Vol. 3, No. 3, Desember 2022, hal. 422-430

Jurnal Computer Science and Information Technology (CoSciTech)

p-ISSN: 2723-567X
e-ISSN: 2723-5661

http://ejurnal.umri.ac.id/index.php/coscitech/index

Tinjauan: Algoritme Random Forest untuk Estimasi $PM_{2.5}$ Berbasis Pengindraan Jauh

Mitra Unik*[1], Sri Nadriati[2]
Email: [1]mitraunik@umri.ac.id [2]srinadriati@gmail.com

[1]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
[2]Sistem Informasi, STMIK Dharmapala

Diterima: 6 Desember 2022 | Direvisi: 20 Desember 2022 | Disetujui: 21 Desember 2022
©2020 Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Muhammadiyah Riau, Indonesia

Abstrak

Studi ini merangkum penelitian tentang estimasi $PM_{2.5}$ menggunakan algoritme pembelajaran mesin random forest (RF), penginderaan jauh, dan keduanya. Tujuan dari tinjauan ini adalah menyajikan studi yang komprehensif untuk memfasilitasi dan menentukan batasan, luas dan kedalaman pengetahuan yang dieksplorasi untuk memperkirakan konsentrasi $PM_{2.5}$ di masa depan menggunakan RF dan pengindraan jauh. $PM_{2.5}$ merupakan parameter lingkungan atmosfer yang penting, terutama karena dampaknya terhadap kesehatan manusia dan lingkungan. Terlepas dari skala spasial-temporal, perkiraan $PM_{2.5}$ yang akurat penting untuk memahami dan menanggapi berbagai efek buruk dari polusi udara. Oleh karena itu, metode penginderaan jauh dan pembelajaran mesin dikembangkan untuk mendapatkan estimasi $PM_{2.5}$ resolusi tinggi dan mengurangi kesalahan penilaian yang disebabkan oleh dislokasi spasial. Sejak penggunaan pertama jaringan saraf (NN) untuk mempelajari hubungan kompleks AOD-$PM_{2.5}$, lebih dari 40 artikel terkait ML telah diterbitkan dalam dekade terakhir, dan lebih dari 90% di antaranya telah diterbitkan dalam lima tahun dan 75% dalam tiga tahun terakhir. Metode validasi yang mempertimbangkan pola spasial dalam validasi model ML mengungkapkan bahwa RF dan BPNN adalah yang paling populer digunakan.

Kata kunci: *random forest, pembelajaran mesin, AOD, $PM_{2.5}$, pengindraan jauh*

Overview: Random Forest Algorithm for $PM_{2.5}$ Estimation Based on Remote Sensing

Abstract

This study includes predictive research on PM2.5 using random forest (RF) machine learning algorithms, remote sensing, and both. The aim of this review is to present a comprehensive study to facilitate and define the boundaries, breadth and depth of knowledge explored for developing future concentrations of PM2.5 using RF and remote sensing. PM2.5 is an important atmospheric environmental parameter, especially because of its impact on human health and the environment. Regardless of the spatial-temporal scale, an accurate PM2.5 estimate is important for understanding and dealing with the adverse effects of air pollution. Therefore, remote sensing and machine learning methods were developed to obtain high-resolution PM2.5 estimates and reduce judgment errors caused by spatial dislocations. Since the first use of neural networks (NNs) to study the AOD-PM2.5 complex relationship, more than 40 ML-related articles have been published in the last decade, and more than 90% of them have been published within five years and 75% within the last three years. Method validation that considers spatial patterns in ML model validation reveals that RF and BPNN are the most popular ones used..

Keywords: *AOD, machine learning, random forest, remote sensing, $PM_{2.5}$*

1. PENDAHULUAN

Peningkatan penilaian paparan polusi $PM_{2.5}$ (*Particulate Matter* /partikel halus dengan diameter < 2,5 μm) dalam kesenjangan temporal dan spasial, membuat *remote sensing* dari data satelit (seperti kedalaman optik aerosol (AOD) digunakan [1]–[6].

Penggunaan sensor pemantauan kualitas udara $PM_{2.5}$ berbasis darat belum mencakup sebagian besar wilayah negara, baik karena tiadanya akses yang memadai untuk mencapai titik lokasi tertentu dan pembatasan keamanan yang dapat terjadi di zona konflik. Oleh karenanya cakupan temporal dan spasial oleh sensor darat menjadi tidak cukup efektif untuk dapat mengukur secara keseluruhan wilayah dengan akurat. Tidak seperti polutan udara gas, $PM_{2.5}$ merupakan campuran komplek bahan kimia yang dihasilkan oleh berbagai sumber emisi langsung serta memiliki reaksi foto kimia di atmosfer, seperti *Amonium sulfate* ($SO4^{2-}$), *Amonium Nitrat* (NO^{3-}), *organic carbon* (OC), *elemental carbon* (EC) [7]. Estimasi $PM_{2.5}$ berbasis satelit telah banyak digunakan untuk menilai dampak kesehatan terkait paparan PM. Nilai Ambang Batas (NAB) yang diperbolehkan untuk konsentrasi $PM_{2.5}$ dalam udara ambient menurut Badan Meteorologi Klimatologi dan Geofisika (BMKG) Indonesia adalah NAB $PM_{2.5} \leq 65$ µgram/m^3. Studi mengidentifikasi terpapar $PM_{2.5}$ secara langsung dan jangka panjang dengan konsentrasi melebihi NAB di udara, secara signifikan dapat berdampak pada perubahan iklim, degradasi visibilitas, kesehatan masyarakat, gangguan ekosistem, sosial, ekologi serta ekonomi [8]–[13].

Sejak 1990-an, algoritme pembelajaran mesin atau *machine learning* (ML) telah digunakan untuk otomatisasi pengumpulan data, memahami data *remote sensing*, dan mengekstrak pola data yang sebelumnya tidak terlihat [14]. Sejak studi pertama yang mengadopsi *Neural Network* (NN) untuk mengatasi hubungan kompleks AOD–$PM_{2.5}$ [15], ML pada bidang *remote sensing* telah muncul bersama teknologi berbasis platform satelit yang dianggap sebagai area *big data* dan komputasi dengan performa tinggi. ML digunakan untuk mengungkap, mengukur dan memahami proses data secara intensif dan memungkinkan untuk memiliki kemampuan belajar tanpa harus diprogram secara ketat [16]. ML memungkinkan untuk memeriksa secara non-parametrik hubungan antara prediktor konsentrasi polutan dan konsentrasi polutan yang diukur [17], [18].

Penelitian oleh Yazdi (2020) melakukan estimasi konsentrasi $PM_{2.5}$ menggunakan model ansambel (RF, GBM, NN, KNN) dengan data spasial dan temporal pada lokasi penelitian, serta menyimpulkan bahwa model RF mengungguli model ML lainnya [5]. Rincian model RF telah diberikan di studi lain [19]. Umumnya, RF adalah metode ensemble berdasarkan pohon keputusan dari *supervised learning* yang memiliki interpretabilitas tinggi. RF telah menjadi bidang penelitian yang populer karena mudah dipahami dan kuat [20]. Selain itu studi [21] menyatakan bahwa hasil validasi silang $PM_{2.5}$ menunjukkan bahwa kinerja model tidak sensitif terhadap perubahan resolusi spasial untuk model LME dan LUR, sebaliknya model RF dapat membuat estimasi $PM_{2.5}$ yang lebih akurat dengan resolusi spasial AOD yang lebih baik.

Karena banyaknya singkatan yang digunakan, Tabel 1 dan 2 mencantumkan beberapa daftar singkatan yang muncul dalam studi ini.

Tabel 1. Singkatan untuk algoritme machine learning pada artikel ini

Singkatan	Algoritme
RF	Random Forest
GBM	Gradient Boosting Machine
BPNN	Back-propagation NN
KNN	K-Nearest Neighbor
LME	Linear Mixed Effect
LUR	Land-use Regression

Tabel 2. Singkatan untuk komponen remote sensing

Abbreviation	Remote Sensing Components
AOD	Aerosol Optical Depth
GEOS-CHEM	Goddard Earth Observing System- chemical
MAIAC	Multi-Angle Implementation of Atmospheric Correction
MISR	Multi-Angle Imaging Spectroradiometer
MODIS	Moderate Resolution Imaging Spectrometer
TOA	Top of Atmosphere

Fokus studi ini untuk meninjau secara sistematis penerapan algoritme *RF* dalam estimasi konsentrasi polutan $PM_{2.5}$ berbasis remote sensing berdasarkan tinjauan studi yang telah ada. Penulis menganggap artikel jurnal dan buku yang diterbitkan oleh penerbit terkemuka sebagai penelitian berkualitas tinggi, dan oleh karena itu memasukkan nya ke dalam rangkuman. Survei ini ditujukan untuk praktisi dan peneliti dengan tujuan menyajikan kajian secara komprehensif untuk memandu peneliti lapangan dalam aplikasi mereka ke bidang penelitian masing-masing dan bidang terkait, memajukan pemahaman terkait potensi penggunaan algoritme RF berbasis *remote sensing* untuk estimasi $PM_{2.5}$, serta untuk mendorong dan mengidentifikasi batas, keluasan, dan kedalaman pengetahuan yang dapat dieksplorasi oleh peneliti masa depan.

2. METODE PENELITIAN

Studi ini dilakukan secara sistematis dengan panduan karya [22]. Studi pengetahuan terkait penggunaan algoritme RF, *remote sensing*, dan keduanya untuk estimasi indeks polutan $PM_{2.5}$ disertakan. Tinjauan tentang topik tertentu dikeluarkan dari penelitian ini.

2.1. *Proses Pencarian*

Sejalan dengan multidisipliner topik tinjauan literatur beberapa disiplin ilmu lain, mulai dari ilmu komputer, kehutanan, pengindraan jarak jauh, atmosfer dan kebencanaan disertakan. Ringkasan statistik mencatat jumlah artikel yang diambil dari proses survei yang didapat, diabaikan, dan disertakan yang kemudian disajikan dalam bentuk alur diagram pada gambar 1. Pada pencarian awal dihasilkan 137 artikel. Setelah penyaringan awal, penilaian kelayakan, 25 artikel dimasukkan ke dalam survei serta 48 artikel lainnya untuk membantu dalam penyusunan penelitian ini.

Gambar 1. Diagram alir proses penelitian

Identifikasi literatur. Pencarian diawali dengan mengidentifikasi serangkaian kata kunci untuk menelusuri *Carrot²Project*, Google *Cendikia*, *IEEXplore*, *Web of Science* dan perpustakaan digital baik dari jurnal-jurnal Nasional maupun Internasional. Setiap manuskrip, relevansi awal ditentukan berdasarkan judul. Dari judul, jika dianggap membahas topik penelitian, maka referensi lengkap, termasuk penulis, tahun, judul, dan abstrak diperoleh untuk evaluasi lebih lanjut. Setelah set awal terbentuk, artikel yang mengutipnya atau yang dikutip dalam set ini dicari. Selanjutnya dokumen yang teridentifikasi ganda dihilangkan. Kemajuan teknologi mengubah metode pengarsipan dan pengambilan informasi, tahun publikasi dibatasi sehingga dapat membangun tinjauan pada literatur terbaru dengan mempertimbangkan pencarian dan sintesis informasi di area digital.

Penyaringan. Artikel yang ditemukan oleh mesin pencarian disaring dengan memberikan label pada artikel yang relevan atau tidak pada penelitian ini. penyaringan awal ini dilakukan dengan membaca abstrak untuk memutuskan relevansinya dengan topik penelitian. Setiap artikel yang relevan dengan ruang lingkup penelitian akan disimpan.

Penilaian Kelayakan. Artikel yang telah melewati tahap penyaringan awal, dinilai kelayakannya dengan membaca temuan utama, penggunaan data serta saran pengerjaan penelitian selanjutnya. *Scimago Journal & Country Rank* (https://www.scimagojr.com/) digunakan untuk memeriksa rangking dari artikel yang di sertakan. Keseluruhan (100%) referensi yang digunakan dalam penyusunan penelitian adalah artikel dari jurnal Internasional ter index scopus . Dari Total 24 artikel inti yang dirangkum sebanyak 100% artikel memiliki rangking jurnal *Quartile* 1 (Q1).

Penyertaan: Literatur yang bersesuaian dengan tujuan penelitian ini ditetapkan yaitu penggunaan *ML* dan pengindraan jarak jauh untuk estimasi indeks polutan $PM_{2.5}$. Namun, beberapa artikel yang beririsan serta dapat memberikan wawasan tentang berbagai teknologi pengindraan jarak jauh pada polutan, polusi udara serta penelitian lainnya secara spesifik juga ditambahkan tanpa dibatasi oleh tahun periode publikasi.

3. HASIL DAN PEMBAHASAN

3.1. Variabel prediktor estimasi $PM_{2.5}$

a *Satelit* AOD

AOD merupakan parameter satelit paling dasar dalam pengukuran PM2.5 yaitu sejauh mana aerosol mencegah transmisi cahaya dengan menyerap atau menyebarkan cahaya melalui seluruh kolom vertikal atmosfer dari tanah ke sensor satelit [23]. Untuk penilaian AOD dan PM, literatur menunjukkan bahwa terdapat hubungan linier antara kedua variabel tersebut. Pendekatan empiris ini paling dikenal untuk membedakan konsentrasi PM dari dataset satelit, di mana tingkat estimasi dapat ditentukan dari tingkat korelasi antara pengukuran PM dan AOD satelit terkait dari variabel yang sama [24], [25]. Pengukuran ketebalan optik aerosol didasarkan pada fakta bahwa partikel mengubah cara atmosfer memantulkan dan menyerap cahaya tampak dan inframerah [26]. Korelasi antara AOD dan $PM_{2.5}$ menunjukkan ketergantungan yang cukup menjanjikan karena AOD mewakili jumlah cahaya yang dihilangkan dari balok melalui penghamburan atau penyerapan partikel aerosol di sepanjang jalur optik. Dengan kata lain AOD adalah variabel yang mencakup perubahan $PM_{2.5}$, yang

merupakan hasil kombinasi komprehensif dari emisi, reaksi kimia, dan lainnya. Namun, masih ada tiga perbedaan utama antara data AOD dan $PM_{2.5}$. Pertama, AOD adalah nilai tanpa unit yang mencerminkan efek pemadaman cahaya total aerosol di kolom, sedangkan $PM_{2.5}$ adalah konsentrasi massa dipermukaan tanah. Kedua, dengan adanya kelembaban, partikel yang larut dalam air akan menjadi lebih besar melalui proses penyerapan air, sehingga mempengaruhi kemampuan pemadaman cahaya aerosol. Ketiga, $PM_{2.5}$ hanya sebagian dari aerosol dengan diameter sama dengan atau kurang dari 2,5 g/m3, tetapi ini tidak berlaku untuk semua aerosol. Mengenai perbedaan utama ini, data tambahan yang berisi beberapa informasi akan membantu untuk memperkirakan $PM_{2.5}$ dari AOD, seperti distribusi vertikal aerosol, kelembaban relatif lingkungan atmosfer, komposisi kimia dan distribusi ukuran (rasio partikel halus dan kasar) [27].

Studi hubungan antara jumlah PM jarak dekat permukaan tanah dan koefisien kepunahan cahaya matahari telah dilaporkan dalam literatur [25]. Instrumen satelit pada dasarnya tidak mengukur $PM_{2.5}$, namun mengamati AOD. Produk AOD yang paling umum digunakan berasal dari Satelit MODIS dan memiliki dua resolusi: 10 km dan 3 km. Para peneliti juga telah berhasil menggunakan AOD dalam Resolusi 1 km (misalnya MAIAC), yang telah banyak digunakan untuk studi polusi udara dan dampak kesehatan [1], [2], [5], [18], [28]–[41]. Hubungan $PM_{2.5}$-AOD berkorelasi dengan konsentrasi $PM_{2.5}$ hasil pengukuran stasiun darat. Dengan asumsi langit bebas awan, *Boundary Layer Height* (BLH) yang tercampur dengan baik tanpa aerosol di atasnya, dan aerosol yang memiliki sifat optik yang serupa, AOD dapat ditulis [42]:

$$AOD = PM2.5\ x\ BLH\ x\ f(RH)\ x\ \frac{3Qext.dry}{4p\ reff} \qquad (1)$$

Dimana BLH adalah Ketinggian lapisan batas, dan fungsi f(RH) menggambarkan peningkatan penampang pemadaman aerosol dengan kelembaban relatif, dan sebanding dengan luas proyeksi rata-rata partikel aerosol. adalah kerapatan massa aerosol (g.m-3) Qext,dry efisiensi kepunahan dalam kondisi kering, reff= Radius efektif partikel and S adalah Efisiensi kepunahan spesifik (m2.g-1) dari aerosol pada ambient *relative humidity* (RH).

b Penggunaan Lahan

Berbagai jenis tutupan lahan mempengaruhi, pada tingkat yang berbeda-beda, perubahan atmosfer yang lebih rendah, radiasi permukaan dan keseimbangan energi yang dihasilkan nya, dan secara tidak langsung mempengaruhi perubahan iklim. Pemanasan atmosfer non-adiabatic ini tidak hanya mendorong pergerakan molekul atmosfer tetapi juga transfer dan difusi partikel halus. Ketergantungan variabel penggunaan lahan pada AOD atau $PM_{2.5}$ diperiksa dan hasilnya menunjukkan bahwa padang rumput, semak belukar, badan air dan permukaan buatan memiliki ketergantungan positif pada AOD (ketergantungan parsial maksimum sekitar 0,63) dan tidak signifikan terhadap $PM_{2.5}$. Hal di atas menunjukkan bahwa tutupan lahan memiliki dampak langsung terhadap AOD dan dampak tidak langsung terhadap $PM_{2.5}$. Selain itu, tutupan lahan tidak berubah terhadap waktu, yang merupakan alasan penting untuk korelasi yang lemah dengan $PM_{2.5}$. Han et al (2020) menunjukkan bahwa sebaran spasial pencemar berkaitan erat dengan penataan ruang kawasan industri dan lokasi perusahaan yang menimbulkan emisi pencemar udara. Ruang hijau dapat mengurangi polusi, dan kontribusi dari emisi transportasi lebih kecil dari emisi industri [43]. NDVI, pengukuran satelit dan pemantauan kerapatan indeks kehijauan vegetasi (tanaman), dapat membatasi kemampuan ruang hijau untuk menjelaskan perubahan $PM_{2.5}$ [44] dan digunakan dalam model pembelajaran mesin. Ketinggian diketahui mempengaruhi distribusi $PM_{2.5}$ karena $PM_{2.5}$ sulit dijangkau pada suatu titik di atas permukaan bumi dari permukaan laut (elevasi) yang lebih tinggi [45], mengakibatkan perbedaan besar antara daerah yang lebih rendah dan lebih tinggi [43]. Han *et al* (2020) mengungkapkan bahwa sebaran spasial polutan berkaitan erat dengan tata ruang lahan industri dan lokasi perusahaan penyebab emisi polutan udara. Ruang hijau dapat mengurangi polusi, dan kontribusi emisi transportasi lebih rendah dibandingkan kontribusi emisi industri [44]. NDVI yang merupakan pengukuran dan pemantauan kerapatan indeks kehijauan vegetasi (tanaman) berbasis satelit, dapat membatasi kemampuan ruang hijau untuk menjelaskan perubahan $PM_{2.5}$ [46] serta digunakan dalam model *machine learning*. Ketinggian diketahui mempengaruhi distribusi $PM_{2.5}$, karena $PM_{2.5}$ sulit dijangkau pada sebuah titik di atas muka bumi dari permukaan laut (elevasi) yang lebih tinggi [45], yang menghasilkan perbedaan besar antara wilayah yang lebih rendah dan yang lebih tinggi.

c Meteorologi

Sejalan dengan penelitian sebelumnya, hubungan $PM_{2.5}$-AOD dapat membentuk fungsi multivariate terkait dengan berbagai faktor meteorologi dan terrestrial yang mempengaruhinya [47], [48]. Variabel meteorologi yang biasa digunakan adalah kelembaban relatif (RH), suhu (TEMP), angin u/v, tekanan permukaan (SP), dan arah angin (WD). Studi observational lainnya memberikan bukti bahwa *Planetary Boundary Layer Height* (PBLH) berdampak pada hubungan $PM_{2.5}$ dan AOD [49]. Semakin tinggi PBLH, semakin tinggi AOD tetapi semakin rendah $PM_{2.5}$ [50]. RH mengubah partikel aerosol, yang kemudian mempengaruhi AOD, yaitu dengan meningkatnya kelembaban, demikian juga besarnya higroskopisitas (kemampuan suatu zat untuk mengambil molekul air dari sekitarnya, baik dengan penyerapan atau adsorpsi) dan partikel aerosol [51]. Selanjutnya penguapan memiliki korelasi positif yang kuat dengan suhu (R> 0,6), dan korelasi negatif yang kuat dengan kelembaban relatif (R< -0,6) [43].

d Geografis

Variabel yang berhubungan dengan jumlah penduduk *Night Light* (NLT), kepadatan penduduk, kepadatan jaringan jalan, tinggi dan jumlah bangunan, dan lain-lain Dapat menunjukkan tingkat aglomerasi penduduk dan urbanisasi, dengan urbanisasi berkembang ke skala pembangunan industri [5]. Misalnya, batu bara dan emisi kendaraan dapat menjadi sumber utama kabut asap, karena komposisi partikel jelaga halus yang lebih besar yang dilepaskan menjadi predisposisi pengukuran AOD dan $PM_{2.5}$ yang lebih tinggi dalam MODIS [18]. Studi oleh Liu (2018) menunjukkan bahwa NLT memiliki peningkatan MSE: sekitar 30 n plot dengan ketergantungan parsial pada $PM_{2.5}$ umumnya meningkat perlahan seiring dengan peningkatan NLT [52]. Hal ini menunjukkan bahwa NLT tinggi cenderung mewakili daerah padat penduduk dan pabrik masih Namun, dampak emisi dalam waktu singkat tidak berpengaruh signifikan terhadap tingginya $PM_{2.5}$. Zhang (2018) menunjukkan bahwa kepadatan penduduk berkorelasi positif secara signifikan dengan AOD, dengan konsentrasi $PM_{2.5}$ di meningkat tajam mendekati kepadatan penduduk = 6 (orang/KM) dan kemudian meningkat perlahan Pola ini menunjukkan kontribusi kepadatan penduduk terhadap konsentrasi $PM_{2.5}$, yang dapat naik di atas batas polutan karena tingginya tingkat aktivitas manusia.

3.2 Rangkuman Studi Relevan

Bagian ini memberikan gambaran umum tentang teknik ML utama dan aplikasinya dalam estimasi $PM_{2.5}$. Tabel 3 merangkum studi penggunaan algoritme *RF* berbasis *remote sensing* untuk estimasi $PM_{2.5}$ pada 42 artikel yang dipilih. Studi ini mengeksplorasi algoritme RF yang digunakan para peneliti terdahulu untuk estimasi $PM_{2.5}$-AOD. Selanjutnya ditampilkan hasil *Cross Validation* (CV) dari model yang digunakan.

Tabel 3. Rangkuman studi estimasi $PM_{2.5}$ berbasis remote sensing menggunakan algoritme RF.

Ref	Area -periode studi	Resolusi AOD	CV R^2
[29]	Italia (2013-2015)	1KM Aqua MAIAC AOD	10-fold CV 0.81
[53]	Tehran, Iran (2015-2018)	10 and 3 Km Terra collection 6 L2 DT AOD (MOD04_L2 and MOD04_3k)	10-fold CV 0.66, 0.67 and 0.63 for RF XGBoost and DL model, respectively (after including 3km AOD Product in models)
[1]	Guangdong – HongKong – Macao Greater Bay Area, China (2016-2018)	1 km Terra and Aqua MAIAC AOD (MCD19A2)	10-fold CV - 0.937, 0.905 and 0.884 for 2016, 2017 and 2018,
[33]	Great Britain, United Kingdom (2008-2018)	1 km Terra and Aqua MAIAC AOD (MCD19A2	10-fold CV - 0.767
[5]	London (2005-2013)	1 Km MODIS	10-fold CV - 0.828
[54]	USA (2011)	10 km Aqua collection 6 L2 DT AOD (MYD04_L2)	10-fold CV - 0.80
[55]	Cincinnati, OH, USA (2000-2015)	3 km Terra and Aqua collection 6 L2 DT AOD (MOD04_L2 And MYD04_L2)	10-fold CV - 0,91
[34]	Iraq and Kuwait (2001-2018)	1 Km Terra and Aqua - (MAIAC) AOD data (MCD19A2)	10 fold CV - 0,71
[48]	Indo-Gangetic Plain- Pakistan across Northern India to the east of the Bay of Bengal and Bangladesh (Juli 2018- Juni 2019)	IKM Modis MCD19A2	CV 0.87
[56]	California (2005-2014)	4.4 Km MISR	4-fold CV 0.72, 0.70, 0.68 and 0.70 for sulfate, nitrate, OCand EC, respectively
[52]	United States (2000-2013)	-	(5-fold CV – 0.720
[35]	China (2018)	IKM HIMAWARI 8	CV 0,80
[57]	Japan (2011-2016)	1 Km MAIAC	10-fold CV - 0.86
[36]	California (2005-2014)	4.4 Km MISR	4-fold CV 0.72, 0.70, 0.68 and 0.70 for sulfate, nitrate, OCand EC, respectively
[58]	Beijing-Tianjin-Hebei (2010–2016)	3Km Aqua MODIS level 2	10-fold CV-0.86
[37]	China (2017)	1Km MODIS MAIAC AOD	CV- 0.74
[38]	North China Plain area, China (2013-2016)	1 Km Terra and Aqua MAIAC AOD	10-fold CV- 0.88
[59]	South Korea, eastern China, Taiwan and Japan (2016)	10km MODIS, GEOS-Chem	0.93(hourly),0.93[b]
[39]	China (2004–2012)	1 Km MODIS	CV - 0.54–0.88
[2]	China (2015-2016)	1 Km MODIS	10-fold CV- 0.85

[60]	China (2018)	6Km daily Level-2 VIIRS AERDB	CV- 0.87
[3]	Central and Eastern China (2016)	Himawari 8 CEMC, AHI	10-fold CV- 0.85,
[61]	mainland China, Hong Kong and Taiwan (2013-2017)	10Km (MODIS) Collection 6 level 2	10-fold CV-0.89.
[62]	Beijing–Tianjin–Hebei (2019-2020)	1Km TOAR, AOD,	R2, RMSE, and MAE- 0.91, 11.60, and 7.34

3.3. Ikhtiar Penggunaan Algoritme RF

RF diperkenalkan oleh [63], merupakan metode ansambel *supervised learning* yang bertindak berdasarkan pohon keputusan. RF dapat digunakan untuk klasifikasi dan regresi, dan sangat fleksibel dan cepat. Untuk melakukan analisis RF, perlu untuk menyesuaikan hyper parameter model. Pencarian grid untuk optimasi kinerja model dilakukan dengan teknik validasi silang 10 kali lipat berdasarkan metrik R^2. Tabel 4 menunjukkan rentang hyper parameter RF dan nilai optimal yang terdeteksi oleh pencarian grid.

Tabel 4. Hyper parameter pencarian grid RF.

Parameter	Range	Optimum Value
n_estimators max_features min_samples_split bootstrap	70 to 150 [Auto, SQRT, Log2] [2,4,8] [True, False]	130 SQRT 2 False

Model berbasis pohon menggunakan beberapa strategi ensemble untuk menciptakan pembelajaran yang kuat (*strong model*) dari beberapa pembelajaran dasar/komponen, dengan strategi menentukan kinerja model. Bagging lebih fokus pada pengurangan varian model, dan efek nya lebih terasa pada data latih yang rentan terhadap bias sampling. RF, perpanjangan dari bagging, terdiri dari beberapa pohon keputusan, dan setiap pohon keputusan adalah unit regresi (untuk memperkirakan masalah). Sebuah sampel dari set pelatihan lengkap dipilih secara acak untuk membangun beberapa pohon keputusan. Keluaran rata-rata dari pohon keputusan digunakan untuk melatih model. Struktur model ditentukan dengan menyesuaikan jumlah dan kedalaman pohon keputusan dengan akurasi yang diinginkan [20]. RF telah menjadi bidang penelitian yang populer karena mudah dipahami dan kuat. Kerangka umum untuk memperkirakan konsentrasi $PM_{2.5}$ dalam RF adalah sebagai berikut:

$$f(x) = \sum_{z=1}^{z} C_z I(x \in R_z) \quad (2)$$

$$\hat{c}_z = \text{mean}(y_i | x \in R_z) \quad (3)$$

dimana, f(x) mewakili pohon regresi, nilai fungsinya adalah $PM_{2.5}$. (xi, yi) adalah sampel untuk i = 1, 2,..., N dari daerah Z (R1, R2,… , Rz), cz adalah estimasi terbaik rata-rata keluaran sampel dari kumpulan data. Strategi pembagian RF dapat dinyatakan sebagai berikut:

$$z_1(m,n) = \{X | X_j \leq N\} \& Z_2(m,n)$$

$$= \{\{X | X_j > N\} \min_{m,n} \left[\min_{m,n} \sum_{x_{i \in R_1(m,n)}} (y_i - C_1)^2 + \min_{m,n} \sum_{x_{i \in R_2(m,n)}} (y_i - C_2)^2\right] \hat{c}_z = \text{mean}(y_i | x_i \in R_1(m,n)) \& \hat{c}_2 = \text{mean}(y_i | x_i \in R_2(m,n)) \quad (4)$$

di mana, m adalah variabel pemisah, n adalah titik pemisah.

Stein *et. al* (2020) memperkirakan konsentrasi $PM_{2.5}$ dan PM_{10} di Italia (2013-2015) dengan data satelit dan pendekatan pemodelan ensambel (GAM+RF, LME dan XGBoost). Tahapan pemodelan berikut digunakan: (1) RF digunakan untuk menghitung nilai AOD berbasis satelit yang hilang melalui spatiotemporal penggunaan lahan yang menggabungkan data AOD dari model gabungan atmosfer. (2) Perkiraan PM harian dihasilkan menggunakan empat pendekatan pemodelan: LME, RF, XGBoost dan *Chemical Transport Model* (CTM), model kualitas udara regional yang fleksibel. MAIAC-AOD yang telah terisi sebelumnya bersama dengan prediktor spasial dan temporal tambahan digunakan sebagai input untuk tiga model pertama. (3) Model ansambel GAM kemudian digunakan untuk menggabungkan perkiraan dari keempat model dengan memungkinkan bobot masing-masing model bervariasi menurut spasial temporal. Ukuran kinerja $PM_{2.5}$ untuk tahun 2013-2015 [29]. Hasil validasi RF dan GAM dapat terlihat pada Tabel 5 berikut.

Zamani Joharestani *et al* (2019) Konfigurasi optimal dari model RF dicapai dengan menggunakan fungsi pencarian grid bawaan Python dengan teknik CV 10 *fold*. Untuk ukuran kumpulan data, kinerja terbaik diamati ketika AOD03 dikeluarkan dari fungsi input model. Metrik prediksi menunjukkan R^2 = 0,78 (R = 0,88), MAE = 10,8 g/m3 dan RMSE = 14,54 g/m3. Prediksi Scatter Points vs. terdistribusi di sekitar garis referensi y = x dan kerapatannya poin lebih dekat ke garis referensi. Ini menunjukkan prediksi $PM_{2.5}$ yang masuk akal. Mengecualikan kedua AOD menunjukkan kinerja yang hampir sama dengan pengujian dengan AOD10 [53].

Tabel 5. Hasil Koefisien Determinan (R^2) dan Root Mean Squared Error (RMSE) algoritme RF dan ansambel GAM [29].

		$PM_{2.5}$		PM_{10}	
Tahun	Evaluasi	GAM	RF	GAM	RF
2013	R^2	0.79	0.77	0.73	0.7
	RMSE	6.56	6.84	9.42	9.91
2014	R^2	0.79	0.77	0.74	0.73
	RMSE	5.29	5.49	8.54	8.74
2015	R^2	0.81	0.79	0.76	0.74
	RMSE	6.34	6.62	8.95	9.32

4. DISKUSI

Sebagian besar metode saat ini dapat bekerja dengan cukup baik untuk estimasi $PM_{2.5}$ bulanan, sementara $PM_{2.5}$ harian dan bahkan $PM_{2.5}$ per jam juga dapat dicapai, terutama dengan model ML. Estimasi ini umumnya berkinerja baik, dengan nilai R2 lebih besar dari 0,80 dan mean square error (RSME) kurang dari 20 g/m. Namun, masih banyak kesulitan dan tantangan. Pertama, missing value dari AOD adalah salah satu masalah utama, terutama untuk estimasi data resolusi spasial tinggi [64]. Beberapa kemajuan telah hadir untuk mengatasi masalah ini, seperti menggunakan reflektans di TOA sebagai pengganti AOD [62] , menggunakan RF untuk memperlakukan jaringan AOD dan non AOD secara terpisah, dan kemudian menggabungkan dua RF menjadi satu RF [55] serta menginterpolasi AOD berdasarkan data, seperti keluaran CTM [65] atau data awan satelit maupun salju [65]. Namun, pekerjaan ini sering didasarkan pada perhitungan matematis murni, yang berbeda dari produk AOD biasa, yang didasarkan pada model fisik. Selain itu, banyak pendekatan hanya diteliti dalam skala kecil [28], [55], [62]. Kedua, observatorium $PM_{2.5}$ di permukaan tanah umumnya tidak merata dan membuat perkiraan $PM_{2.5}$ kurang aman di area dengan lebih sedikit stasiun $PM_{2.5}$ dibandingkan dengan lebih banyak stasiun. Ini juga menimbulkan masalah lain: validasi silang berbasis situs yang umum digunakan dapat menjadi bias karena data pengamatan yang digunakan untuk pelatihan dan validasi terkonsentrasi di area dengan lebih banyak stasiun, membuat hubungan AOD-$PM_{2.5}$ yang divalidasi kurang efektif di area dengan lebih sedikit stasiun $PM_{2.5}$ [64]. Ketiga, rendahnya konsistensi variabel independen dan dependent dalam lingkungan atmosfer yang sama untuk menghasilkan hasil estimasi, sehingga ketersediaan data yang berbeda sebagai prediktor cukup penting untuk meningkatkan kepercayaan terhadap hasil yang diperoleh, meskipun di sisi lain ketersediaan data menjadi tantangan. di beberapa wilayah penelitian, namun gagasan untuk memprediksi variasi temporal $PM_{2.5}$ berdasarkan AOD satelit perlu diselidiki lebih lanjut di masa mendatang.

5. KESIMPULAN

Banyak metode baru dan perbaikan metode yang ada telah diterapkan pada estimasi $PM_{2.5}$ berdasarkan data satelit turunan yang digunakan. Untuk sebagian besar algoritme saat ini, koefisien korelasi pemulihan yang dilaporkan dengan pengamatan lapangan bisa lebih besar dari 0,8. Namun, bisa juga dipengaruhi oleh metode validasi yang digunakan. Ada beberapa pendekatan CV yang diterapkan untuk memvalidasi model estimasi $PM_{2.5}$ berbasis satelit. Metode validasi yang mempertimbangkan pola spasial dalam validasi model ML mengungkapkan bahwa RF dan BPNN adalah yang paling populer digunakan.

DAFTAR PUSTAKA

[1] W. Chen *et al.*, "Estimating PM2.5 with high-resolution 1-km AOD data and an improved machine learning model over Shenzhen, China," *Science of the Total Environment*, vol. 746, p. 141093, 2020, doi: 10.1016/j.scitotenv.2020.141093.

[2] J. Wei *et al.*, "Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach," *Remote Sens Environ*, vol. 231, no. April, 2019, doi: 10.1016/j.rse.2019.111221.

[3] J. Chen, J. Yin, L. Zang, T. Zhang, and M. Zhao, "Stacking machine learning model for estimating hourly PM2.5 in China based on Himawari8 aerosol optical depth data," *Science of the Total Environment*, vol. 697, p. 134021, 2019, doi: 10.1016/j.scitotenv.2019.134021.

[4] X. Meng, J. L. Hand, B. A. Schichtel, and Y. Liu, "Space-time trends of PM2.5 constituents in the conterminous United States estimated by a machine learning approach, 2005–2015," *Environmental International*, vol. 121, no. August, pp. 1137–1147, 2018, doi: 10.1016/j.envint.2018.10.029.

[5] M. D. Yazdi *et al.*, "Predicting fine particulate matter (PM2.5) in the greater london area: An ensemble approach using machine learning methods," *Remote Sens (Basel)*, vol. 12, no. 6, 2020, doi: 10.3390/rs12060914.

[6] Y. A. Aliyu and J. O. Botai, "Appraising city-scale pollution monitoring capabilities of multi-satellite datasets using portable pollutant monitors," *Atmos Environ*, vol. 179, no. November 2017, pp. 239–249, 2018, doi: 10.1016/j.atmosenv.2018.02.034.

[7] U.S. EPA, "The Particle Pollution Report Current Understanding of Air Quality and Emissions through 2003," *Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and standards; Emissions, Monitoring and Analysis Division*, p. 28, 2004.

[8] L. Yang, C. Li, and X. Tang, "The Impact of PM2.5 on the Host Defense of Respiratory System," *Front Cell Dev Biol*, vol. 8, no. March, pp. 1–9, 2020, doi: 10.3389/fcell.2020.00091.

[9] T. Schikowski and H. Altuğ, "The role of air pollution in cognitive impairment and decline," *Neurochem Int*, vol. 136, no. February, p. 104708, 2020, doi: 10.1016/j.neuint.2020.104708.

[10] Greenpeace International, "The Final Countdown: Now or never to reform the palm oil industry," *Greenpeace*, 2018. https://www.greenpeace.org/international/publication/18455/the-final-countdown-forests-indonesia-palm-oil/ (accessed Feb. 20, 2020).

[11] L. Syaufina, "Forest and land fires in Indonesia: Assessment and mitigation," in *Integrating Disaster Science and Management: Global Case Studies in Mitigation and Recovery*, Bogor: Elsevier Inc., 2018, pp. 109–121. doi: 10.1016/B978-0-12-812056-9.00008-7.

[12] M. Gross, "Learning to live with landscape fires," *Current Biology*, vol. 25, no. 16, pp. R693–R696, 2015, doi: 10.1016/j.cub.2015.07.069.

[13] J.-S. Tan-Soo and S. K. Pattanayak, "Seeking natural capital projects: Forest fires, haze, and early-life exposure in Indonesia," *Proceedings of the National Academy of Sciences*, vol. 116, no. 12, pp. 5239 LP – 5245, Mar. 2019, doi: 10.1073/pnas.1802876116.

[14] K. J. Bergen, P. A. Johnson, M. V. De Hoop, and G. C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," *Science (1979)*, vol. 363, no. 6433, 2019, doi: 10.1126/science.aau0323.

[15] P. Gupta and S. A. Christopher, "Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach," *Journal of Geophysical Research Atmospheres*, vol. 114, no. 20, pp. 1–14, 2009, doi: 10.1029/2008JD011497.

[16] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *Int J Inf Manage*, vol. 48, no. January, pp. 63–71, 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[17] Q. Di *et al.*, "Assessing NO(2) Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging.," *Environ Sci Technol*, vol. 54, no. 3, pp. 1372–1384, Feb. 2020, doi: 10.1021/acs.est.9b03358.

[18] Q. Di *et al.*, "An ensemble-based model of PM(2.5) concentration across the contiguous United States with high spatiotemporal resolution.," *Environ Int*, vol. 130, p. 104909, Sep. 2019, doi: 10.1016/j.envint.2019.104909.

[19] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[20] J. Liu, F. Weng, and Z. Li, "Satellite-based PM2.5 estimation directly from reflectance at the top of the atmosphere using a machine learning algorithm," *Atmos Environ*, vol. 208, pp. 113–122, 2019, doi: https://doi.org/10.1016/j.atmosenv.2019.04.002.

[21] H. Bai, Y. Shi, M. Seong, W. Gao, and Y. Li, "Influence of Spatial Resolution on Satellite-Based PM2.5 Estimation: Implications for Health Assessment," *Remote Sens (Basel)*, vol. 14, no. 12, 2022, doi: 10.3390/rs14122933.

[22] Y. Xiao and M. Watson, "Guidance on Conducting a Systematic Literature Review," *J Plan Educ Res*, vol. 39, no. 1, pp. 93–112, 2019, doi: 10.1177/0739456X17723971.

[23] K. N. Liou, *An Introduction to Atmospheric Radiation*, 2nd ed. San Diego: Academic Press, 2002.

[24] X. Q. Yap and M. Hashim, "A robust calibration approach for PM10 prediction from MODIS aerosol optical depth," *Atmos Chem Phys*, vol. 13, no. 6, pp. 3517–3526, 2013, doi: 10.5194/acp-13-3517-2013.

[25] K. Schäfer, A. Harbusch, S. Emeis, P. Koepke, and M. Wiegner, "Correlation of aerosol mass near the ground with aerosol optical depth during two seasons in Munich," *Atmos Environ*, vol. 42, no. 18, pp. 4036–4046, 2008, doi: 10.1016/j.atmosenv.2008.01.060.

[26] NEO, "Aerosol Optical Depth," *NASA EARTH OBSERVATION*, 2020. https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MODAL2_M_AER_OD (accessed Jul. 08, 2020).

[27] C. Zheng *et al.*, "Analysis of influential factors for the relationship between PM$_{2.5}$ and AOD in Beijing," *Atmos Chem Phys*, vol. 17, no. 21, pp. 13473–13489, 2017, doi: 10.5194/acp-17-13473-2017.

[28] B. Chen *et al.*, "An interpretable self-adaptive deep neural network for estimating daily spatially-continuous PM2.5 concentrations across China," *Science of the Total Environment*, vol. 768, p. 144724, 2021, doi: 10.1016/j.scitotenv.2020.144724.

[29] A. Shtein *et al.*, "Estimating Daily PM2.5 and PM10 over Italy Using an Ensemble Model," *Environ Sci Technol*, vol. 54, no. 1, pp. 120–128, Jan. 2020, doi: 10.1021/acs.est.9b04279.

[30] G. Zhang *et al.*, "A framework to predict high-resolution spatiotemporal pm2.5 distributions using a deep-learning model: A case study of shijiazhuang, china," *Remote Sens (Basel)*, vol. 12, no. 17, pp. 1–33, 2020, doi: 10.3390/rs12172825.

[31] Z. Wang, Y. Zhou, R. Zhao, N. Wang, A. Biswas, and Z. Shi, "High-resolution prediction of the spatial distribution of PM2.5 concentrations in China using a long short-term memory model," *J Clean Prod*, vol. 297, p. 126493, 2021, doi: 10.1016/j.jclepro.2021.126493.

[32] F. Liang *et al.*, "MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China," *Science of The Total Environment*, vol. 616–617, pp. 1589–1598, 2018, doi: https://doi.org/10.1016/j.scitotenv.2017.10.155.

[33] R. Schneider *et al.*, "A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain," *Remote Sens (Basel)*, vol. 12, no. 22, 2020, doi: 10.3390/rs12223803.

[34] X. Li, Y. Li, Q. Ma, and S. Wang, "Random Forest Model for PM2.5 Concentration in China Using Himawari-8 Hourly AOD Product," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 1935–1938. doi: 10.1109/IGARSS47720.2021.9554364.

[35] J. Li *et al.*, "Estimation of ambient PM2.5 in Iraq and Kuwait from 2001 to 2018 using machine learning and remote sensing," *Environ Int*, vol. 151, no. January, p. 106445, 2021, doi: 10.1016/j.envint.2021.106445.

[36] G. Geng, X. Meng, K. He, and Y. Liu, "Random forest models for PM2.5 speciation concentrations using MISR fractional AODs," *Environmental Research Letters*, vol. 15, no. 3, 2020, doi: 10.1088/1748-9326/ab76df.

[37] B. Guo *et al.*, "Estimating PM2.5 concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017," *Science of the Total Environment*, vol. 778, 2021, doi: 10.1016/j.scitotenv.2021.146288.

[38] K. Huang *et al.*, "Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain," *Environmental Pollution*, vol. 242, pp. 675–683, 2018, doi: 10.1016/j.envpol.2018.07.016.

[39] T. Li, H. Shen, C. Zeng, and Q. Yuan, "A Validation Approach Considering the Uneven Distribution of Ground Stations for Satellite-Based PM2.5 Estimation," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 13, pp. 1312–1321, 2020, doi: 10.1109/JSTARS.2020.2977668.

[40] K. Gui *et al.*, "Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model," *Environ Int*, vol. 141, p. 105801, 2020, doi: https://doi.org/10.1016/j.envint.2020.105801.

[41] Y. Ding, Z. Chen, W. Lu, and X. Wang, "A CatBoost approach with wavelet decomposition to improve satellite-derived high-resolution PM2.5 estimates in Beijing-Tianjin-Hebei," *Atmos Environ*, vol. 249, no. August 2020, p. 118212, 2021, doi: 10.1016/j.atmosenv.2021.118212.

[42] R. B. A. Koelemeijer, C. D. Homan, and J. Matthijsen, "Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe," *Atmos Environ*, vol. 40, no. 27, pp. 5304–5315, 2006, doi: https://doi.org/10.1016/j.atmosenv.2006.04.044.

[43] R. Zhang *et al.*, "A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient PM2.5 levels," *Environmental Pollution*, vol. 243, pp. 998–1007, 2018, doi: https://doi.org/10.1016/j.envpol.2018.09.052.

[44] L. Han, J. Zhao, Y. Gao, Z. Gu, K. Xin, and J. Zhang, "Spatial distribution characteristics of PM2.5 and PM10 in Xi'an City predicted by land use regression models," *Sustain Cities Soc*, vol. 61, p. 102329, 2020, doi: https://doi.org/10.1016/j.scs.2020.102329.

[45] L. Yang, H. Xu, and Z. Jin, "Estimating ground-level PM2.5 over a coastal region of China using satellite AOD and a combined model," *J Clean Prod*, vol. 227, pp. 472–482, 2019, doi: 10.1016/j.jclepro.2019.04.231.

[46] C.-D. Wu *et al.*, "Land-use regression with long-term satellite-based greenness index and culture-specific sources to model PM2.5 spatial-temporal variability," *Environmental Pollution*, vol. 224, pp. 148–157, 2017, doi: https://doi.org/10.1016/j.envpol.2017.01.074.

[47] J. Zhong *et al.*, "Robust prediction of hourly PM2.5 from meteorological data using LightGBM," *Natl Sci Rev*, vol. 8, no. 10, 2021, doi: 10.1093/nsr/nwaa307.

[48] A. Mhawish *et al.*, "Estimation of High-Resolution PM2.5 over the Indo-Gangetic Plain by Fusion of Satellite Data, Meteorology, and Land Use Variables," *Environ Sci Technol*, vol. 54, no. 13, pp. 7891–7900, 2020, doi: 10.1021/acs.est.0c01769.

[49] Y. Miao, S. Liu, J. Guo, S. Huang, Y. Yan, and M. Lou, "Unraveling the relationships between boundary layer height and PM2.5 pollution in China based on four-year radiosonde measurements," *Environmental Pollution*, vol. 243, pp. 1186–1195, 2018, doi: 10.1016/j.envpol.2018.09.070.

[50] W. Qu, J. Wang, X. Zhang, L. Sheng, and W. Wang, *Opposite seasonality of the aerosol optical depth and the surface particulate matter concentration over the north China Plain*, vol. 127. Elsevier Ltd, 2016. doi: 10.1016/j.atmosenv.2015.11.061.

[51] C. J. Paciorek, Y. Liu, H. Moreno-Macias, and S. Kondragunta, "Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM2.5," *Environ Sci Technol*, vol. 42, no. 15, pp. 5800–5806, 2008, doi: 10.1021/es703181j.

[52] Y. Liu, G. Cao, N. Zhao, K. Mulligan, and X. Ye, "Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach," *Environmental Pollution*, vol. 235, pp. 272–282, 2018, doi: 10.1016/j.envpol.2017.12.070.

[53] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data," *Atmosphere (Basel)*, vol. 10, no. 7, 2019, doi: 10.3390/atmos10070373.

[54] X. Hu *et al.*, "Estimating PM2 . 5 Concentrations in the Conterminous United States Using the Random Forest Approach Department of Environmental Health , Rollins School of Public Health , Emory University , Department of Biostatistics & Bioinformatics , Rollins School of," *Environ Sci Technol*, pp. 1–29, 2017.

[55] C. Brokamp, R. Jandarov, M. Hossain, and P. Ryan, "Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model," *Environ Sci Technol*, vol. 52, no. 7, pp. 4173–4179, 2018, doi: 10.1021/acs.est.7b05381.

[56] G. Geng, X. Meng, K. He, and Y. Liu, "Random forest models for PM2.5 speciation concentrations using MISR fractional AODs," *Environmental Research Letters*, vol. 15, no. 3, p. 34056, 2020, doi: 10.1088/1748-9326/ab76df.

[57] C. R. Jung, W. T. Chen, and S. F. Nakayama, "A national-scale 1-km resolution pm2.5 estimation model over japan using maiac aod and a two-stage random forest model," *Remote Sens (Basel)*, vol. 13, no. 18, 2021, doi: 10.3390/rs13183657.

[58] C. Zhao *et al.*, "Estimating the daily PM2.5 concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01°×0.01° spatial resolution," *Environment International*, vol. 134. 2020. doi: 10.1016/j.envint.2019.105297.

[59] S. Park *et al.*, "Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models," *Science of The Total Environment*, vol. 713, p. 136516, 2020, doi: https://doi.org/10.1016/j.scitotenv.2020.136516.

[60] W. Xue *et al.*, "Inferring near-surface pm2.5 concentrations from the viirs deep blue aerosol product in china: A spatiotemporally weighted random forest model," *Remote Sens (Basel)*, vol. 13, no. 3, pp. 1–17, 2021, doi: 10.3390/rs13030505.

[61] Q. Xiao, H. H. Chang, G. Geng, and Y. Liu, "An Ensemble Machine-Learning Model to Predict Historical PM2.5 Concentrations in China from Satellite Data," *Environ Sci Technol*, vol. 52, no. 22, pp. 13260–13269, 2018, doi: 10.1021/acs.est.8b02917.

[62] Y. Feng, S. Fan, K. Xia, and L. Wang, "Estimation of Regional Ground-Level PM2.5 Concentrations Directly from Satellite Top-of-Atmosphere Reflectance Using A Hybrid Learning Model," *Remote Sens (Basel)*, vol. 14, no. 11, 2022, doi: 10.3390/rs14112714.

[63] T. K. Ho, "Random decision forests," *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, vol. 1, pp. 278–282, 1995, doi: 10.1109/ICDAR.1995.598994.

[64] T. Li, H. Shen, C. Zeng, and Q. Yuan, "A validation approach considering the uneven distribution of ground stations for satellite-based PM2.5 Estimation," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 13, pp. 1312–1321, 2020, doi: 10.1109/JSTARS.2020.2977668.

[65] J. Bi, J. H. Belle, Y. Wang, A. I. Lyapustin, A. Wildani, and Y. Liu, "Impacts of snow and cloud covers on satellite-derived PM2.5 levels," *Remote Sens Environ*, vol. 221, pp. 665–674, 2019, doi: https://doi.org/10.1016/j.rse.2018.12.002.