Penerapan Algoritma Random Forest dengan Cost Complexity Pruning untuk Klasifikasi Kualitas Udara Berdasarkan Indeks AQI

1st Farrell Kamala Apriansyah
Program Studi S1 Informatika
Telkom University
Bandung, Indonesia
farrellka@student.telkomuniversity.ac.id

2nd Aji Gautama Putrada,
Program Studi S1 Informatika
Telkom University
Bandung, Indonesia
ajigps@telkomuniversity.ac.id

3rd Ryan Lingga Wicaksono
Program Studi S1 Informatika
Telkom University
Bandung, Indonesia
wicaksonoryan@telkomuniversity.ac.id

***Abstrak* — Kualitas udara tidak boleh dianggap remeh karena itu adalah masalah yang penting untuk kesehatan dan lingkungan masyarakat. Model random forest seringkali tidak efisien dalam hal komputasi, yang menghasilkan pohon keputusan dengan kedalaman yang besar. Studi ini menggunakan metode Cost-Complexity Pruning (CCP) pada algoritma random forest untuk mengoptimalkan ukuran model tanpan mengurangi akurasi. Dataset yang digunakan pada penelitian ini Air Quality Index (AQI) yang didapat dari website kaggle, yang dimana sudah ada kategori menjadi 4 level, yaitu *Good, Moderate, Unhealthy, and Hazardous*. Percobaan menghasilkan bahwa penerapan pruning dengan nilai pararameter ccp_alpha = 0.001 menunjukkan keseimbangan yang optimal antara akurasi dan efisiensi, dengan akurasi mencapai 99.63% yang hanya sedikit lebih rendah dari model tanpa pruning yaitu 99.95%, tetapi dengan pengurangan ukuran model yang sangat signifikan hingga mencapai sekitar 79% lebih kecil dan jumlah node berkurang hingga 80% dari model non-pruning. Penelitian ini menunjukkan bahwa metode cost-complexity pruning secara efektif dapat mengurangi kompleksitas model tanpa menurunkan akurasi secara signifikan**

***Kata kunci*— machine learning, random forest, index AQI, CCP, kualitas udara, pruning.**

I. PENDAHULUAN

Isi Polusi udara adalah salah satu permasalahan lingkungan yang sangat sulit diselesaikan dan masalah ini juga akan langsung berdampak pada kesehatan manusia dan juga dengan iklim global. Air Quality Index (AQI) adalah salah satu indikator yang dapat menyediakan informasi tentang pencemaran udara berdasarkan konsentrasi zat polutan seperti PM2.5, PM10, $NO2$, dan $O3$, dengan ini masyarakat dapat mengetahui seberapa berbahaya kondisi Udara di lingkungannya [1].

Menurut Laporan WHO pada tahun 2018 yang dikutip dalam studi Alzu'bi [2], lebih dari seperempat kematian orang dewasa di Mediterania Timur Disebabkan oleh Polusi Udara yang dipicu aktivitas antropogenik seperti industri dan lalu lintas kendaraan bermotor. Menurut Menéndez García dkk. [3], polusi udara menyebabkan sekitar 8,5 juta kematian setiap tahun secara global. Oleh karena itu system pemantauan AQI yang andal, efisien, dan akurat sangat penting, khususnya didaerah urban dengan tingkat emisi yang sangat tinggi. di Indonesia, kualitas udara bersih telah menurun hingga ke tingkat kritis di beberapa wilayah metropolitan, termasuk Jakarta dan kota-kota disekitarnya, Sehingga menimbulkan kekhawatiran tentang dampak pada Kesehatan dan lingkungan karena Udara bersih yang menurun. oleh karena itu sistem pemantauan AQI menjadi sangat krusial, dan pendekatan berbasis data sangat bermanfaat untuk mendukung pengambilan keputusan secara cepat.

Seiring berkembangnya zaman dan teknologi canggih pun bermunculan, pendekatan berbasis machine learning sudah lebih umum dipakai untuk menangani tantangan dalam pemantauan kualitas udara. Random forest telah digunakan secara luas dalam studi kualitas udara karena kemampuannya menangani hubungan non- linear yang kompleks antara variabel meteorologi dan konsentrasi polutan, serta menghasilkan performa prediksi yang kompetitif dalam berbagai kondisi spasial dan temporal [4]. Model random forest mampu mendapatkan akurasi yang tinggi 1 untuk data urban, ini membuktikan random forest bisa menjadi pilihan yang sangat ideal untuk sistem pemantauan berbasis data [5], [6]. Tetapi, walaupun random forest bisa mendapatkan akurasi tinggi ada masalah yang muncul juga, yaitu waktu inferensi yang lama dan juga ukuran yang besar, sehingga kurang efisien untuk implementasi dalam perangkat dengan sumber daya yang terbatas atau kebutuhan real-time [7].

Pada penelitian ini mencoba penerapan teknik pruning menggunakan parameter cost-complexity alpha (ccp_alpha) pada algoritma random forest. Teknik ini dapat mengurangi kompleksitas pohon keputusan dengan memangkas node-node yang dianggap tidak penting, sehingga dapat meningkatkan efisiensi dan akurasi model secara bersamaan agar bisa menghemat sumber daya komputasi dan dapat diterapkan menjadi lebih luas untuk sistem pemantauan kualitas udara.

II. KAJIAN TEORI

Studi tentang pemantauan dan klasifikasi kualitas udara sudah banyak berkembang pesat seiring berkembangnya zaman. pada studi Hamami dan Dahlan (2022) menggunakan pendekatan machine learning untuk klasifikasi udara di daerah perkotaan, pada studinya membuktikan bahwa pendekatan machine learning mampu mengenali pola polutan yang kompleks [1]. Lalu ada Chaturvadi (2024) yang membangun sistem prediksi kualitas udara yang komprehensif, mereka menggunakan algoritma supervised learning yang menunjukkan cukup efektif menangani variabel lingkungan yang dinamis [6]. untuk kasus yang lebih spesifik, pada studi Alzu'bi dkk. (2024) melakukan penelitian kasus di kota Amman Zarqa menggunakan RF untuk memprediksi kualitas udara secara spasial, namun ada tantangan dalam resolusi data yang didapat [2]. Kumbhar dan Jagtap (2025) memperbarui pendekatan ini dengan lebih memfokuskan pada prediksi rating indeks kualitas udara, meskipun penelitian mereka lebih merujuk pada implementasi antarmuka pengguna dibandingkan optimasi struktur internal model [5].

Lalu efisiensi komputasi juga menjadi isu krusial selain akurasi prediksi dalam pengembangan model yang siap dipakai. pada studi Menéndez García dkk. (2022). memperkenalkan pendekatan menggunakan metode pruning untuk menangani data yang hilang pada deret waktu kualitas udara, yang membuktikan bahwa penyederhanaan data tidak selalu mempengaruhi performa prediksi secara signifikan [3]. Lalu pada studi Sakthivel dkk. (2022), yang mengembangkan teknik pruning berbasis machine learning untuk komputasi yang lebih rendah (low power approximate computing), yang dimana memangkas node yang tidak berpengaruh pada akurasi terbukti mengurangi beban komputasi secara signifikan [8]. Pruning juga tidak terbatas pada ensemble learning, pada studi Li (2024) yang membahas bahwa teknik pruning pada *Deep Neural Networks* telah berevolusi yang memungkinkan untuk menyimpan model yang jauh lebih hemat (low storage), sebuah konsep yang relevan bisa digunakan pada algoritma random forest [9].

Adapun bukti lainnya kemampuan random forest yang memang terbukti memiliki performa yang tangguh dalam menangani berbagai tipe data dan kasus kompleks di berbagai sektor krusial, dalam bidang kesehatan Mahendra dkk. (2025) dalam pada penelitiannya mengembangkan model untuk memprediksi penyakit kanker serviks berbasis ensemble learning yang melibatkan algoritma random forest yang dimana model tersebut bisa mencapai tingkat akurasi hingga 98,4%, presisi 98,45% dan recall 98,45% [10]. Pada penelitian Triyani dkk. (2020) juga menunjukkan penggunaan random forest yang dikombinasikan dengan seleksi fitur Discrete Wavelet Transform (DWT) yang dapat menghasilkan akurasi untuk klasifikasi penyakit kanker hingga 97,06% pada data *microarray* [11].

Random forest juga memiliki keunggulan dalam menangani data teks dan *Natural Language Processing* yang berhasil medapatkan performa cukup baik. Pada penelitian Purbolaksono (2024) algoritma random forest bisa menghasilkan akurasi sebesar 84,77% dalam menganalisis sentimen ulasan game di platform Steam menggunakan algoritma random forest [12]. Lalu adapun itu, pada kasus klasifikasi multi-label yang lebih kompleks untuk ayat-ayat-Al-Qur'an, random forest juga membuktikan kemampuannya yang tangguh bisa mencapai nilai Subset Accuracy tertinggi hingga 88,75% dengan tingkat kesalahan (Hamming Loss) yang sangat rendah yaitu 0,0075 [13].

Adapun pada studi kasus pengolahan data pola gerak olahraga, Muslim dkk. (2025) menggunakan algoritma random forest untuk klasifikasi pose memanah yang dikombinasikan dengan YOLOv8, yang menghasilkan kinerja klasifikasi yang presisi dalam mengenali postur atlet secara real-time [14].

Terakhir ada pada kasus domain lingkungan yang selaras dengan penelitian ini, Palupi dkk. (2023) menunjukkan ketangguhan random forest dalam memprediksi titik panas kebakaran hutan, model random forest yang dikembangkan bisa menghasilkan skor determinasi $R2$ hingga 0,9639 pada data uji, yang dimana ini membuktikan ketangguhan algoritma random forest dalam menjelaskan secara akurat variansi data lingkungan hingga 96% [15]. Hal ini membuktikan alasan kuat pemilihan algoritma random forest sebagai metode utama dalam penelitian ini.

Namun, menggunakan metode pruning sering kali menghadapi tantangan yang dimana kelas pada data yang dikumpulkan tidak seimbang (*class imbalance*). pada studi Rahmati (2025) menunjukkan adanya *trade-off* antara kompresi model dan keadilan (*fairness*), yang dimana strategi untuk menggunakan metode pruning yang cukup agresif cendrung mengorbankan akurasi pada data kelas minoritas demi mengecilkan model [16]. ini dibuktikan pada studi Alshdaifat dkk. (2025) yang menyimpulkan bahwa dataset yang tidak seimbang pada algoritma pruning standar sering kali memangkas node pada kelas minoritas [17]. Risiko ini juga ditekankan oleh Boyko dan Mokryk (2021) dalam studi deteksi penipuan, model menunjukkan kegagalan dalam mengenali kelas minoritas (kasus positif/bahaya) yang memiliki konsekuensi yang jauh lebih fatal daripada kelas mayoritas [18].

Dalam upaya menungkatkan kinerja algoritma dasar, Sun dkk. [7] mengusulkan adanya perbaikan pada random forest berdasarkan pengukuran korelasi pada decision tree, sementara Hubens dkk. [19] mengeksplorasi dampak pre-training pada efektivitas pruning. Selain itu ada Diaz Resquin dkk. [4] menunjukkan adanya perubahan pola kualitas udara yang drastis yang memerlukan model yang lebih adaptif. Berdasarkan tinjauan pustaka tersebut, penelitian ini berada di posisi untuk mengisi celah pada penelitian dengan menerapkan metode Cost-Complexity Pruning terhadap keseimbangan antara efisiensi ukuran model dan akurasi model, aspek yang belum dibahas secara mendalam pada penelitian prediksi AQI sebelumnya.

A. Random Forest

Random Forest (RF) adalah algoritma machine learning yang terdiri dari sekumpulan *Decision Tree* yang telah dibangun secara acak. prinsip utama algoritma ini adalah untuk menggabungkan prediksi dari banyaknya model individu untuk menghasilkan satu keputusan akhir yang lebih akurat dan stabil (*Ensemble Learning*) [20]. Dalam setiap decision tree yang terbuat, algoritma menggunakan kriteria *splitting (splitting criteria)* seperti Gini Index untuk menentukan fitur terbaik yang memisahkan data pada setiap node yang ada.

B. Classification

pada kasus klasifikasi status kualitas udara, Random Forest menggunakan mekanisme *Majority Voting*. Yang dimana prediksi final ditentukan oleh kelas yang sering muncul dari hasil seluruh prediksi pohon. Berikut rumus *Majority Voting* [2], [7] :

$$\hat{y} = mode\{T_1(x), T_2(x), \dots, T_M(x)\} \quad (1)$$

$\hat{y}$ merupakan hasil prediksi akhir model Random Forest yang diperoleh melalui mekanisme *majority voting*. Variabel x merepresentasikan vektor input yang berisi kumpulan fitur kualitas udara yang dimasukkan ke dalam model. Parameter M menyatakan jumlah pohon keputusan ($n\_estimators$) yang dibangun dalam algoritma Random Forest. Setiap $T_i(x)$ menunjukkan hasil prediksi klasifikasi yang dihasilkan oleh pohon keputusan individual ke-i terhadap data input x. Prediksi akhir ditentukan berdasarkan kelas yang paling sering muncul dari seluruh prediksi pohon keputusan tersebut.

C. *C*ost-Complexity Pruning

Cost-Complexity Pruning adalah metode penyederhanaan model decision tree yang bertujuan menyeimbangkan akurasi model dengan kompleksitas strukturnya. Metode ini bekerja dengan cara memangkas node yang tidak mempengarusi akurasi prediksi pada model, sehingga menghasilkan model yang lebih ringkas [9], [16].

Algoritma ini mencari sub-pohon terbaik dengan meminimalkan Cost-Complexity yang didefinisikan sebagai berikut:

$$R_\alpha(T) = R(T) + \alpha|T| \quad (2)$$

$R_\alpha(T)$ merupakan nilai cost-complexity total dari pohon keputusan T yang dipengaruhi oleh parameter kompleksitas α. Komponen $R(T)$ merepresentasikan tingkat kesalahan klasifikasi atau ukuran impurity dari pohon T sebelum dilakukan *pruning*. Sementara itu, $|T|$ menyatakan jumlah node pada pohon keputusan T, yang merefleksikan tingkat kompleksitas struktur pohon. Parameter $\alpha \geq 0$ berfungsi sebagai pengendali kompleksitas yang mengatur *trade-off* antara kesalahan klasifikasi dan ukuran pohon, di mana nilai α yang lebih besar akan mendorong terbentuknya pohon yang lebih sederhana melalui pengurangan jumlah node.

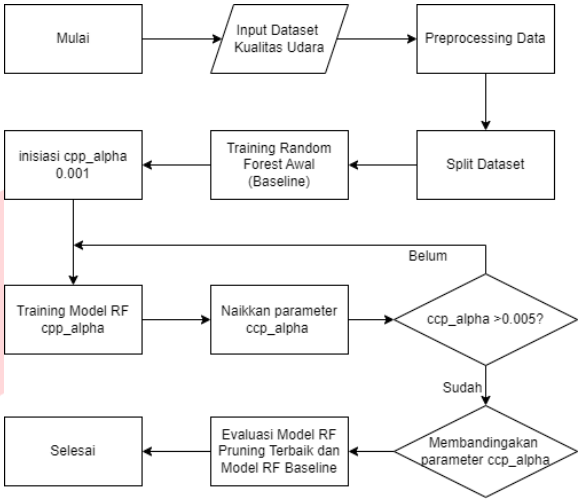
Nilai α (ccp_alpha) berfungsi sebagai pinalti. Jika $\alpha = 0$, fungsi Cost hanya memperhitungkan error (fullygrown tree). Semakin besar penalti untuk kompleksitas, akan memaksa algoritma untuk memangkas pohon menjadi lebih sederhana [9].

D. Imbalance Data

Dalam kasus klasifikasi kualitas udara, distribusi data sering kali tidak seimbang. Merujuk pada penelitian Chen dkk.[21], algoritma Random Forest standar cendrung lebih bias ke kelas mayoritas dibandingkan dengan kelas minoritas untuk memaksimalkan total akurasi model. Ketika metode pruning digunakan, ada risiko untuk nodes kelas minoritas akan dipangkas karena dianggap tidak berkontribusi terhadap akurasi model dan akan dianggap noise oleh model, yang menyebabkan penurunan yang cukup signifikan pada kemampuan model untuk deteksi kelas minoritas [17].

III. METODE

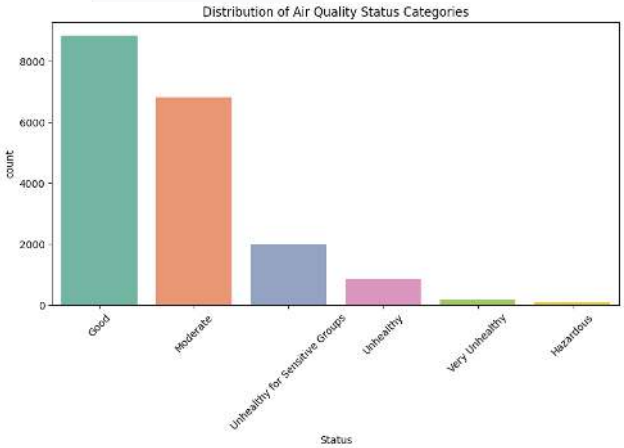
Pada bini meagian menjelaskan alur rancangan model Random Forest dari awal hingga akhir. Secara garis besar, alur penelitian meliputi tahapan: Input Dataset, prapemrosesan data, latih model, evaluasi.



GAMBAR 1

A. Input Dataset

Pada penelitian ini dataset yang didapat ada 4 atribut, yaitu *Date, Country, Status, AQI Value* tetapi hanya 2 atribut yang digunakan dari dataset yang didapat, yaitu nilai AQI sebagai input dan status kategori sebagai target. 2 Atribut ini dipilih dengan tujuan Eksperimen Terkendali (*Controlled Experimentation*) untuk menguji dampak murni dari metode pruning pada algoritma Random Forest standar. Untuk atribut sisanya tidak digunakan karena lemahnya feature important pada model ini. Berikut distribusi atribut status :



GAMBAR 2
(A)

B. Pra-pemrosesan Data

Data yang didapat biasanya ada beberapa yang hilang, outlier, atau formatnya tidak dapat diproses oleh model. Oleh karena itu tahap pra-pemrosesan dilakukan untuk menjaga data agar tetap bersih dan siap untuk pelatihan model.

C. *Split* Dataset

Setelah tahap pra-pemrosesan, tahap selanjutnya adalah membagi dataset dibagi menjadi 2 bagian, data latih dan data uji. Data dibagi 80% untuk data latih dan sisanya 20% untuk data uji. Proses distribusi ini menjamin bahwa model dapat diuji dengan data yang belum dikenali pada latih model, sehingga pada saat uji model dapat performa yang seimbang.

D. Latih Model *Random Forest Baseline*

Setelah membagi data, selanjutnya adalah latih model Random Forest. Model ini merupakan ensemble learning yang menggabungkan beberapa decision tree untuk menghasilkan prediksi yang stabil dan akurat. Parameter dasar seperti jumlah pohon (n_estimators) dan kedalaman maksimum pohon (max_depth) digunakan untuk membentuk struktur awal model.

E. Latih Model dengan *Pruning* ccp_alpha

Tahap Selanjutnya adalah melatih model ke-2 yaitu model RF menggunakan teknik pruning ccp alpha yang dimulai dari nilai parameter 0.001 hingga 0.005, jika nilai parameter belum mencapai 0.005 maka akan terus dinaikkan hingga 0.005.

F. Membandingkan Nilai Parameter ccp_alpha

Pada tahap ini melakukan analisis komparatif yang didapat dari hasil nilai parameter ccp_alpha 0.001-0.005. Yang bertujuan untuk megevaluasi pengaruh setiap nilai alpha terhadap kompleksitas pohon dan akurasi prediksi, untuk menentukan nilai parameter yang terbak. Nilai parameter yang terpilih atau terbaik yaitu yang bisa memberikan keseimbangan yang idean antara kompleksitas model dan performa model yang nantinya akan dibandingkan dengan model RF Baseline.

G. Evaluasi Model RF *Baseline* dan RF *Pruning*

Evaluasi model RF Baseline dengan RF Pruning dengan nilai parameter alpha terbaik, ini dilakukan menggunakan data testing untuk mengukur seberapa baik performa model dalam melakukan klasifikasi kualitas udara. Berikut adalah metrik yang digunakan :

Accuracy, yang menghitung proporsi prediksi yang benar:

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

Precision per kelas, yaitu ketepatan prediksi positif:

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

Recall per kelas, yaitu kemampuan model menemukan semua kasus positif:

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

F1-Score, yaitu rata-rata harmonik antara precision dan recall :

$$F_1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + \ Recall} \quad (6)$$

Selain evaluasi metriks diatas, juga digunakan confusion matrix untuk menganalis kesalahan klasifikasi antar kategori AQI. Confusion matrix menyajikan informasi jumlah prediksi benar dan salah untuk setiap kelas, yang direpresentasikan dalam bentuk tabel. Baris pada confusion matrix menunjukkan label aktual, sedangkan kolom menunjukkan label prediksi. Dengan metriks ini, dapat diketahui jenis kesalahan yang paling sering terjadi, misalnya apakah model sering mengklasifikasikan AQI "*Unhealthy*" sebagai "*Moderate*". Analisis ini sangat membantu dalam mengidentifikasi kelemahan model dan kelas yang membingungkan.

IV. HASIL DAN PEMBAHASAN

Tahap pengujian menggunakan 2 model, yaitu RF *Baseline* dan RF *Pruning* ccp_alpha, dengan tujuan untuk memvalidasi performa dari RF dalam mengklasifikasikan kualitas udara berdasarkan dataset yang didapat.

RF *Baseline* dibangun tanpa menggunakan *pruning.* Pada model ini dilatih dan diuji dengan pengaturan *default* tanpa pembatasan kedalaman pohon atau pemangkasan. Tujuannya untuk mendapatkan hasil akurasi yang maksimal yang bisa dicapai oleh algoritma dengan cara membiarkan decision tree tumbuh secara penuh (fully grown tree) hingga setiap daun bisa mencapai batas minimum sample. Dan model ini diharapkan memiliki kompleksitas yang tinggi dengan jumlah *nodes* yang banyak yang nanti akan digunakan sebagai tolak ukur perbandingan dengan model yang menggunakan metode *pruning*. Model ini menggunakan konfigutasi dengan pembagian data 80% latih dan 20% uji, Parameter *n_estimator* 100 (*default*) yang dimana model membangun 100 *decision tree*, Parameter criterion Gini Impurity *(Default)* untuk mengukur kualitas pemisahan *(split)* dan Parameter ccp_alpha: 0.0 (tanpa *pruning*).

Untuk model kedua berfokus pada optimasi model yaitu Cost-Complexity Pruning Alpha (ccp_alpha), metode ini diterapkan dengan cara menyuntikkan parameter ccp_alpha ke dalam algoritma. Eksperimen dilakukan dengan memvariasikan nilai parameter ccp_alpha dimulai dari 0.001 sampai dengan 0.005. dan nilai terpilih adalah ccp_alpha = 0.001 untuk evaluasi akhir. Tujuan menyuntikkan metode ini adalah untuk memangkas cabang-cabang pohon yang tidak memberikan kontribusi signifikan terhadap penurunan performa model, guna mengurangi ukuran model dan kompleksitas komputasi. Konfigurasi untuk model ini sama seperti RF *Baseline* hanya berbeda dengan menambah parameter *alpha* 0.001 sampai 0.005.

A. Hasil

Pada hasil percobaan dengan model baseline yang sudah dilatih dengan data uji menunjukkan performa yang sangat tinggi. Pada Tabel 1 menampilkan detail klasifikasi untuk tiap kelasnya.

TABEL 1
(A)

Performa Model Baseline			
	Precision	*Recall*	*F1-Score*
Good	*1.00*	*1.00*	*1.00*
Hazardous	*1.00*	*0.90*	*0.95*
Moderate	*1.00*	*1.00*	*1.00*
Unhealthy	*0.99*	*1.00*	*1.00*
Unhealthy for Sensitive Groups	*1.00*	*1.00*	*1.00*
Very Unhealthy	*0.97*	*0.97*	*0.97*
Macro avg	*0.99*	*0.98*	*0.99*
Weight avg	*1.00*	*1.00*	*1.00*
Accuracy	*0.9995*		
Jumlah Nodes	*33,832*		
Ukuran Model (KB)	*3742.93*		

Berdasarkan Tabel 1 menunjukkan bahwa performa model Baseline sangat bagus. Dengan mencapai akurasi hingga 0.9995 dengan jumlah nodes hingga 33,832 dan ukuran model hingga 3742.93. Secara overall model sudah bagus untuk di akurasi tetapi masih kompleks dan juga jumlah nodes masih besar.

Sedangkan model RF ccp_alpha dengan nilai parameter dimulai dari 0.001 hingga 0.005. berikut adalah perbandingan efisiensi model penerapan metode ccp_alpha yang memberikan dampak yang signifikan terhadap performa model yang sebagaimana dirangkum dalam Tabel 2.

TABEL 2
(A)

Perbandingan performa ccp_alpha daro 0.001 hingga 0.005					
ccp_alpha	**Akurasi**	**Total *Node***	**Ukuran Model (KB)**	**F1-Score (Hazardous)**	**F1-Score (Very Unhealthy)**
0.001	**0.9963**	**6728**	**778.43**	**0.46**	**0.85**
0.002	**0.9915**	**4610**	**546.77**	**0.00**	**0.56**
0.003	**0.9877**	**3668**	**443.74**	**0.00**	**0.00**
0.004	**0.9877**	**3088**	**380.30**	**0.00**	**0.00**
0.005	**0.9877**	**2742**	**342.46**	**0.00**	**0.00**

Pada Tabel 2 menunjukkan kalau nilai parameter ccp_alpha 0.001 menghasilkan model ddengan total 6.728 node berkurang hingga 27.104 dibandingkan dengan model RF *Baseline* dan dengan ukuran model hanya 778.43 KB jauh lebih rendah dibanding RF *Baseline*. Dari sisi performa, akurasi mencapai 0.9963 hanya lebih rendah sedikit dibanding dengan RF *Baseline* (0.9995). berdasarkan tabel diatas membuktikan semakin tinggi nilai parameter pruning akan semakin aggresif dan semakin tidak mengenali kelas minoritas.

GAMBAR 3
(A)

Gambar 3 menunjukkan hasil *Confusion Matrix* model RF *Baseline* dan RF ccp_alpha 0.001. Nilai parameter ini dipilih untuk perbandingan dengan model RF *Baseline* karena menghasilkan performa paling optimal diantara nilai parameter lainnya. Walaupun model RF ccp_alpha mengalami penurunan performa dalam mengenali kelas minoritas (*Hazardous, Very Unhealthy*) tetapi masih bisa menghasilkan performa yang sangat bagus untuk mengenali pola utama dataset. Pada model RF *Baseline* sangat mampu mengenali semua kelas, termasuk kelas minoritas yang dimana kesalahannya sangat minim. Terutama pada kelas minoritas *Hazardous* yang dimana datanya sangat sedikit tetapi hanya salah prediksi pada 1 data saja.

B. Analisis

Model RF *Baseline fullygrown* tanpa ccp_alpha untuk menjadi tolak ukur dari model RF ccp_alpha untuk melihat dampak *pruning* pada model RF dan juga mengetahui performa maksimal yang didapat oleh ccp_alpha dalam mengklasifikasikan kualitas udara, serta untuk mengukur kompleksitas struktur pohon yang terbentuk secara alami.

Berdasarkan hasil evaluasi pada tabel 1, model RF *Baseline* menunjukkan performa yang sangat *superior* dengan akurasi sangat tinggi hingga mencapai 0.9995. Hal ini menunjukkan bahwa tanpa pruning, model mampu mengenali kelas minoritas walaupun datanya sangat sedikit. Sedangkan model RF ccp_alpha 0.001 berhasil mempertahankan akurasi total tetap tinggi (0.9963) dan dapat menurunkan total node

dari 33,832 sampai dengan 6,728 penurunan ini sampai 80% dan juga menurunkan ukuran model dari 3742.93 sampai dengan 778.43 penurunan ini sampai 79%. Hal ini membuktikan bahwa dampak *Pruning* sangat signifikan terhadap efisiensi model.

Lalu pada *Confusion Matrix* pada gambar 3 menunjukkan bahwa RF *Baseline* menghasilkan prediksi yang sangat presisi. Hampir seluruh data uji dikenali hingga ke kelas minoritas masih sangat baik dan memiliki kesalahan sangat minim, ini menunjukkan bahwa model RF *Baseline* memiliki performa sangat baik antar kelas. Sedangkan model RF ccp_alpha berbeda dengan model *baseline* yang nyaris sempurna, *Confusion Metrix* menunjukkan kalau *pruning* menurunkan performa model pada kelas minoritas terutama *hazardous* yang dimana hanya mengenali 30% data aktual dengan benar. Hal ini menandakan kalau kelas minoritas telah terpangkas karena atanya sangat sedikit sehingga dianggap menjadi noise oleh *pruning.*

Meskipun akurasi secara keseluruhan model *baseline* masih lebih tinggi dibandingkan dengan model *pruning* dan minim kesalahan, tetapi ada *trade-off* yang cukup signifikan dari sisi kompleksitas model. Untuk total nodes *baseline* hingga 33,832 node dan ukuran model hingga 3742,93 KB. Sedangkan model *pruning* dapat menurunkan total node sampai 80% dan total ukuran sampai 79% saja. Efisiensi ini dapat dicapai dikarenakan *pruning* memangkas node yang tidak berkontribusi besar pada keseluruhan akurasi.

Walaupun *pruning* berdampak positif pada efisiensi dan total akurasi model, ada kekurangan yaitu penurunan performa dalam mengenali kelas minoritas. Fenomena ini dapat dijelaskan memalui karakteristik dataset yang bisa dilihat pada gambar 2 bisa dilihat terdapat *class imbalance*. *Class imbalance* ini terjadi karena terdapat perbedaan jumlah data antar kelas khususnya pada penelitian ini pada kelas *Hazardous* dan *Very Unhealthy* yang memiliki data sangat sedikit. Dikarenakan cara kerja Cost-Complexity Pruning dengan cara memangkas nodes yang paling dianggap tidak berkontribusi besar pada akurasi model. Sehingga kelas minoritas tidak efisien karena tidak memberikan kontribusi besar pada akurasi model yang berdampak pada penurunan kemampuan model untuk mengenali kelas minoritas.

Meskipun terdapat penurunan pada prediksi kelas minortas, model RF ccp_alpha mencapai tujuan efisiensi dan menurunkan kompleksitasnya dengan mengurangi jumlah node hingga 80% dari model baseline. Dan berhasil mengecilkan ukuran model hingga 79% lebih kecil dari model baseline. Penggunaan ccp_alpha sangat efektif untuk mendapatkan efisiensi yang tinggi dan menghemat sumberdaya komputasi. Namun, bukan tanpa kekurangan yang dimana model ini menjadi kurang sensitif terhadap kelas minoritas.

Untuk penelitian selanjutnya harus lebih fokus mengidentifikasi dampak fatal dari kesalahan prediksi pada kelas minoritas dan menyesuaikan dengan teknologi mutakhir dengan mengadopsi tekni yang lebih maju untuk menangani data yang tidak seimbang.

V. KESIMPULAN

Random Forest dengan menggunakan metode Cost-Complexity Pruning terbukti dapat meningkatkan efisiensi model tanpa mengurangi akurasi secara signifikan. Nilai ccp_alpha sebesar 0.001 menghasilkan keseimbangan yang optimal dengan akurasi sebesar 99.63%, yang hanya sedikit lebih rendah dibandingkan model tanpa pruning (99.95%). Penggunaan nilai tersebut juga mampu mengurangi ukuran model sekitar 79% serta menurunkan jumlah node hingga sekitar 80%. Walaupun kinerja pada kelas minoritas menurun, model ini tetap konsisten dalam mengenali pola utama pada data kualitas udara. Hal ini menunjukkan bahwa pruning efektif untuk menyederhanakan sambil mempertahankan keandalan klasifikasi pada sebagian besar kategori AQI.

REFERENSI

[1] F. Hamami and I. A. Dahlan, “Air Quality Classification in Urban Environment using Machine Learning Approach,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 986, no. 1, p. 012004, Feb. 2022, doi: 10.1088/1755-1315/986/1/012004.

[2] F. Alzu’bi, A. Al-Rawabdeh, and A. Almagbile, “Predicting air quality using random forest: A case study in Amman-Zarqa,” *Egypt. J. Remote Sens. Space Sci.*, vol. 27, no. 3, pp. 604–613, Sep. 2024, doi: 10.1016/j.ejrs.2024.07.004.

[3] L. A. Menéndez García *et al.*, “A Method of Pruning and Random Replacing of Known Values for Comparing Missing Data Imputation Models for Incomplete Air Quality Time Series,” *Appl. Sci.*, vol. 12, no. 13, p. 6465, Jun. 2022, doi: 10.3390/app12136465.

[4] M. Diaz Resquin *et al.*, “A machine learning approach to address air quality changes during the COVID-19 lockdown in Buenos Aires, Argentina,” *Earth Syst. Sci. Data*, vol. 15, no. 1, pp. 189–209, Jan. 2023, doi: 10.5194/essd-15-189-2023.

[5] S. S. Kumbhar and D. S. Jagtap, “Air Quality Index Rating Prediction Using Machine Learning,” 2025.

[6] P. Chaturvedi, “Air Quality Prediction System Using Machine Learning Models,” *Water. Air. Soil Pollut.*, vol. 235, no. 9, p. 578, Sep. 2024, doi: 10.1007/s11270-024-07390-0.

[7] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, “An improved random forest based on the classification accuracy and correlation measurement of decision trees,” *Expert Syst. Appl.*, vol. 237, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.

[8] B. Sakthivel, K. Jayaram, N. Manikanda Devarajan, S. Mahaboob Basha, and S. Rajapriya, “Machine Learning-Based Pruning Technique for Low Power Approximate Computing,” *Comput. Syst. Sci. Eng.*, vol. 42, no. 1, pp. 397–406, 2022, doi: 10.32604/csse.2022.021637.

[9] X. Li, “The Advancements of Pruning Techniques in Deep Neural Networks:,” in *Proceedings of the 1st International Conference on Data Science and Engineering*, Singapore, Singapore: SCITEPRESS - Science and Technology Publications, 2024, pp. 177–181. doi: 10.5220/0012837000004547.

[10] A. N. Ahmad, “Cervical Cancer Prediction with Stacked Tabnet, Random Forest, and LightGBM,” vol. 1, no. 2, 2025.

[11] M. Triyani, A. Adiwijaya, and A. Aditsania, “Discrete Wavelet Transform (DWT) and Random Forest for

Cancer Detection Based on Microarray Data Classification," *J. INFOTEL*, vol. 12, no. 3, pp. 97–104, Aug. 2020, doi: 10.20895/infotel.v12i3.484.

[12] M. Dwifebri Purbolaksono, "Sentiment Analysis of Game Review in Steam Platform using Random Forest," *Int. J. Inf. Commun. Technol. IJoICT*, vol. 10, no. 2, pp. 161–169, Dec. 2024, doi: 10.21108/ijoict.v10i2.1007.

[13] R. A. Mu'allim and K. M. Lhaksmana, "Klasifikasi Multi-Label Ayat-Ayat Al-Qur'an Menggunakan Random Forest dan Word Centrality".

[14] Y. A. Muslim, B. Purnama, and B. Erfianto, "Pose Classification in Archery Sports Based on YoloV8 Using SVM and Random Forest Methods," *Int. J. Inf. Commun. Technol. IJoICT*, vol. 11, no. 1, pp. 1–12, Jun. 2025, doi: 10.21108/ijoict.v11i1.8996.

[15] I. Palupi, B. A. Wahyudi, N. Al Mamuda, and A. Shabrina, "Predicting Forest Fire Hotspots with Carbon Emission Insights Using Random Forest and Gradient Boosting Regression," *Int. J. Inf. Commun. Technol. IJoICT*, vol. 9, no. 2, pp. 137–149, Dec. 2023, doi: 10.21108/ijoict.v9i2.865.

[16] E. Rahmati, "Pruning Strategies in Random Forests: The Interplay Between Model Compression and Fairness".

[17] E. Alshdaifat, A. Al-Shdaifat, and F. Hussein, "The effect of pruning on the efficiency and effectiveness of hybrid imbalanced multiclass classification models," *Array*, vol. 28, p. 100610, Dec. 2025, doi: 10.1016/j.array.2025.100610.

[18] N. Boyko and Y. Mokryk, "Detecting Fraud in Banking Transactions with Random Forest Models," 2021.

[19] N. Hubens, M. Mancas, B. Gosselin, M. Preda, and T. Zaharia, "An Experimental Study of the Impact of Pre-training on the Pruning of a Convolutional Neural Network," in *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, 2020, pp. 1–6. doi: 10.1145/3378184.3378224.

[20] R. Liu *et al.*, "Air Quality—Meteorology Correlation Modeling Using Random Forest and Neural Network," *Sustainability*, vol. 15, no. 5, p. 4531, Mar. 2023, doi: 10.3390/su15054531.

[21] L. Chen, S. Gong, X. Shi, and M. Shang, "Dynamical Conventional Neural Network Channel Pruning by Genetic Wavelet Channel Search for Image Classification," *Front. Comput. Neurosci.*, vol. 15, p. 760554, Oct. 2021, doi: 10.3389/fncom.2021.760554.