemodelan DT dapat dilihat pada Gambar 9. Pada kedua gambar puncak spektrum terlihat sangat jelas sehingga memudahkan untuk model untuk mengklasifikasikan.



Gambar 8. Hasil transformasi spektral terbaik pada SVM



Gambar 9. Hasil transformasi spektral terbaik pada DT

Tabel 4 merupakan hasil perbandingan 4 skenario pada algoritma SVM dan DT dengan pengukuran kinerja berdasarkan akurasi, presisi dan recall. Dengan model yang dibangun oleh SVM dan DT kinerja terbaik dengan penerapan transformasi spektral kemudian penerapan oversampling SMOTE dan transformasi spektral, kemudian penerapan oversampling SMOTE dan terakhir tanpa preprocessing.

BAB 5. KESIMPULAN DAN SARAN

Untuk meningkatkan akurasi model klasifikasi 4 kelas kultivar mangga berdasarkan spektroskopi NIR, diperlukan perlakuan spektral. Perlakuan dengan oversampling SMOTE dapat meningkatkan akurasi klasifikasi pada algoritma SVM dan DT. Perlakuan dengan transformasi spektral menggunakan 5 kombinasi metode diperoleh akurasi klasifikasi yang optimal sebesar 100% pada SVM menggunakan metode clipping dan koreksi scatter yaitu LSNV, sedangkan pada DT menggunakan metode clipping, koreksi scatter dan smoothing. Ketika penerapan transformasi spektral optimal dengan penggunaan oversampling SMOTE, akurasi klasifikasi tidak lebih baik. Perlakuan transformasi spektral dalam pemodelan klasifikasi sangat mempengaruhi akurasi karena spektrum NIR yang tidak spesifik, tumpang tindih, luas dan adanya noise yang timbul dari instrumentasi perangkat selama pengumpulan data. Meskipun efek oversampling SMOTE menyebabkan lebih banyak kumpulan data dan peluang yang lebih beragam untuk prediksi yang salah antar kelas, itu telah berhasil meningkatkan akurasi klasifikasi dibandingkan tanpa pra-pemrosesan. Teknik oversampling SMOTE dapat digunakan ketika terjadi ketidakseimbangan kelas data dalam klasifikasi.

DAFTAR PUSTAKA

[1] P. Osinenko et al., "Application of non-destructive sensors and big data analysis to predict physiological storage disorders and fruit firmness in 'Braeburn' apples," Comput. Electron. Agric., vol. 183, p. 106015, Apr. 2021, doi: 10.1016/j.compag.2021.106015.

[2] M. Arunkumar, A. Rajendran, S. Gunasri, M. Kowsalya, and C. K. Krithika, "Non-destructive fruit maturity detection methodology - A review," Mater. Today Proc., Mar. 2021, doi: 10.1016/j.matpr.2020.12.1094.

[3] B. Nugraha, P. Verboven, S. Janssen, Z. Wang, and B. M. Nicolaï, "Non-destructive porosity mapping of fruit and vegetables using X-ray CT," Postharvest Biol. Technol., vol. 150, pp. 80–88, Apr. 2019, doi: 10.1016/j.postharvbio.2018.12.016.

[4] B. M. Nicolaï et al., "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," Postharvest Biol. Technol., vol. 46, no. 2, pp. 99–118, Nov. 2007, doi: 10.1016/j.postharvbio.2007.06.024.

[5] A. Ibrahim, N. El-Bialee, M. Saad, and E. Romano, "Non-Destructive Quality Inspection of Potato Tubers Using Automated Vision System," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2419, Dec. 2020, doi: 10.18517/ijaseit.10.6.13079.

[6] H. Lee, M. S. Kim, H.-S. Lim, E. Park, W.-H. Lee, and B.-K. Cho, "Detection of cucumber green mottle mosaic virus-infected watermelon seeds using a near-infrared (NIR) hyperspectral imaging system: Application to seeds of the 'Sambok Honey' cultivar," Biosyst. Eng., vol. 148, pp. 138–147, Aug. 2016, doi: 10.1016/j.biosystemseng.2016.05.014.

[7] J. M. S. Netto, F. A. Honorato, P. M. Azoubel, L. E. Kurozawa, and D. F. Barbin, "Evaluation of melon drying using hyperspectral imaging technique in the near infrared region," LWT, vol. 143, p. 111092, May 2021, doi: 10.1016/j.lwt.2021.111092.

[8] F. D. Anggraeni, N. Khuriyati, M. A. F. Falah, H. Nishina, K. Takayama, and N. Takahashi, "Non-destructive Measurement of Lycopene Content in High Soluble Solids Stored Tomato (Solanum Lycopersicum Mill. cv Rinka 409)," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2567, Dec. 2020, doi: 10.18517/ijaseit.10.6.9478.

[9] R. Hayati, A. A. Munawar, and F. Fachruddin, "Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango," Data Br., vol. 30, p. 105571, Jun. 2020, doi: 10.1016/j.dib.2020.105571.

[10] P. Mishra, E. Woltering, and N. El Harchioui, "Improved prediction of 'Kent' mango firmness during ripening by near-infrared spectroscopy

supported by interval partial least square regression," Infrared Phys. Technol., vol. 110, p. 103459, Nov. 2020, doi: 10.1016/j.infrared.2020.103459.

[11] G. Ren, Y. Liu, J. Ning, and Z. Zhang, "Assessing black tea quality based on visible–near infrared spectra and kernel-based methods," J. Food Compos. Anal., vol. 98, p. 103810, May 2021, doi: 10.1016/j.jfca.2021.103810.

[12] J. Li, H. Zhang, B. Zhan, Y. Zhang, R. Li, and J. Li, "Nondestructive firmness measurement of the multiple cultivars of pears by Vis-NIR spectroscopy coupled with multivariate calibration analysis and MC-UVE-SPA method," Infrared Phys. Technol., vol. 104, p. 103154, Jan. 2020, doi: 10.1016/j.infrared.2019.103154.

[13] C. Liu, S. X. Yang, X. Li, L. Xu, and L. Deng, "Noise level penalizing robust Gaussian process regression for NIR spectroscopy quantitative analysis," Chemom. Intell. Lab. Syst., vol. 201, p. 104014, Jun. 2020, doi: 10.1016/j.chemolab.2020.104014.

[14] J. Torniainen, I. O. Afara, M. Prakash, J. K. Sarin, L. Stenroth, and J. Töyräs, "Open-source python module for automated preprocessing of near infrared spectroscopic data," Anal. Chim. Acta, vol. 1108, pp. 1–9, Apr. 2020, doi: 10.1016/j.aca.2020.02.030.

[15] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 04, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.

[16] N. S. Sani, M. Abdul Rahman, A. Abu Bakar, S. Sahran, and H. Mohd Sarim, "Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4–2, p. 1698, Sep. 2018, doi: 10.18517/ijaseit.8.4-2.6829.

[17] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential targeting: A continual learning approach for data imbalance in text classification," Expert Syst. Appl., vol. 179, p. 115067, Oct. 2021, doi: 10.1016/j.eswa.2021.115067.

[18] A. A. Munawar, Kusumiyati, and D. Wahyuni, "Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits," Data Br., vol. 27, p. 104789, Dec. 2019, doi: 10.1016/j.dib.2019.104789.

[19] Åsmund Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," TrAC Trends Anal. Chem., vol. 28, no. 10, pp. 1201–1222, Nov. 2009, doi: 10.1016/j.trac.2009.07.007.

[20] D. D. Silalahi, H. Midi, J. Arasan, M. S. Mustafa, and J.-P. Caliman, "Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data," *Vib. Spectrosc.*, vol. 97, pp. 55–65, Jul. 2018, doi: 10.1016/j.vibspec.2018.05.002.

[21] D. Syvilay et al., "Evaluation of the standard normal variate method for Laser-Induced Breakdown Spectroscopy data treatment applied to the

discrimination of painting layers," Spectrochim. Acta Part B At. Spectrosc., vol. 114, pp. 38–45, Dec. 2015, doi: 10.1016/j.sab.2015.09.022.
[22] Q. Guo, W. Wu, and D. . Massart, "The robust normal variate transform for pattern recognition with near-infrared data," Anal. Chim. Acta, vol. 382, no. 1–2, pp. 87–103, Feb. 1999, doi: 10.1016/S0003-2670(98)00737-5.
[23] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," Inf. Sci. (Ny)., vol. 512, pp. 1214–1233, Feb. 2020, doi: 10.1016/j.ins.2019.10.048.
[24] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," J. King Saud Univ. - Comput. Inf. Sci., Feb. 2021, doi: 10.1016/j.jksuci.2021.01.014.
[25] M. Sinambela, M. Situmorang, K. Tarigan, S. Humaidi, and M. Sirait, "Waveforms Classification of Northern Sumatera Earthquakes for New Mini Region Stations Using Support Vector Machine," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 11, no. 2, p. 489, Apr. 2021, doi: 10.18517/ijaseit.11.2.12503.
[26] V. Vapnik and R. Izmailov, "Reinforced SVM method and memorization mechanisms," Pattern Recognit., vol. 119, p. 108018, Nov. 2021, doi: 10.1016/j.patcog.2021.108018.
[27] I. Hasanah, E. Purwanti, and P. Widiyanti, "Design and Implementation of an Early Screening Application for Dengue Fever Patients Using Android-Based Decision Tree C4.5 Method," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2237, Dec. 2020, doi: 10.18517/ijaseit.10.6.5771.
[28] J. P. Pinder, "Decision Trees," in Introduction to Business Analytics using Simulation, Elsevier, 2017, pp. 47–69.
[29] P. Arumugam and P. Jose, "Efficient Decision Tree Based Data Selection and Support Vector Machine Classification," Mater. Today Proc., vol. 5, no. 1, pp. 1679–1685, 2018, doi: 10.1016/j.matpr.2017.11.263.
[30] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf. Sci. (Ny)., vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 3
ISSN: 2088-5334

Effects of Oversampling Smote and Spectral Transformations in the Classification of Mango Cultivars Using Near-Infrared Spectroscopy

Ali Khumaidi [a,*], Ridwan Raafi'udin [b]

[a] *Faculty of Engineering, Universitas Krisnadwipayana, Jakarta, 17411, Indonesia*
[b] *Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta, 12450, Indonesia*
*Corresponding author: *alikhumaidi@unkris.ac.id*

***Abstract*— Near-Infrared spectroscopy (NIR) is a non-destructive analytical technique that can provide chemical and structural information on samples in a speedy and accurate time. NIR has a wavelength of 750-2500 nm. However, the absorbance bands of the NIR spectrum are often broad, non-specific, and overlapping. NIR spectrum analysis requires a multivariate method which is very subjective to noise arising from instrumentation. There is no standard protocol in modeling for classification and prediction using NIR spectra. Several models have been developed with and without pre-processing techniques. The SMOTE technique can improve the model to predict all class responses accurately. This research contributes to creating a multiclass classification model for grouping mango cultivars by finding the best pre-processing technique and using SMOTE oversampling. The results of the four test scenarios on the model's performance built using the Support Vector Machine (SVM) that the best model is obtained using spectral transformations with LSNV and CLIP operations with 100% accuracy, precision, and recall values. The Decision Tree (DT) has the performance results in 100% model was obtained by using spectral transformation with LSNV, CLIP and SAVGOL operations with parameters {'deriv_order': 0,1, 2, 'filter_win': 11, 13, 'poly_order': 3}. Using SMOTE has better accuracy than without pre-processing, with an accuracy of 92% on SVM and 94% on DT. In comparison, the combination of SMOTE and Spectral Transformation gives classification results for SVM and DT with the same accuracy of 96%, better than using SMOTE only.**

***Keywords*— Classification; cultivar mango; near-infrared; spectral transformation; oversampling SMOTE.**

Manuscript received 9 Aug. 2021; revised 13 Sep. 2021; accepted 13 Dec. 2021. Date of publication 30 Jun. 2022.

I. INTRODUCTION

Technological advances have encouraged innovation in technology development to determine fruit's characteristics and quality non-destructively [1]. This technology can quantify fruit quality based on the classification of external and internal factors that the human senses cannot detect. The current quality determination method is mostly carried out destructively, which requires time, effort, cost, and there is a bias factor due to human subjectivity [2]. So, measuring quality and destructive detection is not suitable to be applied in the industry. The non-destructive quality measurement method is more effective based on the correlation between the physical properties of the fruit associated with the fruit quality factor. The use of non-destructive equipment produces more consistent results than human labor, thereby minimizing the chance of errors in deciding fruit quality [3].

The application of Near Infrared Spectroscopy (NIR) by utilizing infrared rays is not new, and NIR spectroscopy was developed in 1950 in the industrial field, which focused on analyzing the chemical content of materials [4]. NIR spectroscopy has begun to be widely used to analyze moisture, protein, and fat content in agricultural and food products. NIR spectroscopy is a non-destructive analytical technique capable of providing chemical and structural information on certain samples in a swift time (less than 1 minute). NIR has a wavelength of 750-2500 nm, the target sample is illuminated with light, and the reflected light or backscatter is measured with a spectrometer. Horticultural products can also use this NIR method in grading, sorting, internal quality, and determining harvest time [5]. Determination of the quality of horticultural products can be done non-destructively by using NIR spectroscopy that has been applied to watermelon [6] and melon [7], using spectral for detection Lycopene Content in Tomato [8], detection of mango quality [9], [10].

There is no standard protocol in modeling for classification and prediction using the NIR spectrum. Several models have been developed with and without pre-processing techniques. Support Vector Machine (SVM) is the most widely used

algorithm in prediction models, classification, and regression for fruit quality detection and fairly good accuracy. Detection of black tea quality using standard normal variate (SNV) spectral transformation and Savitzky Golay combination with first derivative and SVM algorithm [11]. Pear hardness measurement using SNV spectral transformation and first-derivative with SVM [12].

Spectral transformation technique can improve model performance [13]. These techniques include Smoothing, Scatter Correction, Trimming, Clipping, Resampling, and Derivatives. The order of pre-processing operations applied can affect the performance of the model [14]. Smoothing aims to smooth the spectral and help remove noise. Scatter correction aims to counteract the effects of particle size. Trimming allows the extraction of continuous and non-continuous wavelength regions from full spectral data. Clipping aims to remove or replace data points with values that exceed a user-defined threshold. Resampling processes a new spectral resolution using the Fourier method, combining the obtained spectral with several devices having different spectral resolutions.

A Balance class is a condition of unbalanced distribution between classes in a dataset, where one class has a very large amount of data (majority class) compared to the other class (minority class) [15]. The large difference in the amount of data between classes can result in the classification model often not being able to predict the minority class correctly so that many test data that should be in the minority class is predicted wrongly by the classification model [16]. One of the methods used to overcome the imbalanced class problem is sampling. The sampling method modifies the distribution of data between the majority and minority classes in the training dataset to balance the amount of data for each class [17]. One of the sampling methods that is often used is the Synthetic Minority Oversampling Technique (SMOTE).

The aims of this study are (1) to investigate the effect of the performance of the spectral transformation method and the most optimal operation on the dataset; (2) to find out the effect of data balance on the model; (3) to explore the optimal machine learning classification model based on the application of spectral transformation and data balance.

II. Materials and Method

The research stages include dataset preparation, pre-processing or spectral transformation to obtain the most optimal data to support the model, SMOTE to deal with class imbalances, splitting the data into training data and testing data, developing models, and evaluating the model to find out the best model. The modeling compared the support vector machine (SVM) and the decision tree algorithm. To determine the performance of the classification model by evaluating the model with four scenarios are (1) without any treatment; (2) apply SMOTE; (3) apply spectral transformation; (4) apply SMOTE and spectral transformation. The relationship between stages can be seen in Figure 1.

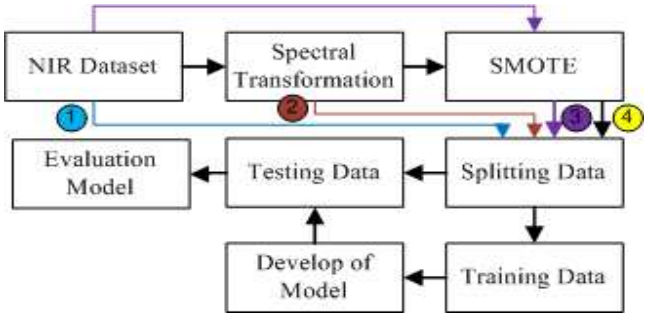


Fig. 1 Research Method

A. Dataset

The dataset used is derived from research results [18]. A total of 186 whole mango samples from 4 different cultivars (Cengkir, Kweni, Kent, and Palmer) were taken using near-infrared spectral obtained in the form of absorbance with wavelengths from 1000 to 2500 nm. The number of samples of cultivar Cengkir 18, Kweni 29, Kent 85, and Palmer 54. This dataset is accessible https://data.mendeley.com/datasets/b9d6s7hr33/1.

B. Pre-processing

The purpose of data pre-processing is to reduce the physical properties in the spectrum to reduce the variability caused by light scattering nonlinearity and improve the model to be used [19]. The spectral transformations used are resampling, clipping, smoothing, derivative, and scatter correction in smoothing using Savitsky Golay filtering with parameters: filter windows (7, 11, 13), order of the polynomial = 3, and order of the derivative (0, 1, 2). Scatter correction uses several operations are multiplicative scatter correction (MSC), standard normal variate (SNV), robust normal variate (RNV), normalization, baseline, detrend, localized version of SNV (LSNV), and the extended version of MSC (EMSC). The use of the spectral transformation method in more detail can be seen in Table 1.

The MSC method is to draw spectral sample points to approach the reference spectrum by utilizing the results of estimating simple linear regression parameters and can eliminate the variation between spectra by correcting the scatter position of each intensity value of each replication to the scattering position of the average intensity of the entire replication [20]. The SNV method removes the scattering effect from the spectrum by centering and adjusting the scale of each spectrum [21]. The RNV method is more suitable for data with much noise by using the concept of correction based on the median value and the interval between quartiles [22]. The spectral normalization method uses a certain range of values and usually applies to Euclidean. In principle, the baseline method uses the average of the central values of the spectral. The concept of the LSNV method is SNV with the principle of division operation on the spectral window. The EMSC method is almost the same as the MSC, but in EMCS, it considers linear and quadratic corrections. All methods and operations of this spectral transformation were compared to the model to obtain the most optimal accuracy—the use of spectral transformation techniques is one factor in improving the model's performance.

TABLE I
SPECTRAL TRANSFORM METHODS

Method	Operation	Parameter	Value
Resampling	RESAMPLE	Rasio	0.8
Clipping	CLIP	Threshold	1e4
		substitute	None
Smoothing	SAVGOL	filter_win	7, 11, 13
		poly_order	3
		deriv_order	0, 1, 2
Scatter Correction	MSC		
	SNV		
	RNV	iqr	75-25, 90-10
	NORML		
	BASELINE		
	LSNV		
	EMSC		

C. SMOTE

This study identified that the dataset used has class imbalance problems, so an over-sampling method is needed to overcome the imbalanced class problem. The method that can be used is SMOTE. SMOTE is an over-sampling method in which the data in the minority class is reproduced using synthetic data derived from data replication in the minority class. Over-sampling in SMOTE takes an instance of the minority class and then looks for the k-nearest neighbor of each instance, then generates a synthetic instance instead of replicating the minority class instance; therefore, it can avoid the problem of excessive overfitting [23]. The algorithm that works on the first SMOTE differentiated the vectors of the features in the minority class and the nearest neighbor values from the minority class and then multiplied that value by a random number between 0 to 1. Next, the calculation results are added to the feature vector so that the vector value results are obtained from the new one [24].

The proposed model was validated by two experimental scenarios that were carried out, namely using the SVM and Decision Tree algorithm approaches, and each was used for modeling without considering class imbalance, and secondly, SMOTE oversampling was carried out to increase the number of datasets in order to achieve a balanced dataset.

D. Modeling

The modeling developed in this research is classification. The developing model group was based on NIR spectral in 4 classes of mango cultivar. Classification is a multivariate technique for separating different sets of objects and allocating new objects into predefined groups. A good classification method resulted in less misclassification. It is necessary to use the right method to perform the classification accuracy. Support Vector Machine (SVM) is one method that can perform classification. SVM is a technique for finding hyperplanes that can separate two data sets from two different classes [25]. SVM has advantages, including determining the distance using a support vector so that the computational process becomes fast. The learning process in SVM aims to obtain a hypothesis in the form of the best dividing field that minimizes empirical risk, namely the average error in the training data and provides good generalization. Generalization is the ability of a hypothesis to be able to classify data that is not contained in the training data correctly. The principle of SVM is a linear classifier and then redeveloped so that it can work on non-linear problems using the kernel trick method, which looks for a hyperplane by transforming the dataset into a vector space with larger dimensions (feature space) using a kernel function which then was classified and performed on the feature space. Determination of the kernel function used affected the classification results [26].

The decision tree is one of the most popular classification methods because it is easy for humans to interpret. The decision tree is a predictive model using a tree structure or hierarchical structure [27]. The concept of a decision tree is to convert data into a decision tree and decision rules. A Decision Tree is used to study the classification and prediction of patterns from data and describe the relationship of the attribute variable x and the target variable y in the form of a tree. The decision tree resembles a flowchart where each internal node (a node that is not a leaf or the outermost node) is a test of attribute variables; each branch is the test result, while the outermost node, namely the leaf, is the. The main benefit of using a decision tree is its ability to break down complex decision-making processes into simpler ones to interpret solutions to problems [28] better.

Decision trees are also useful for exploring data finding hidden relationships between many potential input variables and a target variable. Decision trees combine data exploration and modeling, so they are great as a first step in the modeling process, even when used as the final model of some other technique. Another advantage of this method is eliminating unnecessary calculations or data. The existing samples are usually only tested based on certain criteria or classes [29].

E. Model evaluation

A confusion matrix was used to measure the classification model's performance using a. The confusion matrix, also known as the error matrix, provides information on the comparison of the classification results performed by the model with the actual classification results [30]. There are four terms representing the results of the classification process in the confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP is positive data that is predicted to be correct, TN is negative data that is predicted to be correct, FP is negative data but is predicted to be positive data, and FN is positive data but is predicted to be negative data.

The confusion matrix can calculate various performance metrics to measure the performance of the model that has been made, some of which are often used, namely accuracy, precision, and recall. Accuracy describes how accurately the model can classify correctly. Precision describes the level of accuracy between the requested data and the prediction results provided by the model. Recall describes the success of the model in retrieving information.

III. RESULTS AND DISCUSSION

The NIR spectral dataset of 186 samples with a wavelength of 1000 to 2500 nm can be seen in Figure 4a. Several absorption peaks can be found from the original transmittance spectrum. The mango dataset consists of 4 cultivars, so the multiclass classification method was used in this study. The following are the results of measuring the model performance from the four scenarios with the distribution of training data and testing data of 70 and 30.

A. Modelling without pre-processing

The classification of NIR spectrum data processing using SVM without pre-processing has resulted in a fairly good accuracy of 90%, class classification errors of 6 Kent cultivars, which are actually Palmer cultivars. The results of the classification using DT without pre-processing resulted in higher accuracy than SVM, which was 94%, with an incorrect guess of 4 Cengkir cultivars, which should have been Palmer cultivar. The confusion matrix results in more detail can be seen in Table 2.

TABLE II
CONFUSION MATRIX RESULTS WITHOUT PRE-PROCESSING AND USING SPECTRAL TRANSFORMATION

Algorithm	Class	Prediction				Without Pre-Processing			Prediction				Using Spectral Transformation		
		C	K	Kw	P	Accuracy	Precision	Recall	C	K	Kw	P	Accuracy	Precision	Recall
SVM	Cengkir (C)	6	0	0	0	0,90	0,96	0,92	5	0	0	0	1	1	1
	Kent (K)	0	28	0	0				0	26	0	0			
	Kweni (Kw)	0	0	10	0				0	0	13	0			
	Palmer (P)	0	6	0	12				0	0	0	18			
DT	Cengkir (C)	6	0	0	0	0,94	0,9	0,94	5	0	0	0	1	1	1
	Kent (K)	0	28	0	0				0	26	0	0			
	Kweni (Kw)	0	0	10	0				0	0	13	0			
	Palmer (P)	4	0	0	14				0	0	0	18			

B. Modeling with oversampling SMOTE

By applying SMOTE oversampling, the amount of data between classes is balanced with the number of 85 mangoes in each class. Cengkir cultivars, which originally had 18 data, Kweni had 29 data, and Palmer originally had 54 data, were equated with the total data of Kent, which was 85 data. The results of the classification of NIR spectrum data processing using SVM with the application of SMOTE oversampling improved the classification accuracy of 92% compared to those without pre-processing, the class grouping error of 1 Palmer cultivar, which was actually a Cengkir cultivar, 1 Cengkir cultivar which was actually a Kent cultivar and the supposed Palmer cultivar was predicted to be 2 cultivars Cup and 5 Kent.

The results of the classification using DT with the application of SMOTE oversampling improved the classification accuracy of 94% compared to those without pre-processing, class classification errors were 1 Palmer cultivar and 1 Kent cultivar, which was Cengkir cultivar, 3 Palmer cultivar which was Kent cultivar and 1 Cengkir cultivar which was supposed to be Palmer. The confusion matrix results in more detail can be seen in Table 3.

TABLE III
CONFUSION MATRIX RESULTS USING OVERSAMPLING SMOTE AND SPECTRAL TRANSFORMATION

Algorithm	Class	Prediction				Using Oversampling SMOTE			Prediction				Using Oversampling SMOTE and Spectral Transformation		
		C	K	Kw	P	Accuracy	Precision	Recall	C	K	Kw	P	Accuracy	Precision	Recall
SVM	Cengkir (C)	27	0	0	1	0,92	0,92	0,92	25	3	0	0	0,96	0,97	0,96
	Kent (K)	1	27	0	0				0	28	0	0			
	Kweni (Kw)	0	0	29	0				0	0	29	0			
	Palmer (P)	2	5	0	21				0	1	0	27			
DT	Cengkir (C)	26	1	0	1	0,94	0,94	0,94	27	0	0	1	0,96	0,96	0,96
	Kent (K)	0	25	0	3				0	27	0	1			
	Kweni (Kw)	0	0	29	0				0	0	29	0			
	Palmer (P)	1	0	0	22				0	3	0	25			

C. Modeling with spectral transformation

The NIR spectrum data processing classification results using SVM with the application of spectral transformation methods, namely Smoothing, Scatter Correction, Clipping, Resampling, and Derivatives and their combinations, operations, and parameters of these methods, can be seen in Table 1. Using the Nippy library with Python, the results obtained classification with an accuracy value of 100%, where the use of the Clipping and Scatter Correction method with LSNV operation provides the most optimal results for SVM. The results of the DT classification with spectral transformation using the Clipping and Scatter Correction method with LSNV operation also produce 100% accuracy without class prediction errors using the Clipping, Scatter Correction, and Smoothing methods. The confusion matrix results in more detail can be seen in Table 2.

The classification accuracy calculation process using SVM, and DT combined with 5 spectral transformation methods gave varying results. Some even had worse accuracy with accuracy values without pre-processing. Therefore, it is very important to adapt spectral transformation methods and operations to the data. Experiments were carried out on all spectral transformation operations to obtain the most optimal accuracy results in the classification using SVM. First, testing is carried out on each operation and its parameters. The results of the spectral transformation operation with the best accuracy are then combined with other operations. The results of 2 combinations of spectral transformation operations are then combined with other operations. And so on until the most optimal accuracy value is obtained.

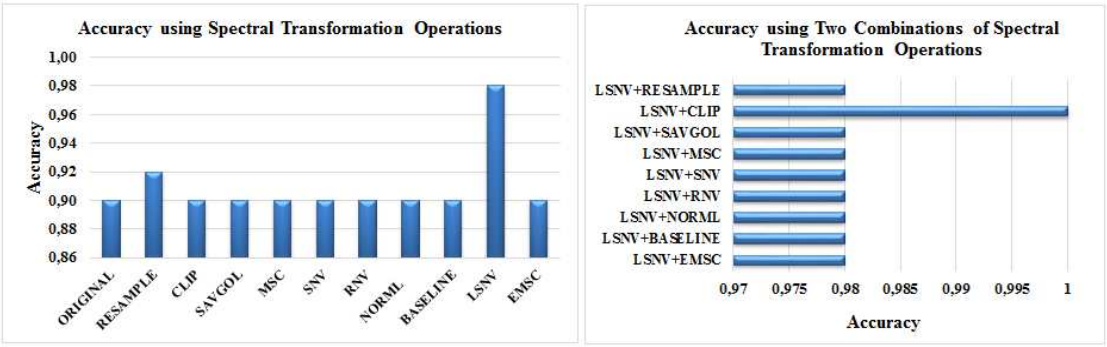


Fig. 2 Comparison of Combination of Spectral Transformation toward Accuracy on SVM

Figure 2a describes the results of applying one spectral transformation operation on SVM. Spectral transformation operations that produce higher accuracy values than those without pre-processing are RESAMPLE with 92% accuracy and LSNV with 98% accuracy. The accuracy value is 90%, the same as without pre-processing for other operations. The LSNV operation as the spectral transformation operation with the highest value is then combined with other operations. Accuracy results from 2 combinations of spectral transformation operations on SVM have reached 100% accuracy using LSNV and CLIP operations with threshold=10000. While the other combinations with a maximum accuracy value of 98%. By obtaining the most optimal spectral transformation operation, the next combination achieved optimal again, at least by using two combinations of spectral transformation operations, namely LSNV and CLIP. The accuracy results of 2 combinations of spectral transformation operations on SVM can be seen in Figure 2b.

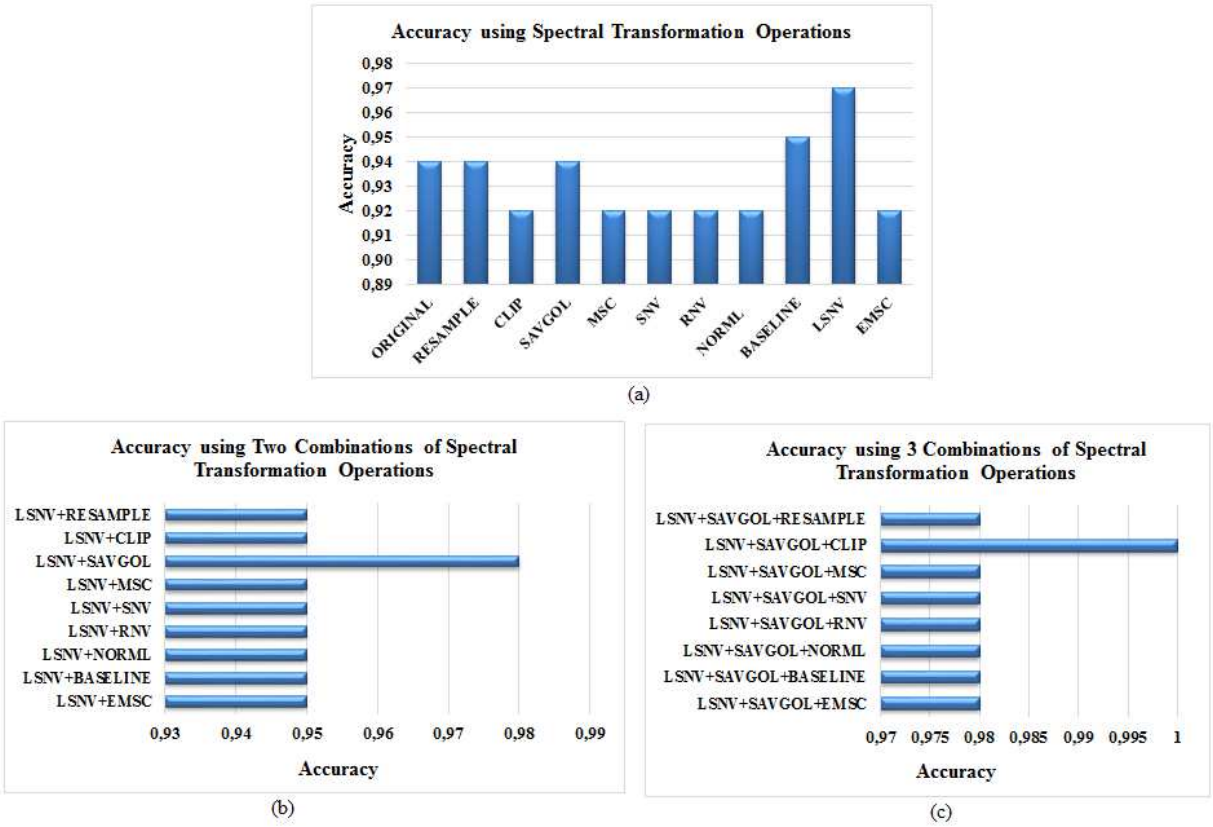


Fig. 3 Comparison of Combination of Spectral Transformation toward Accuracy on DT

Figure 3a describes the results of applying one spectral transformation operation to DT. Spectral transformation operations that produce higher accuracy values than those without pre-processing are BASELINE with 95% accuracy and LSNV with 97% accuracy. For other spectral transformation operations, the accuracy value is lower than the accuracy value without pre-processing, which is 92%. The LSNV operation as the spectral transformation operation with the highest value is then combined with other operations. Accuracy results from 2 combinations of spectral transformation operations on DT that have increased, namely LSNV and SAVGOL operations with parameters {'deriv_order': 2, 'filter_win': 11, 'poly_order': 3} with an accuracy value of 98%. While the combination of LSNV with

other spectral transformation operations, the maximum classification accuracy value is 95%. Accuracy results with two combinations of spectral transformation operations on DT can be seen in Figure 3b. Next is to perform three combinations of spectral transformation operations using LSNV and SAVGOL as operations with the best accuracy on two combinations of spectral transformation operations combined with other spectral transformation operations. The combination of 3 spectral transformation operations has achieved 100% accuracy by using LSNV, CLIP with threshold=10000 and SAVGOL with parameters {'deriv_order': 2, 'filter_win': 11, 'poly_order': 3}, {'deriv_order' : 1, 'filter_win': 11, 'poly_order': 3}, {'deriv_order': 2, 'filter_win': 13, 'poly_order': 3}, and {'deriv_order': 0, 'filter_win': 13, 'poly_order': 3}. Three other combinations of spectral transformations with a maximum accuracy value of 98%. By having obtained the most optimal spectral transformation operation, the next combination achieved optimal again, at least by using three combinations of spectral transformation operations, namely LSNV, CLIP, and SAVGOL. For the accuracy results of 3 combinations of spectral transformation operations on DT, it can be seen in Figure 3c.

D. Modeling with Oversampling SMOTE and Spectral Transformation

The results of the classification of NIR spectrum data processing using SVM with the application of SMOTE oversampling and spectral transformation, where there is a balance of the amount of data in the class and the use of Clipping and Scatter Correction methods with LSNV operations, produces a classification accuracy value of 96%. The class classification error is 3 Kent cultivar, Cengkir cultivar, and 1 Kent cultivar, Palmer cultivar. The classification results using DT also produce a classification accuracy value of 96%. Class classification errors are 1 Palmer cultivar, Cengkir cultivar, 1 Palmer cultivar, which is Kent cultivar, and 3 Kent cultivar, which should be Palmer. The confusion matrix results in more detail can be seen in Table 3. The performance of classification accuracy with SVM and DT has the same value, and both are more or less good than using spectral transformation alone. However, the accuracy value is higher than without pre-processing and the use of oversampling SMOTE.

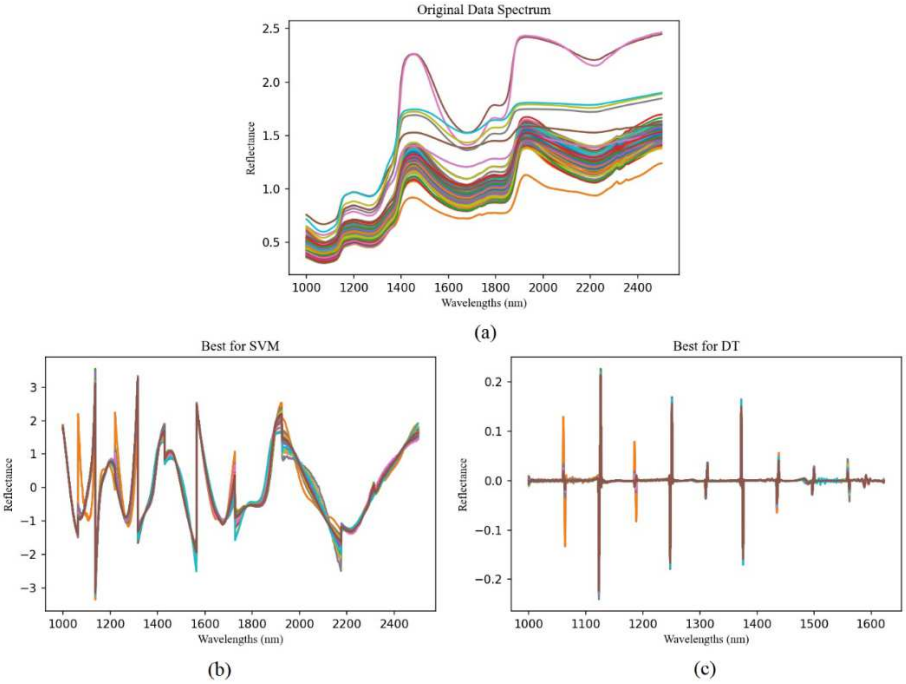


Fig. 4 Comparison of Spectral Curves after using Spectral Transformation

The spectral transformation has a significant influence in increasing the accuracy of the classification model. In the four test scenarios, the classification results with appropriate spectral transformation methods and operations achieved 100% accuracy. The best spectral transformation results with Clipping and Scatter Correction methods with LSNV operation for modeling with SVM can be seen in Figure 4b, and for DT modeling can be seen in Figure 4c. The spectrum peaks are clearly visible in both images, making it easier for the model to classify.

Table 3 compares four SVM and DT algorithms scenarios with performance measurements based on accuracy, precision, and recall. With the model built by SVM and DT, the best performance with the application of spectral transformation, the application of SMOTE oversampling and spectral transformation, the application of SMOTE oversampling, and finally without pre-processing.

IV. CONCLUSION

Spectral treatment is needed to improve the accuracy of the classification model of 4 classes of mango cultivars based on NIR spectroscopy. Treatment with SMOTE oversampling can improve classification accuracy in SVM and DT algorithms. Treatment with spectral transformation using five combinations of methods obtained an optimal classification

accuracy of 100% on SVM using clipping and scatter correction methods, namely LSNV, while in DT using clipping, scatter correction, and smoothing techniques. When the application of the spectral transformation is optimal with SMOTE oversampling, the classification accuracy is not better. Spectral transformation treatment in classification modeling significantly affects accuracy due to the non-specific, overlapping, broad nature of the NIR spectrum and the presence of noise arising from the instrumentation of the device during data collection. Although the SMOTE oversampling effect causes more data sets and more diverse opportunities for incorrect predictions between classes, it has increased classification accuracy compared to no pre-processing. The SMOTE oversampling technique can be used when data classes are imbalanced in the classification.

ACKNOWLEDGMENT

The authors are grateful to Mr. Munawar and his colleagues who provided research data and the dean of the engineering faculty at Krisnadwipayana University, who has supported research and publications.

REFERENCES

[1] P. Osinenko et al., "Application of non-destructive sensors and big data analysis to predict physiological storage disorders and fruit firmness in 'Braeburn' apples," Comput. Electron. Agric., vol. 183, p. 106015, Apr. 2021, doi: 10.1016/j.compag.2021.106015.

[2] M. Arunkumar, A. Rajendran, S. Gunasri, M. Kowsalya, and C. K. Krithika, "Non-destructive fruit maturity detection methodology - A review," Mater. Today Proc., Mar. 2021, doi: 10.1016/j.matpr.2020.12.1094.

[3] B. Nugraha, P. Verboven, S. Janssen, Z. Wang, and B. M. Nicolaï, "Non-destructive porosity mapping of fruit and vegetables using X-ray CT," Postharvest Biol. Technol., vol. 150, pp. 80–88, Apr. 2019, doi: 10.1016/j.postharvbio.2018.12.016.

[4] B. M. Nicolaï et al., "Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," Postharvest Biol. Technol., vol. 46, no. 2, pp. 99–118, Nov. 2007, doi: 10.1016/j.postharvbio.2007.06.024.

[5] A. Ibrahim, N. El-Bialee, M. Saad, and E. Romano, "Non-Destructive Quality Inspection of Potato Tubers Using Automated Vision System," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2419, Dec. 2020, doi: 10.18517/ijaseit.10.6.13079.

[6] C. Ding, D. Wang, Z. Feng, W. Li, and D. Cui, "Integration of vibration and optical techniques for watermelon firmness assessment," *Comput. Electron. Agric.*, vol. 187, p. 106307, Aug. 2021, doi: 10.1016/j.compag.2021.106307.

[7] J. M. S. Netto, F. A. Honorato, P. M. Azoubel, L. E. Kurozawa, and D. F. Barbin, "Evaluation of melon drying using hyperspectral imaging technique in the near infrared region," LWT, vol. 143, p. 111092, May 2021, doi: 10.1016/j.lwt.2021.111092.

[8] F. D. Anggraeni, N. Khuriyati, M. A. F. Falah, H. Nishina, K. Takayama, and N. Takahashi, "Non-destructive Measurement of Lycopene Content in High Soluble Solids Stored Tomato (Solanum Lycopersicum Mill. cv Rinka 409)," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2567, Dec. 2020, doi: 10.18517/ijaseit.10.6.9478.

[9] R. Hayati, A. A. Munawar, and F. Fachruddin, "Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango," Data Br., vol. 30, p. 105571, Jun. 2020, doi: 10.1016/j.dib.2020.105571.

[10] P. Mishra, E. Woltering, and N. El Harchioui, "Improved prediction of 'Kent' mango firmness during ripening by near-infrared spectroscopy supported by interval partial least square regression," Infrared Phys. Technol., vol. 110, p. 103459, Nov. 2020, doi: 10.1016/j.infrared.2020.103459.

[11] G. Ren, Y. Liu, J. Ning, and Z. Zhang, "Assessing black tea quality based on visible–near infrared spectra and kernel-based methods," J. Food Compos. Anal., vol. 98, p. 103810, May 2021, doi: 10.1016/j.jfca.2021.103810.

[12] J. Li, H. Zhang, B. Zhan, Y. Zhang, R. Li, and J. Li, "Non-destructive firmness measurement of the multiple cultivars of pears by Vis-NIR spectroscopy coupled with multivariate calibration analysis and MC-UVE-SPA method," Infrared Phys. Technol., vol. 104, p. 103154, Jan. 2020, doi: 10.1016/j.infrared.2019.103154.

[13] C. Liu, S. X. Yang, X. Li, L. Xu, and L. Deng, "Noise level penalizing robust Gaussian process regression for NIR spectroscopy quantitative analysis," Chemom. Intell. Lab. Syst., vol. 201, p. 104014, Jun. 2020, doi: 10.1016/j.chemolab.2020.104014.

[14] J. Torniainen, I. O. Afara, M. Prakash, J. K. Sarin, L. Stenroth, and J. Töyräs, "Open-source python module for automated pre-processing of near infrared spectroscopic data," Anal. Chim. Acta, vol. 1108, pp. 1–9, Apr. 2020, doi: 10.1016/j.aca.2020.02.030.

[15] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 04, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.

[16] N. S. Sani, M. Abdul Rahman, A. Abu Bakar, S. Sahran, and H. Mohd Sarim, "Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4–2, p. 1698, Sep. 2018, doi: 10.18517/ijaseit.8.4-2.6829.

[17] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential targeting: A continual learning approach for data imbalance in text classification," Expert Syst. Appl., vol. 179, p. 115067, Oct. 2021, doi: 10.1016/j.eswa.2021.115067.

[18] A. A. Munawar, Kusumiyati, and D. Wahyuni, "Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits," Data Br., vol. 27, p. 104789, Dec. 2019, doi: 10.1016/j.dib.2019.104789.

[19] Mishra, J. M. Roger, D. N. Rutledge, and E. Woltering, "SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials," Postharvest Biol. Technol., vol. 168, p. 111271, Oct. 2020, doi: 10.1016/j.postharvbio.2020.111271..

[20] D. D. Silalahi, H. Midi, J. Arasan, M. S. Mustafa, and J.-P. Caliman, "Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data," *Vib. Spectrosc.*, vol. 97, pp. 55–65, Jul. 2018, doi: 10.1016/j.vibspec.2018.05.002.

[21] B. Lu et al., "Quantitative NIR spectroscopy determination of coco-peat substrate moisture content: Effect of particle size and non-uniformity," Infrared Phys. Technol., vol. 111, p. 103482, Dec. 2020, doi: 10.1016/j.infrared.2020.103482.

[22] Q. Guo, W. Wu, and D. . Massart, "The robust normal variate transform for pattern recognition with near-infrared data," Anal. Chim. Acta, vol. 382, no. 1–2, pp. 87–103, Feb. 1999, doi: 10.1016/S0003-2670(98)00737-5.

[23] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," Inf. Sci. (Ny)., vol. 512, pp. 1214–1233, Feb. 2020, doi: 10.1016/j.ins.2019.10.048.

[24] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," J. King Saud Univ. - Comput. Inf. Sci., Feb. 2021, doi: 10.1016/j.jksuci.2021.01.014.

[25] M. Sinambela, M. Situmorang, K. Tarigan, S. Humaidi, and M. Sirait, "Waveforms Classification of Northern Sumatera Earthquakes for New Mini Region Stations Using Support Vector Machine," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 11, no. 2, p. 489, Apr. 2021, doi: 10.18517/ijaseit.11.2.12503.

[26] V. Vapnik and R. Izmailov, "Reinforced SVM method and memorization mechanisms," Pattern Recognit., vol. 119, p. 108018, Nov. 2021, doi: 10.1016/j.patcog.2021.108018.

[27] I. Hasanah, E. Purwanti, and P. Widiyanti, "Design and Implementation of an Early Screening Application for Dengue Fever Patients Using Android-Based Decision Tree C4.5 Method," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2237, Dec. 2020, doi: 10.18517/ijaseit.10.6.5771.

[28] J. P. Pinder, "Decision Trees," in Introduction to Business Analytics using Simulation, Elsevier, 2017, pp. 47–69.

[29] P. Arumugam and P. Jose, "Efficient Decision Tree Based Data Selection and Support Vector Machine Classification," Mater. Today Proc., vol. 5, no. 1, pp. 1679–1685, 2018, doi: 10.1016/j.matpr.2017.11.263.

[30] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf. Sci. (Ny)., vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.