Makalah Kolokium

Implementasi Algoritma *Generative Adversarial Networks* (GAN) dalam Menangani *Imbalance Data* Citra Multispektral Lahan Sawah Untuk Klasifikasi Kesuburan Lahan

ANDI MUHAMMAD ALIFIKRI (G64190005)[1*], Dr. KARLISA PRIANDANA, S.T., M.Eng dan MEDRIA KUSUMA DEWI HARDHIENATA, S.Komp., Ph.D.

ABSTRAK

Kasus *imbalance data* sangat sering dijumpai ketika melakukan kegiatan pengumpulan dan pengambilan data. Proses pengambilan data pada lahan sawah kerap mengalami masalah *imbalance data* karena disebabkan oleh persebaran kondisi tanaman padi yang tidak merata pada lahan sawah tersebut. Distribusi data yang tidak seimbang akan sangat mempengaruhi kinerja machine learning dalam melatih model klasifikasi. Oleh karena itu, dibutuhkan sebuah metode untuk mengatasi masalah *imbalance data* ini. Dalam penelitian ini, masalah tersebut akan diatasi dengan menggunakan algoritma CTGAN *(Conditional Tabular Generative Adversarial Network)*. Hasil akhir yang diharapkan pada penelitian ini adalah membuat sebuah model *Generator* yang mampu memproduksi data sintetis untuk menambah data yang *imbalance* sehingga dapat memberikan hasil klasifikasi yang akurat.

Kata Kunci: CTGAN, *generator model*, *imbalance data*, klasifikasi, *machine learning*, tabular.

ABSTRACT

Cases of data imbalance are very often encountered when collecting data. The data collection process on paddy fields often experiences data imbalance problems due to the uneven distribution of rice crop conditions in the paddy fields. Unbalanced distribution of data will greatly affect the performance of machine learning in training classification models. Therefore, we need a method to overcome this data imbalance problem. In this study, this problem will be solved by using the CTGAN *(Conditional Tabular Generative Adversarial Network) algorithm. The expected result of this research is to create a Generator model that is able to generate synthetic data to balance the existing data in the dataset so that it can give an accurate classification result.*

Keywords: CTGAN, generator model, imbalance data, classification, machine learning, tabular.

[1]Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi S1 Ilmu Komputer, FMIPA-IPB; Surel: andimuhammadalifikri@apps.ipb.ac.id

PENDAHULUAN

Latar Belakang

Kasus *imbalance data* sangat sering dijumpai ketika peneliti melakukan kegiatan pengambilan data. Seperti pada penelitian yang dilakukan oleh Wang *et al* (2021), pada penelitian ini ingin memprediksi status ketersediaan nitrogen tanaman padi pada fase vegetatif dengan menggunakan UAV *multispectral*. Dalam studi tersebut didapati bahwa UAV *multispectral* cukup efisien ketika digunakan untuk memantau status nitrogen pada padi. Adanya perbedaan kadar nutrisi padi menyebabkan tingkat kesuburan tanaman padi berbeda (Wang *et al.* 2021). Salah Satu tantangannya yaitu ketika ingin dilakukan pengembangan model estimasi kesuburan lahan pada sawah berdasarkan data UAV tersebut, tingkat kesuburan tanaman yang berbeda dapat akan menghasilkan kecenderungan kepada salah satu kelas, dan mengakibatkan imbalance data (Ali *et al.* 2019)

Data yang tidak seimbang (*imbalance*) akan menjadi masalah ketika seorang peneliti ingin melatih model *machine learning.* Di mana setiap kelas data tidak memiliki jumlah yang sama maka akan membuat tingkat akurasi klasifikasi pada model menjadi tidak maksimal (Zhang *et al.* 2021). Distribusi kelas yang tidak seimbang akan sangat mempengaruhi kinerja *machine learning* dalam membuat model klasifikasi. Dampaknya adalah model akan menjadi bias dan terjadi *overfitting* terhadap *record* data yang jumlah kelasnya lebih banyak (Patel *et al.* 2020).

Ada beberapa pendekatan dalam mengatasi data yang tidak seimbang yaitu pendekatan pada tingkat data dan pendekatan pada tingkat algoritma. Pendekatan tingkat data menggunakan teknik *oversampling* dan undersampling. Disebut sebagai pendekatan tingkat data karena metode ini secara langsung memanipulasi data yang ada pada dataset demi menyeimbangkan sampel data, dengan cara mengurangi sampel data mayoritas ataupun menghapus kelas minoritas. Kekurangan dari teknik *oversampling* yaitu bisa saja terjadi *overfitting* ketika melatih model sedangkan untuk teknik undersampling dapat menyebabkan kehilangan beberapa sampel data yang penting (Rout *et al*. 2018). Metode yang kedua yaitu pendekatan tingkat algoritme dengan memodifikasi algoritme yang ada demi mengatasi bias akibat ketidakseimbangan data (Spelmen dan Porkodi 2018). Salah satu contohnya yaitu Algoritma *Random Forest.* Algoritma *Random Forest* efektif digunakan untuk mengatasi imbalance data dengan dengan menambah jumlah *classifier* yang mewakili kelas minoritas (Bader-El-Den *et al.* 2019). Lalu digunakan *Confusion Matrix* untuk mengevaluasi output dari *Random Forest*.

Berdasarkan latar belakang di atas, fokus utama dalam penelitian ini adalah ingin mengatasi masalah *imbalance data* dengan cara menambahkan sampel data pada dataset yang akan digunakan. Oleh karena itu, pendekatan yang akan digunakan adalah metode pendekatan tingkat data (*Data-Level Approaches*) dengan teknik *oversampling*. Teknik *oversampling* yang populer digunakan saat ini adalah SMOTE (*Synthetic Minority Oversampling Technique*) (Douzas *et al.* 2019). Algoritma SMOTE diyakini telah terverifikasi efektif dan valid dalam memproduksi sampel data sintetis dan telah banyak variasi SMOTE yang dibuat oleh beberapa peneliti (Xie dan Zhang 2018). Selain SMOTE, terdapat teknik *oversampling* baru yaitu menggunakan algoritme GAN (*Generative Adversarial Networks).* GAN merupakan sebuah algoritme deep learning yang yang digunakan untuk memproduksi data sintetis dari hasil pelatihan dua buah model algoritme yang saling bersaing (*adversarial*). Berbeda dengan SMOTE yang melakukan pendekatan pembelajaran berdasarkan informasi lokal dari dataset yang ada, GAN melakukan proses pembelajaran pada keseluruhan distribusi sampel data (Xie dan Zhang 2018). Perbedaan lainnya yaitu SMOTE memproduksi sampel data sintetis di sepanjang garis yang menghubungkan antar sampel data minoritas, sedangkan GAN memproduksi sampel data

sintetis dengan memperkirakan distribusi dari keseluruhan data aslinya (Douzas dan Bacao 2018).

Pada awalnya algoritme *Generative Adversarial Network* (GAN) diciptakan untuk memproduksi citra buatan/sintetis dengan mempelajari distribusi nilai pixel pada citra aslinya. Hingga pada tahun 2019, Xu *et al* melakukan penelitian dengan judul "*Modeling Tabular Data using Conditional GAN*". Dalam penelitian tersebut ingin dikembangkan algoritma GAN untuk dapat digunakan untuk memproduksi data sintetis dari data tabular. Penelitian tersebut berhasil mengembangkan algoritma CTGAN *(Conditional Tabular Generative Adversarial Network).* Oleh karena itu, penulis tertarik untuk meneliti apakah dengan menerapkan algoritma CTGAN tersebut mampu memproduksi data sintetis yang valid dan dapat digunakan untuk menambah data *imbalance* pada dataset citra multispektral lahan sawah untuk pemetaan kesuburan lahan.

Perumusan Masalah

Data citra lahan sawah mengalami masalah *imbalance data* sehingga data yang ada tidak cukup untuk digunakan dalam pemodelan pemetaan kesuburan lahan sawah, maka penulis merumuskan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan algoritma CTGAN *(Conditional Tabular Generative Adversarial Network)* untuk memproduksi data sintetis demi mengatasi *Imbalance Data* Citra Multispektral Lahan Sawah dalam pemetaan Kesuburan Lahan?
2. Apakah data sintetis yang dihasilkan oleh model algoritma CTGAN ini dapat digunakan sebagai data training yang valid?

Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan algoritma CTGAN *(Conditional Tabular Generative Adversarial Network)* untuk membangkitkan data tambahan/sintesis citra multispektral lahan sawah.
2. Membuat model klasifikasi kesuburan lahan sawah menggunakan data citra multispektral yang sudah ditambah dengan data sintetis.
3. Menganalisis pengaruh penambahan data sintetis terhadap akurasi model klasifikasi.

Manfaat Penelitian

Dengan menggunakan model dari CTGAN tersebut dapat membantu para peneliti dalam menambah data citra multispektral lahan sawah yang tidak seimbang sehingga para peneliti tidak mengulang pengambilan data ketika mengalami ketidakseimbangan data.

Ruang Lingkup Penelitian

Ruang lingkup pada penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan data sekunder citra multispektral yang diambil dari kamera UAV multispectral di atas lahan persawahan di Kecamatan Dramaga yang telah diambil dari penelitian sebelumnya.
2. Masing-masing *record* data menunjukkan rata-rata fitur multispectral pada suatu grid sawah berukuran 4 x 4 meter.
3. Label data adalah level kebutuhan pupuk pada grid-grid sawah tersebut, yang diukur menggunakan Bagan Warna Daun (BWD) atau Leaf Color Chart (LCC), di mana warna daun padi menjadi parameter kebutuhan pupuk tanaman padi.
4. Implementasi algoritma GAN (Generative Adversarial Network) menggunakan bahasa pemrograman python dengan memanfaatkan library CTGAN.

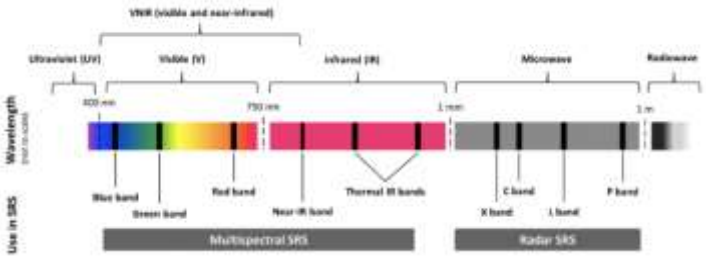
TINJAUAN PUSTAKA

Citra Multispektral untuk Pemetaan Lahan Sawah

Citra multispektral dapat diartikan sebagai citra yang berisi data gelombang frekuensi secara detail yang direpresentasikan dalam spektrum gelombang elektromagnetik. Frekuensi penglihatan manusia hanya terletak pada rentang panjang gelombang 400-700 nm, dan akibatnya manusia hanya dapat melihat dan membedakan warna mulai dari ungu hingga merah (Santoso 2009). Sedangkan citra multispektral menangkap data dari berbagai panjang gelombang pada seluruh spektrum elektromagnetik sehingga memberikan lebih banyak informasi yang diperlukan dalam menganalisis karakteristik dan mengidentifikasi komponen penyusun pada citra yang ditangkap (ElMasry *et al.* 2019). Oleh karena itu citra multispektral mampu menampilkan spektrum cahaya lain yang tidak tampak oleh penglihatan manusia tetapi sangat berguna untuk bermacam kebutuhan penelitian.

Citra multispektral dapat diperoleh dengan bermacam alat salah satunya dengan menggunakan UAV multispektral. UAV (*Unmanned Aerial Vehicle*) sendiri merupakan sebuah pesawat tanpa awak yang dapat dikendalikan oleh manusia menggunakan remot kontrol (Fahlstrom *et al.* 2022). UAV memiliki berbagai sensor yang dapat menyamai sensor yang ada pada sebuah satelit, sehingga dapat digunakan pada bidang pertanian. UAV Multispectral memiliki efisiensi yang tinggi untuk skala lapangan dalam pengambilan data (Shofiyanti 2011). Dengan efisiensi yang ditawarkan, maka UAV multispectral banyak dimanfaatkan untuk memetakan dan menganalisis lahan persawahan.

Hanya sebagian kecil dari gelombang elektromagnetik yang dapat dilihat oleh mata manusia. Berbeda dengan mata manusia, sensor kamera multispektral yang biasanya dijumpai pada satelit dan/atau pesawat tanpa awak (UAV) mampu menangkap lebih banyak informasi gelombang elektromagnetik misalnya gelombang inframerah, ultraviolet, dan bahkan gelombang mikro.



Gambar 1 Spektrum Gelombang Elektromagnetik (Prerona 2020)

Dalam dunia pertanian, citra multispektral sering digunakan untuk memetakan kesuburan lahan persawahan padi. Pemetaan ini menggunakan Bagan Warna Daun (BWD) atau Leaf Color Chart (LCC), di mana warna daun padi menjadi parameter untuk menentukan kebutuhan pupuk tanaman padi (Nasution *et al.* 2022). Citra multispektral dapat pula digunakan sebagai data latih untuk membuat sebuah model *machine learning* yang nantinya digunakan untuk melakukan klasifikasi tingkat kesuburan lahan pertanian, tetapi tentunya hal ini membutuhkan *record* data citra yang cukup banyak untuk melatih model tersebut (Wijayanto *et al.* 2020).

Di dalam sebuah penelitian yang dilakukan oleh Rokhmatuloh *et al.* (2019) dengan judul "Pemetaan Sawah Padi menggunakan citra UAV Multispektral". Penelitian ini ingin meneliti seberapa efektif sensor citra multispektral yang ada pada UAV untuk menghasilkan nilai NDVI (Normalized Difference Vegetation Index) dalam memetakan persawahan. UAV Multispektral yang digunakan memiliki sensor multispectral empat bands yaitu green, red, red edge, dan near-infrared. Dalam melakukan pemetaan sawah padi mereka menggunakan

nilai NDVI sebagai alat untuk membedakan berbagai jenis dan karakteristik tanaman. Untuk menghitung nilai NDVI ini digunakan formula perhitungan sebagai berikut:

$$NDVI = \frac{NIR - red}{NIR + red}, \quad (1)$$

di mana nilai *red* dan NIR merupakan nilai radiasi pada cahaya merah dan near-infrared yang ditangkap oleh sensor. Dari hasil penelitian ini diperoleh kesimpulan bahwa sensor multispektral pada UAV telah berhasil menghasilkan nilai NDVI untuk membantu memetakan lahan sawah dengan dapat membedakan jenis jenis tanaman yang ada pada lahan sawah yang diteliti. Selain itu nilai NDVI yang dihasilkan pula cukup akurat dalam membedakan umur tanaman padi, tetapi nilai NDVI ini cukup bervariasi dan tidak jarang beberapa tanaman yang sama memiliki nilai NDVI yang berbeda.

Masalah pada Imbalance Data

Masalah *Imbalance data* pada sebuah dataset dapat terjadi karena adanya kelas yang jumlah datanya sedikit (kelas minoritas) dibandingkan kelas lain (kelas mayoritas). Sebagai contoh, dari 1000 jumlah sampel terdapat 99% data yang tergolong kelas mayoritas dan hanya terdapat 1% data yang tergolong kelas minoritas. Ketika dataset ini dipakai untuk melakukan pelatihan model maka didapatkan akurasi 99% yang sebenarnya akurasi itu hanya berasal dari kelas mayoritas. Sedangkan 1% dari kelas minoritas itu bisa saja 1% mengandung informasi yang tidak kalah penting. Sehingga ketika menjalankan model klasifikasi, sebagian besar akan terjadi *misclassification* pada sampel data kelas minoritas, sedangkan sampel data pada kelas mayoritas akan jarang untuk terjadi kesalahan klasifikasi (Ali *et al.* 2019). Kesalahan klasifikasi pada model seperti ini memberikan dampak pada kerugian biaya dan waktu komputasi (Longadge *et al.* 2013).

Data yang tidak seimbang *(imbalance)* dapat menjadi masalah pada saat peneliti ingin melakukan training model untuk mendapatkan model klasifikasi machine learning. Karena setiap kelas data tidak memiliki jumlah yang sama, sehingga membuat tingkat akurasi klasifikasi pada setiap kelas menjadi tidak maksimal (Zhang *et al.* 2021). Klasifikasi data dengan distribusi kelas yang tidak seimbang akan sangat mempengaruhi kinerja machine learning dalam menghasilkan model klasifikasi. Sehingga akibatnya adalah pengklasifikasian menjadi bias terhadap *record* data yang jumlah kelasnya lebih banyak (Patel *et al.* 2020).

Generative Adversarial Networks

Algoritme *Generative Adversarial Networks* (GAN) adalah salah satu jenis algoritma *oversampling*. Disebut sebagai algoritma *oversampling* karena algoritme ini berusaha memproduksi atau meng-*generate* data sintetis yang menyerupai data asli yang ada dengan tujuan untuk menambah jumlah sampel data. GAN bekerja dengan mempelajari distribusi dari data asli baik data mayoritas maupun data minoritas, lalu memanfaatkan distribusi yang telah dipelajari untuk menghasilkan data sintetis (Sampath *et al.* 2021).

Algoritma GAN pertama kali diperkenalkan oleh Goodfellow *et al* pada tahun 2014 di dalam papernya yang berjudul *"Generative Adversarial Nets"*. Sesuai namanya, GAN (*Generative Adversarial Network*) merupakan dua buah model machine learning yang saling bersaing *(adversarial)* untuk mengoptimalkan kinerja masing-masing. Kedua model ini disebut sebagai *Generator* dan *Discriminator*. Model *Generator* akan membuat data sintetis dengan tujuan untuk menipu *Discriminator* sedangkan Model *Discriminator* akan mencoba melakukan klasifikasi antara data yang asli dengan data sintetis yang dibuat oleh *Generator*. Kedua jaringan ini memiliki kemampuan untuk mengoptimalkan kinerja masing-masing dengan konsep backpropagation. Ketika *Discriminator* berhasil membedakan antara data asli dengan data sintetis, maka dikembalikan sebagai feedback ke *Generator* dan *Generator* akan berusaha meningkatkan kinerja modelnya dalam memproduksi data sintetis yang

semirip mungkin dengan data asli. Begitu pula sebaliknya, *Discriminator* akan terus melakukan pembelajaran untuk meningkatkan kinerjanya dalam melakukan klasifikasi antara data sintetis yang berasal dari *Generator* dengan data asli secara akurat. Proses pelatihan selesai ketika *Discriminator* tidak lagi dapat membedakan antara data asli dan data sintetis. Secara matematis Goodfellow *et al.* (2014) menggambarkan GAN dalam Persamaan (2).

$$V(D,G) = E_{x\sim p_{data}(x)}[log\ log\ D(x)] + E_{z\sim p_z(z)}[log\ log\ (1 - D(G(z)))] \quad (2)$$

dimana z adalah variabel input dengan sampel data berupa random noise untuk melatih jaringan *Generator*, $D(x)$ merepresentasikan probabilitas x merupakan data asli bukan merupakan data hasil dari *Generator* (p_g). Diskriminator (D) dilatih untuk memaksimalkan probabilitas penentuan sampel data asli dan data sintetis. Secara simultan, Generator (G) dilatih untuk meminimalkan nilai $[log\ log\ (1 - D(G(z)))]$ agar D dapat terkecoh. Dengan kata lain, D dan G memainkan permainan minimax-game.

Pada paper ini disebutkan pula mengenai sebuah pengembangan dari GAN yaitu *Conditional Generative Adversarial Network model* atau CGAN dengan tujuan mendapatkan data sintetis sesuai kelas yang diinginkan. Dengan secara sederhana menambahkan variabel y sebagai input tambahan pada D dan G.

$$V(D,G) = E_{x\sim p_{data}(x)}[log\ log\ D(x|y)] + E_{z\sim p_z(z)}[log\ log\ (1 - D(G(z|y)))] \quad (3)$$

Conditional model ini mampu membatasi proses generasi data sehingga output dari G dapat dikontrol sesuai yang diinginkan. Nilai y merupakan informasi tambahan dapat berupa label kelas setiap data (Mirza dan Osindero 2014).

Secara umum, algoritme *Generative Adversarial Network* (GAN) sering digunakan untuk memproduksi citra buatan/sintetis dengan mempelajari distribusi data nilai pixel pada citra aslinya. Sebuah fungsi aktivasi seperti tanh ataupun sigmoid ditempatkan pada lapisan terakhir pada algoritma GAN untuk menghasilkan nilai output dengan rentang [-1,1] atau [0,1] sebagai nilai pixel pada citra. Kita ketahui bersama bahwa nilai *pixel* pada sebuah citra seperti mengikuti sebaran normal atau distribusi gauss, di mana nilai *pixel* ini dapat dinormalisasi dengan metode transformasi min-max. Berbeda dengan nilai pixel pada citra, data tabular biasanya memiliki nilai kontinu yang bukan merupakan sebaran normal. Sehingga ketika dilakukan transformasi min-max pada data tabular, akan menyebabkan *vanishing gradient problem* pada saat diimplementasikan ke-GAN (Xu *et al.* 2019)

Untuk mengatasi masalah *vanishing gradient problem*, Xu *et al. (*2019) melakukan penelitian dengan judul "*Modeling Tabular Data using Conditional GAN*". Mereka mengembangkan algoritma GAN sehingga dapat digunakan untuk memproduksi data sintetis dari data tabular. Kemudian mereka berhasil mengembangkan algoritma GAN yang mereka sebut sebagai CTGAN *(Conditional Tabular Generative Adversarial Network).* Mereka menunjukkan bahwa CTGAN mampu mempelajari distribusi data secara lebih baik dibandingkan *Bayesian Network*. Sedangkan mereka mengamati bahwa sejauh ini belum ada *deep learning generative model* yang mampu mengungguli Bayesian Network selain CTGAN ini.

Random Forest Classifier

Random forest merupakan salah satu algoritma *machine learning* yang umum digunakan dalam melakukan klasifikasi. *Random forest* merupakan gabungan dari beberapa *Decision Tree* di mana setiap *tree* bergantung pada nilai acak yang diambil pada dataset dengan distribusi yang sama pada setiap *tree*-nya (Breiman 2001). *Random forest* mempunyai keunggulan yaitu mampu mengatasi overfitting, tidak sensitif terhadap pencilan, dan memiliki akurasi yang baik (Ali *et al.* 2012). Algoritma Random Forest juga efektif dalam menangani data yang tidak seimbang dengan memberikan hasil performa yang baik (Bader-El-Den *et al.* 2019). Selain itu algoritma *Random Forest* pula dapat digunakan untuk memprediksi hasil panen secara efektif dan akurat dengan ketelitian yang tinggi (Geetha *et al.* 2020).

Confusion Matrix

Confusion matrix merupakan salah satu metode populer untuk mengukur performa dari suatu model klasifikasi dan mengevaluasi model (Siringoringo 2018)

Tabel 2.1 Confusion matrix untuk mengevaluasi performa model

CLASS	*Predictive Positive*	*Predictive Negative*
Actual Positive	TP	FN
Actual Negative	FP	TN

Untuk mengetahui tingkat keberhasilan model dapat dilakukan dengan menghitung tingkat akurasi, presisi, *recall* serta F1 *Score*-nya.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

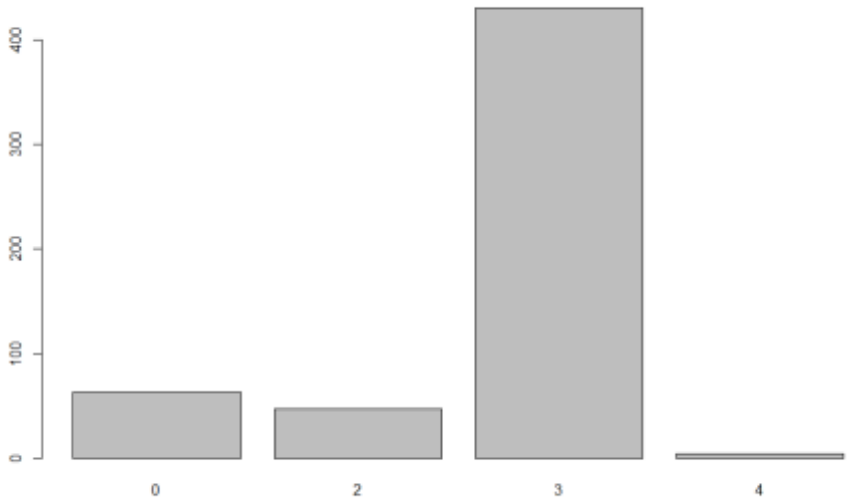
$$F1\ Score = \frac{2(Pecission*Recall)}{(Pecission*Recall)} \quad (6)$$

METODE

Data Penelitian

Data yang digunakan pada penelitian ini merupakan data sekunder yang diambil pada sebuah penelitian oleh Kahfi Gunardi seorang mahasiswa Program Magister IPB University yang meneliti tentang Perbandingan Algoritma Klasifikasi untuk Mendeteksi Kebutuhan Nitrogen Tanaman Padi berdasarkan Data Citra Multispektral Drone. Pengambilan data citra multispektral ini menggunakan sensor kamera *Unmanned Aerial Vehicle* (UAV) atau Drone pada lahan sawah di desa Margajaya Kecamatan Dramaga, Kabupaten Bogor.

Dataset ini memiliki total 544 record data dan memiliki 13 atribut yaitu "rgb_b1, rgb_b2, rgb_b3, green_b1, green_b2, ndvi_b1, nir_b1, nir_b2, red_b1, red_b2, redEdge_b1, redEdge_b2, Label". Terdapat atribut label terdiri dari tiga kelas yaitu 2, 3, dan 4 yang menggambarkan jumlah kebutuhan pupuk pada tanaman padi.

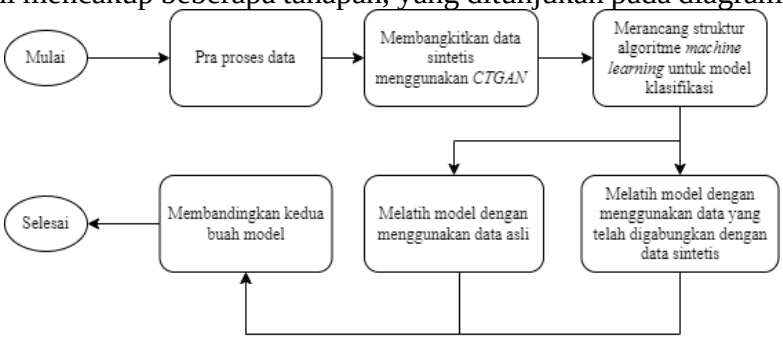


Gambar 2 Plot jumlah data pada setiap label kelas

Pada Gambar 2 diatas dapat kita lihat ada total 544 record data. Jumlah data paling banyak terdapat pada label 3 sebanyak 430 data, label 2 sebanyak 47 data, dan data yang paling sedikit terdapat pada label 4 sebanyak 4 data. Sedangkan yang berlabel 0 merupakan data yang tidak diketahui label sebenarnya sebanyak 63 data.

Tahapan Penelitian

Penelitian ini mencakup beberapa tahapan, yang ditunjukan pada diagram alir berikut:



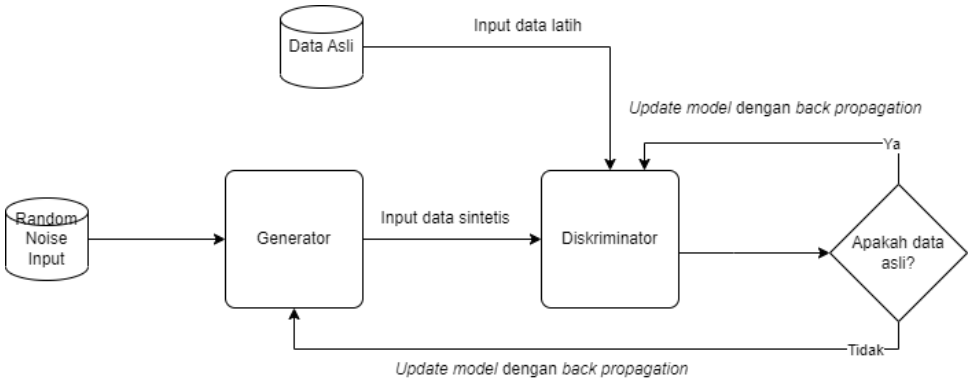
Gambar 3 Skema metode penelitian untuk mengetahui pengaruh penambahan data sintetis

Pra Proses Data

Melakukan data *pre-processing* untuk membuat sampel data yang akan diteliti lebih optimal dan siap digunakan. Tahap *pre-processing* pada penelitian ini meliputi pembersihan data (data *cleaning)* dan reduksi data. Pembersihan data ditujukan untuk menghilangkan *missing value* dan menghapus data yang tidak konsisten. Sedangkan reduksi data ditujukan untuk menghapus baris data yang tidak memiliki label kelas.

Membangkitkan Data Sintetis Dengan CTGAN

Pada tahap ini data yang telah dibersihkan akan digunakan pada algoritma CTGAN untuk menghasilkan data sintetis yang mirip dengan data aslinya. Algoritma CTGAN akan melatih dua model yaitu *Discriminator* dan *Generator*. Data yang asli akan dipelajari oleh *Discriminator*, selanjutnya *Generator* akan memproduksi data sintetis untuk diklasifikasikan oleh *Discriminator* dengan menentukan apakah itu merupakan data asli atau data sintetis. Ketika *Discriminator* berhasil menebak data tersebut sebagai data sintetis maka *Generator* akan meningkatkan performanya untuk memproduksi data sintetis yang semakin mirip dengan data aslinya. Proses ini akan terus berulang hingga *Discriminator* tidak lagi dapat membedakan antara data asli dengan data sintetis. Selanjutnya *Generator* akan memproduksi data sintetis sebanyak yang diinginkan. Hasil dari *Generator* ini akan digabungkan dengan data aslinya untuk digunakan pada tahap pelatihan model *Random Forest*.



Gambar 4 Ilustrasi alur kerja algoritma CTGAN

Merancang Struktur Model Klasifikasi

Model klasifikasi yang akan digunakan pada penelitian ini yaitu *Random Forest* dengan memanfaatkan library pada python yaitu sklearn.

Pelatihan Model

Pada tahap ini, akan dibuat dua buah model *Random Forest*. Model yang pertama menggunakan data sebagai data latih. Model kedua menggunakan data campuran di mana data ini merupakan perpaduan antara data asli dan data yang telah dihasilkan oleh CTGAN.

Evaluasi dan Membandingkan Hasil

Melakukan evaluasi dan membandingkan performa hasil pengujian antara kedua buah model *Random Forest*. Hasil pengujian kemudian dilakukan analisis untuk diambil kesimpulan. Pada tahapan penelitian ini menggunakan *confusion matrix* dalam menghitung akurasi, precision, F1 score, dan recall untuk mengukur kinerja suatu model. Dari hasil ini kemudian dapat diambil kesimpulan apakah dengan menggunakan CTGAN dapat menghasilkan data sintetis yang valid dan bisa digunakan untuk menambah data citra multispektral lahan sawah dalam pemetaan kesuburan lahan.

Lingkungan Pengembangan (atau Peralatan Penelitian)

Spesifikasi perangkat keras dan perangkat lunak yang digunakan untuk penelitian ini adalah sebagai berikut:

1. Perangkat keras dengan spesifikasi:
    - Processor AMD Ryzen 5 4600H CPU @ 3.0Ghz.
    - Graphic Processor NVIDIA® GeForce® GTX 1050.
    - RAM 16 GB.
    - SSD 512 GB.
2. Perangkat lunak yang digunakan:
    - Sistem Operasi Windows 11 Home 64-bit
    - Bahasa Pemrograman Python 3.7.15
    - Lingkungan Pengembangan Google Collabs.
    - Library python SKLEARN
    - Library python CTGAN

JADWAL PENELITIAN

Penelitian ini akan dilaksanakan mulai bulan Januari hingga Juni tahun 2019. Jadwal penelitian dapat dilihat pada Tabel 2.

Tabel 2 Jadwal Penelitian

No	Kegiatan	Tahun 2023															
		Januari				Februari				Maret				April			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Pra Proses data																
2	Perancangan dan pelatihan Model CTGAN																
3	Pembuatan data sintetis dari model CTGAN																
4	Perancangan struktur model klasifikasi random forest																
5	Melatih model *Random Forest* menggunakan data asli																
6	Melatih model *Random Forest* menggunakan data gabungan asli dan sintetis																
7	Evaluasi dan Membandingkan model																
8	Seminar Hasil																
9	Revisi																
10	Sidang																

DAFTAR PUSTAKA

Ali H, Salleh M, Saedudin R, Hussain K, Mushtaq M. 2019. Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*. 14.doi:10.11591/ijeecs.v14.i3.pp1552-1563.

Ali J, Khan R, Ahmad N, Maqsood I. 2012. Random Forests and Decision Trees. *International Journal of Computer Science Issues(IJCSI)*. 9.

Bader-El-Den M, Teitei E, Perry T. 2019. Biased Random Forest For Dealing With the Class Imbalance Problem. *IEEE Transactions on Neural Networks and Learning Systems*. 30(7):2163–2172.doi:10.1109/TNNLS.2018.2878400.

Douzas G, Bacao F. 2018. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*. 91:464–471.doi:10.1016/j.eswa.2017.09.030.

Douzas G, Bacao F, Fonseca J, Khudinyan M. 2019. Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the

Geometric SMOTE Algorithm. *Remote Sensing*. 11(24):3040.doi:10.3390/rs11243040.

ElMasry G, Mandour N, Al-Rejaie S, Belin E, Rousseau D. 2019. Recent Applications of Multispectral Imaging in Seed Phenotyping and Quality Monitoring—An Overview. *Sensors*. 19(5):1090.doi:10.3390/s19051090.

Fahlstrom PG, Gleason TJ, Sadraey MH. 2022. *Introduction to UAV Systems*. John Wiley & Sons. Ed ke-Google-Books-ID: s8Z6EAAAQBAJ.

Geetha V, Punitha A, Abarna M, Akshaya M, Illakiya S, Janani AP. 2020. An Effective Crop Prediction Using Random Forest Algorithm. Di dalam: *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*. *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*;. hlm. 1–5.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative Adversarial Nets. Di dalam: *Advances in Neural Information Processing Systems*. [internet] Vol. 27. Curran Associates, Inc. [diunduh 2022 Okt 14]. Tersedia pada: https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Longadge MR, Dongre SS, Malik DL. 2013. Class Imbalance Problem in Data Mining: Review. 2(1):6.

Mirza M, Osindero S. 2014. Conditional Generative Adversarial Nets. [diunduh 2022 Okt 8]. Tersedia pada: http://arxiv.org/abs/1411.1784

Nasution EKI, Ritonga EN, Siregar ES, Harahap S. 2022. Pengaruh Olah Tanah dan Pemberian Pupuk N Berdasarkan BWD (Bagan Warna Daun) terhadap Pertumbuhan dan Produksi Padi Sawah Varietas Mekongga (Oryza sativa L.). *Formosa Journal of Multidisciplinary Research*. 1(3):455–468.doi:10.55927/fjmr.v1i3.717.

Patel H, Singh Rajput D, Thippa Reddy G, Iwendi C, Kashif Bashir A, Jo O. 2020. A review on classification of imbalanced data for wireless sensor networks. *International Journal of Distributed Sensor Networks*. 16(4):1550147720916404.doi:10.1177/1550147720916404.

Prerona. 2020. A technical deep-dive into Satellite Imaging, Multispectral, SAR and GAN. *AiDash*. [diunduh 2022 Nov 18]. Tersedia pada: https://www.aidash.com/a-technical-deep-dive-into-satellite-imaging-multispectral-sar-and-gan/

Rokhmatuloh, Supriatna, Giok Pin T, Hernina R, Ardhianto R, Putut Ash Shidiq I, Adi W. 2019. Paddy Field Mapping Using Uav Multi-Spectral Imagery. *International Journal of GEOMATE*. 17(61):242–247.doi:https://doi.org/10.21660/2019.61.icee408.

Rout N, Mishra D, Mallick MK. 2018. Handling Imbalanced Data: A Survey. Di dalam: Reddy MS, Viswanath K, K.M. SP, editor. *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*. Singapore. Singapore: Springer. (Advances in Intelligent Systems and Computing). hlm. 431–443.

Sampath V, Maurtua I, Aguilar Martín JJ, Gutierrez A. 2021. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *J Big Data*. 8(1):27.doi:10.1186/s40537-021-00414-0.

Santoso I. 2009. *Interaksi Manusia dan Komputer*. Ed ke-2. C.V. Andi Offset. Ed ke-Google-Books-ID: _pXa7CvwTC0C.

Shofiyanti R. 2011. TEKNOLOGI PESAWAT TANPA AWAK UNTUK PEMETAAN DAN PEMANTAUAN TANAMAN DAN LAHAN PERTANIAN. *Informatika Pertanian*. 20(2):58–64.

Siringoringo R. 2018. KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR. *Journal Information System Development (ISD)*. 3(1). [diunduh 2022 Des 15]. Tersedia pada: https://ejournal-medan.uph.edu/index.php/isd/article/view/177

Spelmen VS, Porkodi R. 2018. A Review on Handling Imbalanced Data. Di dalam: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*;. hlm. 1–11.

Wang Y-P, Chang Y-C, Shen Y. 2021. Estimation of nitrogen status of paddy rice at vegetative phase using unmanned aerial vehicle based multispectral imagery. *Precision Agric*. 23(1):1–17.doi:10.1007/s11119-021-09823-w.

Wijayanto AW, Wahyu Triscowati D, Marsuhandi AH. 2020. Maize field area detection in East Java, Indonesia: An integrated multispectral remote sensing and machine learning approach. Di dalam: *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)*. *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)*;. hlm. 168–173.

Xie Y, Zhang T. 2018. Imbalanced Learning for Fault Diagnosis Problem of Rotating Machinery Based on Generative Adversarial Networks. Di dalam: *2018 37th Chinese Control Conference (CCC)*. *2018 37th Chinese Control Conference (CCC)*;. hlm. 6017–6022.

Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. 2019. Modeling Tabular data using Conditional GAN. Di dalam: *Advances in Neural Information Processing Systems*. [internet] Vol. 32. Curran Associates, Inc. [diunduh 2022 Nov 10]. Tersedia pada: https://papers.nips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Zhang C, Li J, Zhao Y, Li T, Chen Q, Zhang X, Qiu W. 2021. Problem of data imbalance in building energy load prediction: Concept, influence, and solution. *Applied Energy*. 297:117139.doi:10.1016/j.apenergy.2021.117139.