Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 5, October 2023, pp. 1099-1107

DOI: https://doi.org/10.52436/1.jutif.2023.4.5.970
p-ISSN: 2723-3863
e-ISSN: 2723-3871

SYSTEMATIC LITERATURE REVIEW OF THE CLASS IMBALANCE CHALLENGES IN MACHINE LEARNING

Rifqi Fitriadi[*1], Deni Mahdiana[2]

[1]Computer Science Master's Study Program, Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia
[2]Information Systems Study Program, Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia
Email: [1]rifqi0587@gmail.com, [2]deni.mahdiana@budiluhur.ac.id

(Article received: March 24, 2023; Revision: April 28, 2023; published: October 15, 2023)

Abstract

The significant growth of data poses its own challenges, both in terms of storing, managing, and analyzing the available data. Untreated and unanalyzed data can only provide limited benefits to its owner. In many cases, the data we analyze is imbalanced. An example of natural data imbalance is in detecting financial fraud, where the number of non-fraudulent transactions is usually much higher than fraudulent ones. This imbalance issue can affect the accuracy and performance of machine learning classification models. Many machine learning classification models tend to learn more general patterns in the majority class. As a result, the model may overlook patterns that exist in the minority class. Various research has been conducted to address the problem of imbalanced data. The objective of this systematic literature review is to provide the latest developments regarding the cases, methods used, and evaluation techniques in handling imbalanced data. This research successfully identifies new methods and is expected to provide more choices for researchers so that imbalanced data can be properly handled, and classification models can produce unbiased, accurate, and consistent results.

Keywords: *Class Imbalance, Handling Method, Machine Learning, Systematic Literature Review.*

TINJAUAN LITERATUR SISTEMATIS DARI TANTANGAN KETIDAKSEIMBANGAN KELAS DALAM *MACHINE LEARNING*

Abstrak

Pertumbuhan data secara signifikan menjadikan tantangan tersendiri, baik dalam hal menyimpan, mengelola dan menganalisis data yang tersedia. Data yang tidak diolah dan dianalisis hanya sebatas data saja, tidak bisa memberikan manfaat bagi pemiliknya. Dalam banyak kondisi, data yang kita analisa bersifat *imbalance*. Salah satu contoh ketidakseimbangan data yang terjadi secara alami adalah data untuk mendeteksi kecurangan (*fraud*) di bidang keuangan, dimana jumlah transaksi yang tidak curang biasanya jauh lebih banyak dibandingkan transaksi yang curang. Masalah ketidakseimbangan data dapat mempengaruhi akurasi dan kinerja model klasifikasi *machine learning*. Banyak model klasifikasi *machine learning* cenderung mempelajari pola yang lebih umum pada kelas mayoritas. Hasilnya, model mungkin mengabaikan pola yang terdapat pada kelas minoritas. Berbagai penelitian telah dilakukan untuk mengatasi permasalahan data yang tidak seimbang. Tujuan dari penelitian Tinjuan Literatur Sistematis ini adalah untuk memberikan gambaran perkembangan terbaru tentang kasus yang terjadi, metode yang digunakan dan teknik evaluasi dalam menangani ketidakseimbangan data. Penelitian ini berhasil mengidentifikasi metode-metode baru dan diharapkan dapat memberikan lebih banyak pilihan bagi para peneliti sehingga data *imbalance* bisa ditangani dengan baik dan model klasifikasi menghasilkan model yang tidak bias, akurat dan konsisten.

Kata kunci: *Ketidakseimbangan Kelas, Machine Learning, Metode Penanganan, Tinjauan Literatur Sistematis.*

1. PENDAHULUAN

Beberapa dekade terakhir, penggunaan teknologi informasi berkembang sangat pesat [1]. Penggunaan teknologi informasi memberikan banyak kemudahan dalam kehidupan sehari-hari [2]. Teknologi informasi mengubah cara kita dalam bekerja, belajar, berperilaku maupun berinteraksi antara satu individu dengan individu yang lain [3]. Penggunaan teknologi informasi telah diterapkan di

berbagai bidang, termasuk bisnis, sosial, kesehatan, pendidikan, dan masih banyak lagi [4].

Akibat dari penerapan teknologi informasi di berbagai bidang adalah tumbuhnya data secara signifikan [5]. Data yang tumbuh sangat cepat menjadi tantangan baru, baik untuk menyimpan, mengelola, dan menganalisis data dalam jumlah besar [6]. Istilah yang biasa digunakan untuk volume data yang sangat besar, kompleks, dan bervariasi yang tidak dapat diproses menggunakan metode tradisional adalah *Big Data* [7]. Karakteristik lain dari *Big Data* selain jumlah data yang sangat besar, antara lain kecepatan data dihasilkan dan diproses (*velocity*), beragamnya format dan tipe data (*variety*), kepastian dan keakuratan data (*veracity*) dan manfaat sosial dan ekonomi terkait tentang data (value) [8][9]. Tantangan muncul untuk melakukan pengolahan dan analisis data dalam skala besar sehingga data tersebut dapat menghasilkan wawasan dan informasi yang bermanfaat [6].

Dalam banyak kondisi, data yang kita olah dan analisa bersifat *imbalance*, artinya dataset yang memiliki jumlah yang tidak seimbang antara kelas mayoritas dan kelas minoritas [10]. Sebagai contoh, dalam kasus klasifikasi *machine learning*, jumlah sampel kelas positif (kelas hasil yang diinginkan) lebih sedikit daripada jumlah sampel kelas negatif.

Ketidakseimbangan data diakibatkan oleh berbagai macam faktor. Faktor yang paling umum terjadi adalah kondisi alami data. Contoh dari ketidakseimbangan data yang terjadi secara alami adalah data untuk mendeteksi kecurangan (*fraud*) di bidang keuangan [11], dimana jumlah transaksi yang tidak curang biasanya jauh lebih banyak dibandingkan transaksi yang curang. Kondisi alami lain yang menyebabkan ketidakseimbangan data adalah deteksi penyakit berat [12], dimana populasi yang memiliki penyakit stadium berat, jumlahnya lebih sedikit dibandingkan dengan populasi yang memiliki penyakit stadium ringan dan sedang [13]. Faktor lain yang menyebabkan ketidakseimbangan data adalah kesalahan dalam proses pengumpulan data. Dimana kelompok dalam klasifikasi, disebut dengan kelas, tidak terwakili dengan baik oleh jumlah sampel yang diambil.

Secara umum, *machine learning* dapat didefinisikan sebagai proses pengembangan algoritma atau model yang belajar dari data untuk mengidentifikasi pola dan menghasilkan prediksi tanpa adanya instruksi eksplisit [14]. Oleh karena itu, kondisi data yang baik sangat mempengaruhi model berserta hasil prediksinya. Model *machine learning* mengalami kesulitan dalam melakukan generalisasi pada dataset yang tidak seimbang, sehingga dapat mempengaruhi kinerja model pada data yang tidak dikenal. Model *Machine Learning* cenderung mempelajari pola yang lebih umum pada kelas mayoritas. Hasilnya, model mungkin mengabaikan pola yang terdapat pada kelas minoritas. Ini dapat menyebabkan model menjadi bias atau tidak akurat pada kelas minoritas [10]. Selain itu, model memiliki kecenderungan memiliki akurasi yang tinggi pada kelas mayoritas tetapi akurasinya kurang pada kelas minoritas.

Berbagai penelitian telah dilakukan untuk mengatasi permasalahan data yang tidak seimbang. Penting untuk menyadari ketidakseimbangan dalam data dan mengambil tindakan untuk menanganinya agar model *machine learning* dapat mempelajari pola data dengan baik dan memberikan hasil prediksi yang lebih akurat dan konsisten [15]. Tujuan dari penelitian ini adalah untuk mengidentifikasi tentang metode dan pendekatan apa saja yang digunakan pada penelitian yang dijadikan sumber literatur, gambaran tentang perkembangan tentang kasus yang terjadi, metode yang digunakan dan cara evaluasi dalam menangani ketidakseimbangan data.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *Systematic Literature Review* (SLR) dimana dilakukan penelitian literatur secara sistematis dan obyektif untuk mengumpulkan, mengevaluasi, dan mengambil suatu kesimpulan atas semua bukti yang relevan dan valid yang tersedia tentang masalah yang telah diidentifikasi [16]. Penelitian ini diharapkan dapat menjadi sumber rujukan untuk mendukung dalam pengambilan keputusan yang bersifat teknik maunpun strategis serte mendorong pengembangan penelitian baru.

Untuk menimimalkan kesalahan dalam penelitian ini, peneliti melakukan serangkaian tahapan sebagaimana ditampilkan pada Gambar 1. Tahapan yang dilakukan secara ketat dimulai dari perencanaan, memformulasikan pertanyaan penelitian, pencarian literatur yang sistematis, seleksi hasil pencarian secara ketat, ekstraksi informasi berdasarkan pertanyaan penelitian, dan analisis data dari literatur yang terpilih [17].

Gambar 1. Metodologi Penelitian

2.1. Perencanaan

Sub bagian ini menjelakan tahapan awal dari penelitian. Mulai dari menetapkan konteks dan ruang lingkup penelitian, melakukan identifikasi masalah atau pertanyaan penelitian, menentukan metode penelitian, mengidentifikasi sumber-sumber literatur yang akan dianalisa, merencakanan metode untuk ekstraksi informasi dan metode analisis yang digunakan. Selain itu, ditetapkan objek peneltian serta batasan waktu yang digunakan. Pada penelitian ini ditetapkan bahwa batasan litaratur yang akan dianalisa adalah literatur yang diterbitkan oleh ScienceDirect dari tahun 2019 sampai dengan bulan Maret 2023. Penerbit ScienceDirect dipilih karena ScienceDirect merupakan salah satu penerbit jurnal ilmiah terbesar dan terpercaya di dunia. ScienceDirect juga memberikan fitur yang memudahkan peneliti untuk mencari, menelusuri, dan mengakses referensi sesuai dengan penelitian ini. Selain itu ScienceDirect menerbitkan jurnal-jurnal ilmiah yang berkualitas dan terindeks di berbagai database ilmiah ternama, seperti Scopus dan Web of Science.

2.2. Formulasi Pertanyaan

Sub bagian ini menjelaskan pertanyaan penelitian yang digunakan pada penelitian ini. Tujuan penelitian ini adalah melakukan tinjauan literatur secara sistematis atas metode terkini untuk menyelesaikan tantangan ketidakseimbangan data pada *machine learning.* Adapun pertanyaan penelitian atau *reserach question* (RQ) pada penelitian ini sebagai berikut:

RQ1. Bagaimana gambaran sebaran literatur penanganan ketidakseimbangan data dari tahun 2019?
RQ2. Apa saja bidang dan keadaan yang menyebabkan data latih untuk *machine learning* menjadi tidak seimbang?
RQ3. Bagaimana pendekatan yang digunakan untuk menangani tantangan ketidakseimbangan data?
RQ4. Bagaimana metode yang digunakan peneliti untuk mengatasi tantangan ketidakseimbangan data agar kinerja model *machine learning* menjadi lebih akurat?
RQ5. Apa saja teknik yang digunakan untuk mengevaluasi kinerja model dalam menghadapi tantangan ketidakseimbangan data?

RQ1 bertujuan untuk mengetahui tren literatur terbaru yang membahas hasil implementasi penanganan ketidakseimbangan data. Sedangkan RQ2 bertujuan untuk mengetahui bidang dan keadaan yang menjadikan data latih (*data training*) yang akan digunakan pada tahap pemodelan menggunakan *machine learning* mengalami ketidakseimbangan data serta faktor-faktor penyebab yang menjadikan data tersebut menjadi tidak seimbang. RQ3 bertujuan untuk mengetahui tren terbaru dari pendekatan yang digunakan, apakah pada pendekatan level data (*preprocessing)*, pendekatan level algoritma atau pendekatan *hybrid* (gabungan antara metode *preprocessing* dan metode algoritma). Selanjutnya, RQ4 bertujuan untuk memahami metode apa saja yang dilakukan untuk mengatasi tantangan ketidakseimbangan data. Yang terakhir, RQ5 bertujuan untuk mengetahui teknik yang digunakan untuk mengevaluasi kinerja model ketika menghadapi tantangan data yang tidak seimbang.

2.3. Pencarian Literatur

Pencarian awal dilakukan dengan mencari dan mengidentifikasi literatur yang sesuai untuk dilakukan tinjauan literatur sistematik pada penelitian ini. Literatur dari ScienceDirect dicari dengan menggunakan kata kunci imbalance AND ”machine learning” pada kolom inputan *Title, abstract or author-specified keywords* dan memasukkan filter tahun 2019-2023 pada kolom inputan *Years*.

Hasil dari pencarian awal dengan filter kata kunci dan tahun sebagaimana disebutkan di atas mendapatkan artikel sebanyak 891. Namun, berdasarkan hasil pengamatan awal, hasil ini belum sesuai dengan kebutuhan penelitian ini karena tidak semua artikel menjelaskan tentang kondisi data yang tidak seimbang, metode dan pendekatan yang digunakan serta evaluasi dari hasil pemodelan.

2.4. Seleksi Hasil Pencarian

Hasil pencarian awal menunjukkan bahwa banyak artikel yang tidak sesuai dengan penelitian ini, sehingga perlu dilakukan pemilihan secara ketat dan sistematis agar artikel yang masuk sebagai bahan tinjauan literatur sistematis ini sesuai dengan konteks yang telah ditetapkan.

Tahapan seleksi pertama dengan menyaring artikel dengan kata kunci yang terdapat pada judul artikel saja. Hal ini dilakukan agar artikel hasil seleksi merupakan artikel yang benar-benar membahas tentang cara menangani data tidak seimbang pada *machine learning*, bukan artikel yang hanya mengandung kata kunci tersebut.

Tahap seleksi kedua, peneliti melakukan penyaringan artikel yang hanya berupa *research article*. Motivasi mengapa filter ini dilakukan agar artikel yang dilakukan tinjauan literatur sistematis hanya artikel yang berjenis artikel penelitian, bukan merupakan artikel tinjauan sistematis, konferensi abstrak maupun artikel editorial.

Di tahap seleksi yang ketiga, peneliti melakukan penyaringan serta memvalidasi secara mendalam terhadap isi dari artikel sehingga artikel yang dipakai sebagai bahan literatur benar-benar berisi tentang pendekatan dan metode penanganan ketidakseimbangan kelas data dalam *machine learning*. Sebagai contoh, ketika didapatkan artikel dengan judul yang mengandung kata *imbalance* dan

machine learning, tetap harus dilakukan proses penyaringan secara mendalam terhadap isi dari artikel tersebut karena belum tentu membahas tentang pendekatan dan metode penanganan ketidakseimbangan kelas data dalam *machine learning*.

2.5. Ektraksi Informasi

Pada tahap ekstraksi informasi, peneliti melakukan pengambilan data dan informasi dari literatur sumber berdasarkan pertanyaan penelitian yang ditetapkan sub bagian 2.2. Peneliti membaca secara seksama setiap literatur kemudian mengekstrak informasi yang berhubungan dengan penelitian. Informasi yang diambil berupa judul dan tahun penerbitan jurnal, keadaan atau kondisi yang menyabkan studi kasus ketidakseimbangan data, metode yang digunakan peneliti untuk mengatasi permasalahan ketidakseimbangan data, model *machine learning* yang digunakan untuk klasifikasi data serta teknik evaluasi atas model *machine learning* yang digunakan. Hasil proses ektraksi informasi akan dibahas secara mendalam pada bagian berikutnya.

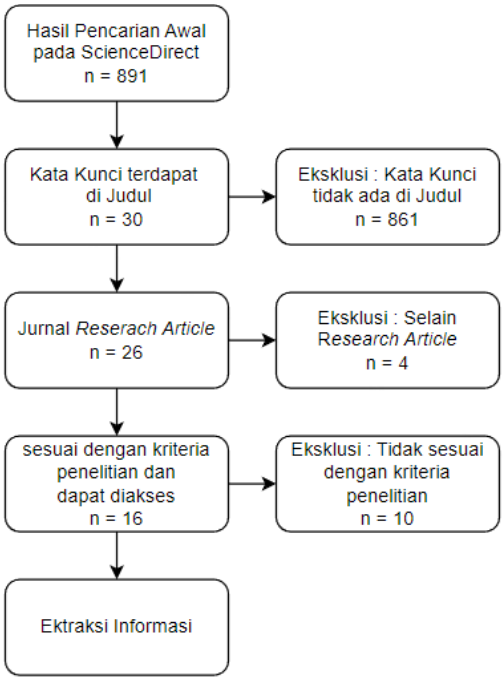
2.6. Analisis

Tahapan analisis dilakukan dengan menggunakan hasil dari tahapan ekstraksi informasi pada sub bagian 2.4 untuk dapat menjawab pertanyaan penelitian yang telah ditetapkan. Metode statistik digunakan untuk proses analisa agar dapat memberikan gambaran tentang tren fenomena yang terjadi secara menyeluruh. Kemudian, hasil analisis diinterpretasikan untuk memberikan jawaban terhadap setiap pertanyaan penelitian dan menyusun kesimpulan yang relevan dengan topik penelitian. Proses analisis beserta hasilnya dipaparkan secara menyeluruh pada bagian berikutnya.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Seleksi Pencarian

Seleksi akhir menghasilkan 16 artikel yang dijadikan sebagai bahan tinjauan literatur. Hasil tersebut didapatkan melalui proses pencarian dan penyaringan secara ketat. Pencarian literatur awal menghasilkan literatur sebanyak 891 artikel. Seleksi awal dilakukan dengan menerapkan kata kunci *imbalance* dan *machine learning* yang dicari hanya terdapat pada judul saja dengan filter tahun yang sama yaitu tahun 2019 sampai dengan 2023. Sampai dengan penelitian ini dibuat (bulan Maret 2023) dihasilkan 30 artikel yang sesuai dengan kriteria tersebut. Namun, hasil seleksi tersebut masih menghasilkan literatur berjenis *review article*, *conference abstracts*, *editorials* dan *research article* sehingga perlu dilakukan seleksi lebih lanjut. Seleksi yang kedua dengan menerapkan penyaringan jenis literatur hanya yang berjenis *research article* dan mendapatkan hasil 26 artikel yang sesuai. Kemudian dari 26 artikel, ditahap seleksi terakhir, peneliti memilih artikel yang dapat diunduh secara gratis untuk kemudian diidentifikasi serta divalidasi isinya sehingga dihasilkan kumpulan literatur yang sesuai dengan konteks penelitian ini. Jumlah literatur akhir yang dihasilkan dari proses seleksi ini sebanyak 16 literatur. Tahapan seleksi disajikan secara sistematis pada Gambar 2.



Gambar 2. Seleksi Literatur Sistematik

3.2. Hasil Ekstraksi Informasi

Dari 16 artikel yang dilakukan tinjauan, dilakukan sistesa informasi sesuai dengan pertanyaan penelitian. Hasilnya dibuat dalam bentuk tabel yang terdiri dari lima kolom, yaitu nama penulis dan tahun terbit artikel, data yang digunakan, metode penangangan data *imbalance* yang digunakan, pendekatan yang digunakan dan teknik evaluasinya. Hasil ektraksi secara detail, ditampilkan pada Tabel 1.

Tabel 1. Ektraksi Informasi

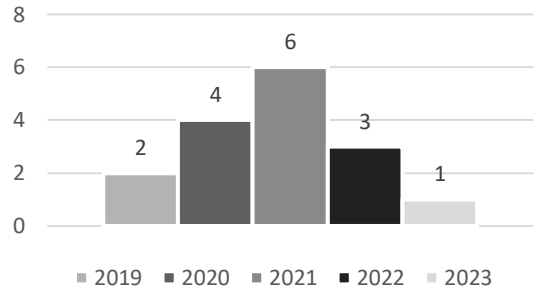
Penulis dan Tahun	Data	Metode Penangannan	Pendekatan	Evaluasi
Chen et al. 2021 [18]	Customers' future purchases	Improved Factoriazation Model using Random Forest (RFFM), Improved XGBoost	Hybrid Approach	Precision, Recall, F1 Score
Prado et al. 2020 [19]	Mineral prospectivity mapping	SMOTE, SVM	Data-level Approach	F1 Score

Wang et al. 2020 [20]	Recurrence in patients	SMOTE, SVM, Adaboost	Hybrid Approach	AUC, ROC
Pirizadeh et al. 2021 [21]	Enhanced Oil Recover	Bagging, Boosting, and Stacking (B2S) Model	Algorithm-level Approach	Accurary
Bae et al. 2021 [22]	Genotoxicity dataset	SMOTE, Gradient Boosting Tree (GBT)	Data-level Approach	F1 Score
Sarkar et al. 2020 [23]	Injury severity	KMSMOTE	Data-level Approach	Recall, F1 score
Liu et al. 2019 [24]	Cerebral stroke dataset	AutoHPO-based DNN	Hybrid Approach	FNR, FPR, accuracy
Keshavarzi et al. 2020 [25]	Cow and Bull Abortion incidence	RUS, ROS, Naïve Bayes and Bayesian network	Data-level Approach	F1 Score, AUC
Bourel et al. 2021 [26]	Faecal contamination in beach waters	RUS, ROS, SMOTE	Data-level Approach	Accuracy, AUC, TPR, TNR
Wang et al. 2019 [27]	Intelligent operation of heavy haul train	KNN based Denoising (EMKD)	Hybrid Approach	F1 Score
Novaes et al. 2021 [28]	Secondary testosterone deficiency	RUS, ROS, XGBoost, weighted average (wAVG)	Hybrid Approach	AUC
Alkharabsheh et al. 2022 [29]	Software design smell detection	SMOTE, LGBM, XGBoost, CatBoost	Algorithm-level Approach	Accuracy, Kappa, ROC, F1 Score
Sambasivam et al. 2021 [30]	Cassava disease detection	class-weight, SMOTE, focal loss, CNN	Data-level Approach	Precision, Recall, F1 Score
Kaisar et al. 2022 [31]	Dyslexia screening tests	ROS, SMOTE, XGBoost, AdaBoost	Data-level Approach	Accuracy, Precision, Recall, ROC
Ahmed et al. 2022 [32]	Data disk drive failures	EasyEnsemble, Balanced Random Forest (BRF), Weighted Logistic Regression (WLR)	Algorithm-level Approach	ROC, G-mean
Thiyam et al. 2023 [33]	CIC-DDoS2019 and Edge-IIoT Dataset	SMOTE and TOMEK link, Random Forest	Data-level Approach	Accuracy, ROC, MCC

3.3. Sebaran Literatur *Imbalance*

Sub bagian ini menjelaskan tentang sebaran artikel per tahun mengenai metode penanganan data *imbalance* dari tahun 2019 sampai dengan Maret 2023 sebagaimana menjadi pertanyaan RQ1. Untuk dapat menemukan metode terbaik, perlu tinjauan literatur terbaru yang dapat melihat secara sistematis tentang metode-metode yang digunakan, bagaimana metode tersebut diimplementasikan untuk kemudian diambil hasil yang terbaik.

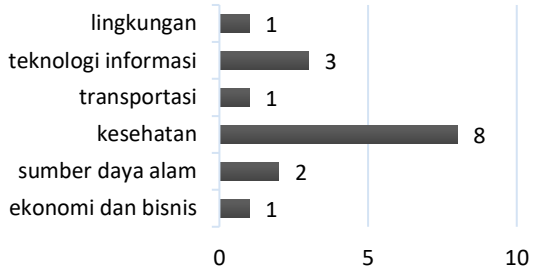
Seperti yang digambarkan pada Gambar 3, penelitian dengan kategori *research articles*, secara tren meningkat dari tahun 2019 sampai tahun 2021. Dimana pada tahun 2019, artikel penelitian penangangan data *imbalance* hanya sebanyak dua artikel, kemudian meningkat menjadi empat artikel pada tahun 2020, kemudian meningkat kembali pada tahun 2021 yaitu sebanyak enam artikel. Namun pada tahun 2022 terjadi penurunan menjadi tiga artikel saja. Sedangkan untuk tahun 2023, masih belum dapat dilihat pertumbuhannya karena masih perlu dilihat sampai dengan akhir tahun 2023. Semakin banyak artikel penelitian yang diterbitkan, terlebih lagi apabila didalamnya menemukan metode baru atau metode pengembangan dari metode yang sudah ada, maka akan menjadikan lebih banyak pilihan peneliti selanjutnya sehingga diharapkan mampu memberikan solusi terbaik terkait tantangan data *imbalance*.



Gambar 3. *Research Article(s)* Berdasarkan Tahun

3.4. Kasus Data *Imbalance*

Sub bagian ini menjelaskan tentang kasus apa saja yang menyebabkan data latih *machine learning* menjadi *imbalance* sebagaimana menjadi pertanyaan RQ2. Hasil tinjauan literatur menggambarkan bahwa kasus imbalance terjadi di berbagai bidang. Berdasarkan kolom Data pada Tabel 1 tentang Ekstraksi Informasi, setiap penelitian dikelompokkan berdasarkan bidangnya. Penelitian [18] termasuk dalam bidang Ekonomi dan Bisnis, penelitian [19], [21] termasuk dalam bidang Sumber Daya Alam, penelitian [27] termasuk dalam bidang transportasi, penelitian [29], [32], [34] termasuk dalam bidang teknologi informasi dan penelitian [20], [22]–[25], [28], [30], [31] termasuk dalam bidang kesehatan.



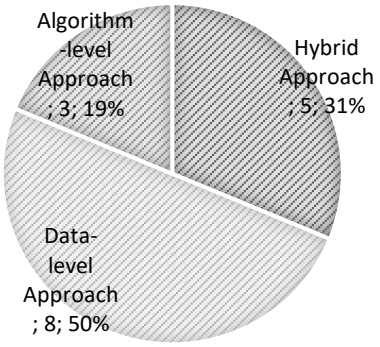
Gambar 4. Jumlah *Kasus Berdasarkan Bidang*

Seperti ditampilkan pada Gambar 4, bidang kesehatan paling dominan dengan delapan artikel, kemudian disusul bidang teknologi informasi dengan tiga artikel. Apabila dianalisa lebih mendalam menggunakan Tabel 1, didapatkan bahwa ketidakseimbangan data pada literatur bidang kesehatan didominasi oleh kasus deteksi penyakit. Ketidakseimbangan data pada data latih deteksi penyakit merupakan kondis alami, karena kelas positif yaitu orang yang memiliki penyakit pasti jumlahnya lebih sedikit dibanding dengan kelas negatif. Dengan kondisi alami data seperti itu, peran penanganan data *imbalance* menjadi sangat penting agar model yang dihasilkan tidak memiliki kecenderungan ke kelas negatif yang jumlah datanya lebih besar.

3.5. Pendekatan Penyelesaian Data *Imbalance*

Pendekatan untuk penyelesaian kondisi data *imbalance* secara umum dibagi menjadi tiga, yaitu *Data Level Approach*, *Algorithm Level Approach* dan *Hybrid Approach* [35]. Pada pendekatan Level Data, data asal dimodifikasi sedemikian rupa agar data latih yang akan dilakukan pemodelan telah menjadi data dengan kelas yang seimbang. Pada pendekatan Level Algoritma, modifikasi langsung dilakukan pada algoritma pembelajaran yang digunakan agar model yang dihasilkan tidak bias. Sedangkan pada pendekatan *Hybrid*, pendekatan yang digunakan adalah pendekatan *ensemble learning*, artinya menggunakan beberapa algoritma secara bersamaan. Tujuan penggunaan lebih dari satu algoritma agar model menghasilkan prediksi yang lebih akurat. Namun untuk pendekatan Level Algoritma dan pendekatan *Hybrid* membutuhkan sumber daya dan komputasi yang lebih besar bila dibandingkan pendekatan Level Data.

Sebagaimana digambarkan pada Gambar 5, pendekatan Level Data menjadi cara yang paling banyak digunakan dengan porsi 50% dari keseluruhan literatur, kemudian disusul Pendekatan *Hybrid* dengan 31% dan pendekatan Level Algoritma dengan 19%. pendekatan Level Data menjadi paling banyak digunakan karena memiliki tingkat kesulitan yang relatif rendah, konsepnya mudah dipahami dan tidak membutuhkan sumber daya yang besar untuk melakukan pemodelan.



Gambar 5. Pendekatan Penyelesaian Data Imbalance

3.6. Metode Penanganan Data *Imbalance*

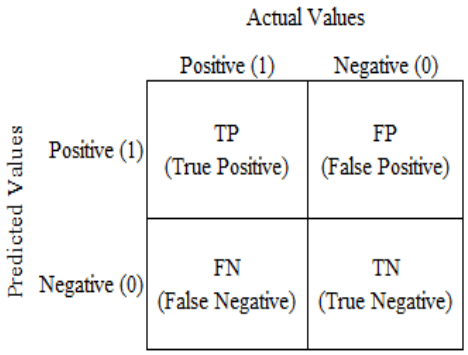
Setelah di sub bagian sebelumnya telah dijelaskan pendekatan yang digunakan untuk menyelesaikan kasus data *imbalance.* Pada sub bagian ini dipaparkan metode apa saja yang banyak digunakan pada masing-masing pendekatan.

Metode yang digunakan pada Pendekatan Level Data antara lain *Synthetic Minority Oversampling Technique* (SMOTE), *Random Under-Sampling* (RUS), *Random Over-Sampling* (ROS), KMSSMOTE dan SMOTE and TOMEK. Berdasarkan Tabel 1, SMOTE menjadi metode yang paling banyak digunakan pada tujuh literatur. SMOTE mengatasi masalah ketidakseimbangan kelas dengan membuat sampel sintetis baru berdasarkan tetangga terdekat di sekitarnya, selain itu SMOTE mudah diimplementasikan pada berbagai algoritma *machine learning*. Meskipun SMOTE juga memiliki kekurangan karena SMOTE cenderung menghasilkan sampel sintetis yang saling terkait sehingga menjadikan *noise* ke dalam data. Pada pendekatan Level Data ROS dan RUS juga cukup banyak digunakan karena kemudahan dalam pemahaman dan kemudahan dalam penerapan. Untuk Pendekatan Level Algoritma dan Pendekatan Hibrid, berdasarkan Tabel 1, masing-masing literatur menggunakan metode yang berbeda-beda. Terdapat literatur yang menggunakan model yang dikembangkan sendiri berbasiskan model yang sudah ada seperti pada literatur [18], [21] dan ada juga yang menggunakan *ensemble learning* seperti pada literatur [28].

3.7. Teknik Evaluasi

Pada sub bagian ini akan dijelaskan tentang teknik evaluasi yang digunakan pada literatur yang dilakukan tinjuan sistematis. Berdasarkan Tabel 1, teknik evaluasi yang paling banyak digunakan adalah *F1 Score*, *Precision*, *Recall*, *Area Under the Curve* (AUC*)*, *Recevier Operating Characteristic Curve* (ROC *Curve)*, *Accuracy*, *True Positive Rate* (TPR*)*, *False Positive Rate* (FPR), *True Negative Rate* (TNR) dan *False Negative Rate* (FNR) dan Selain matriks evaluasi tersebut, beberapa literatur juga menggunakan teknik evaluasi *Geometric Mean* (*G-*

Mean) dan *Cohen's Kappa*. Pada dasarnya teknik tersebut merupakan teknik evaluasi yang dibuat berdasarkan *Confusion Matrix*. *Confusion Matrix* digunakan untuk mengambarkan kinerja model klasifikasi, dimana menghasilkan angka yang menggambarkan jumlah *True Positive, False Positive, True Negative* dan *False Negative*, sebagaimana pada Gambar 6.



Gambar 6. *Confusion Matrix*

4. DISKUSI

Hasil penelitian ini telah menjawab semua pertanyaan penelitian sebagaimana disebutkan pada bagian Metodologi Penelitian. Mulai dari menjelaskan sebaran literatur penelitian untuk menjawab RQ1, menjelaskan kasus dan kondisi apa saja yang menyebabkan ketidakseimbangan data berserta penyebabnya untuk menjawab pertanyaan RQ2, menjelaskan terkini pendekatan apa saja yang digunakan pada literatur terbaru dalam hal menangani ketidakseimbangan data untuk menjawab pertanyaan RQ3, menjelaskan metode apa saja yang digunakan pada literatur terbaru untuk menjawab pertanyaan RQ4 dan yang terakhir menjelaskan teknik evaluasi yang digunakan untuk menjawab pertanyaan RQ5.

Penelitian ini menggunakan batasan waktu literatur yang diterbitkan dari tahun 2019 sampai dengan bulan Maret 2023, dengan tujuan agar ditemukan metode-metode penanganan data *imbalance* terbaru. Bila dibandingkan dengan penelitian lain dengan konteks yang sama [35], penelitian ini hasilnya selaras dengan penelitian tersebut. Namun peneliti menemukan metode-metode baru yang sebelumnya belum ada, antara lain metode Improved Factoriazation Model using Random Forest (RFFM), Improved XGBoost [18], Bagging, Boosting, and Stacking (B2S) Model [21] dan AutoHPO-based DNN [24].

5. KESIMPULAN

Penelitian ini menggunakan metode tinjuan literatur secara sistematis (SLR) untuk mengidentifikasi pendekatan, metode terbaru, teknik evaluasi serta tren penanganan permasalahan ketidakseimbangan kelas data dalam *machine learning* yang disimpulkan dari 16 artikel dari tahun 2019 sampai bulan Maret 2023. Adapun kesimpulan penelitian ini antara lain (1) Tren jumlah penelitian tentang metode penanganan data *imbalance* yang akan digunakan dalam pemodelan menggunakan *machine learning* dari tahun 2019 sampai dengan tahun 2021 terus meningkat, meskipun di tahun 2022 terjadi penurunan. (2) Ketidakseimbangan kelas data pada bidang kesehatan menjadi yang paling banyak dijadikan sumber data, khususnya ketidakseimbangan kelas data pada kondisi penyakit langka. (3) Pendekatan level data menjadi pendekatan yang paling banyak digunakan untuk menangani permasalahan ketidakseimbangan data yaitu sebanyak delapan penelitian, kemudian pendekatan hybrid dengan lima penelitian dan terakhir pendekatan level algoritma dengan tiga penelitian, Pendekatan level data lebih banyak digunakan karena mudah diimplementasikan dan tidak membutuhkan sumber daya yang banyak. (4) Ditemukan metode-metode baru yang dikembangkan antara lain Improved Factoriazation Model using Random Forest (RFFM), Improved XGBoost, Bagging, Boosting, and Stacking (B2S) Model dan AutoHPO-based DNN, sehingga memberikan lebih banyak pilihan metode untuk menyelesaikan permasalahan ini. (5) Teknik evaluasi yang digunakan merupakan teknik evaluasi yang berbasis *Confusion Matrix*. Teknik tersebut merupakan teknik evaluasi terbaik untuk mengukur kinerja model ketika terjadi ketidakseimbangan kelas data.

Meskipun telah banyak metode yang dibuat, namun setiap metode memiliki kelebihan dan kekurangan masing-masing ketika diterapkan di suatu data, sehingga tidak menutup kemungkinan di kemudian hari dibuat metode baru atau dilakukan pengembangan metode yang sudah ada. Hal tersebut mendorong peneliti untuk dapat terus memperbarui penelitian tentang permasalahan ketidakseimbangan kelas data.

DAFTAR PUSTAKA

[1] R. F. Daud and A. Novrimansyah, "Strategi Komunikasi Pemasaran Jamu Tradisional di Era Teknologi Digitalisasi 4.0," *Formosa Journal of Applied Sciences (FJAS)*, vol. 1, no. 3, pp. 233–248, 2022, doi: 10.55927.

[2] S. Sumayah, F. Sembiring, and W. Jatmiko, "Analysis of Sentiment of Indonesian Community On Metaverse Using Support Vector Machine Algorithm," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 1, 2023, doi: 10.20884/1.jutif.2023.4.1.417.

[3] R. Anggraeni and I. E. Maulani, "Pengaruh Teknologi Informasi Terhadap Perkembangan Bisnis Modern," *Jurnal Sosial dan Teknologi (SOSTECH)*, vol. 3, no. 2, 2023.

[4] Y. Jumaryadi and D. Mahdiana, "Usability Testing of Budi Luhur University E-Earning System Using System Usability Scale,"

Jurnal Teknik Informatika (JUTIF), vol. 3, no. 4, 2022.

[5] D. Sawitri, "Revolusi Industri 4.0 : Big Data Menjawab Tantangan Revolusi Industri 4.0," *Jurnal Ilmiah Maksitek*, vol. 4, no. 3, 2019.

[6] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, "Data-Driven Materials Science: Status, Challenges, and Perspectives," *Advanced Science*, vol. 6, no. 21. John Wiley and Sons Inc., Nov. 01, 2019. doi: 10.1002/advs.201900808.

[7] A. K. Kar and Y. K. Dwivedi, "Theory building with big data-driven research – Moving away from the 'What' towards the 'Why,'" *Int J Inf Manage*, vol. 54, Oct. 2020, doi: 10.1016/j.ijinfomgt.2020.102205.

[8] M. Mohammadpoor and F. Torabi, "Big Data analytics in oil and gas industry: An emerging trend," *Petroleum*, vol. 6, no. 4. KeAi Communications Co., pp. 321–328, Dec. 01, 2020. doi: 10.1016/j.petlm.2018.11.001.

[9] H. Tamiminia, B. Salehi, M. Mahdianpari, L. Quackenbush, S. Adeli, and B. Brisco, "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164. Elsevier B.V., pp. 152–170, Jun. 01, 2020. doi: 10.1016/j.isprsjprs.2020.04.001.

[10] A. S. Ashraf and T. Ahmed, "Machine Learning Shrewd Approach For An Imbalanced Dataset Conversion Samples," *Journal of Engineering and Technology*, 2020, [Online]. Available: https://journal.utem.edu.my/index.php/jet/index

[11] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTe based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 277–286, 2021, doi: 10.12785/IJCDS/100128.

[12] Y. Azhar, A. Khoiriyah Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *SINTECH JOURNAL*, vol. 5, no. 2, 2022, [Online]. Available: https://doi.org/10.31598

[13] S. Afzal *et al.*, "A Data Augmentation-Based Framework to Handle Class Imbalance Problem for Alzheimer's Stage Detection," *IEEE Access*, vol. 7, pp. 115528–115539, 2019, doi: 10.1109/ACCESS.2019.2932786.

[14] S. Mazurenko, Z. Prokop, and J. Damborsky, "Machine Learning in Enzyme Engineering," *ACS Catalysis*, vol. 10, no. 2. American Chemical Society, pp. 1210–1223, Jan. 17, 2020. doi: 10.1021/acscatal.9b04321.

[15] P. Vuttipittayamongkol and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," *Inf Sci (N Y)*, vol. 509, pp. 47–70, Jan. 2020, doi: 10.1016/j.ins.2019.08.062.

[16] W. Mengist, T. Soromessa, and G. Legese, "Method for conducting systematic literature review and meta-analysis for environmental science research," *Science of the Total Environment*, vol. 702. Elsevier B.V., Feb. 01, 2020. doi: 10.1016/j.scitotenv.2019.134581.

[17] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," *IET Software*, vol. 13, no. 6. Institution of Engineering and Technology, pp. 479–496, Dec. 01, 2019. doi: 10.1049/iet-sen.2018.5193.

[18] S. xia Chen, X. kang Wang, H. yu Zhang, and J. qiang Wang, "Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine," *Expert Syst Appl*, vol. 173, Jul. 2021, doi: 10.1016/j.eswa.2021.114756.

[19] E. M. G. Prado, C. R. de Souza Filho, E. J. M. Carranza, and J. G. Motta, "Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data," *Ore Geol Rev*, vol. 124, Sep. 2020, doi: 10.1016/j.oregeorev.2020.103611.

[20] L. Wang *et al.*, "Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods," *Comput Methods Programs Biomed*, vol. 196, Nov. 2020, doi: 10.1016/j.cmpb.2020.105567.

[21] M. Pirizadeh, N. Alemohammad, M. Manthouri, and M. Pirizadeh, "A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods," *J Pet Sci Eng*, vol. 198, Mar. 2021, doi: 10.1016/j.petrol.2020.108214.

[22] S. Y. Bae, J. Lee, J. Jeong, C. Lim, and J. Choi, "Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints," *Computational Toxicology*, vol. 20, Nov. 2021, doi: 10.1016/j.comtox.2021.100178.

[23] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, "Predicting and analyzing injury severity: A machine learning-based approach

using class-imbalanced proactive and reactive data," *Saf Sci*, vol. 125, May 2020, doi: 10.1016/j.ssci.2020.104616.

[24] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif Intell Med*, vol. 101, Nov. 2019, doi: 10.1016/j.artmed.2019.101723.

[25] H. Keshavarzi, A. Sadeghi-Sefidmazgi, A. Mirzaei, and R. Ravanifard, "Machine learning algorithms, bull genetic information, and imbalanced datasets used in abortion incidence prediction models for Iranian Holstein dairy cattle," *Prev Vet Med*, vol. 175, Feb. 2020, doi: 10.1016/j.prevetmed.2019.104869.

[26] M. Bourel *et al.*, "Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters," *Water Res*, vol. 202, Sep. 2021, doi: 10.1016/j.watres.2021.117450.

[27] X. Wang, S. Li, T. Tang, X. Wang, and J. Xun, "Intelligent operation of heavy haul train with data imbalance: A machine learning method," *Knowl Based Syst*, vol. 163, pp. 36–50, Jan. 2019, doi: 10.1016/j.knosys.2018.08.015.

[28] M. T. Novaes *et al.*, "Prediction of secondary testosterone deficiency using machine learning: A comparative analysis of ensemble and base classifiers, probability calibration, and sampling strategies in a slightly imbalanced dataset," *Inform Med Unlocked*, vol. 23, Jan. 2021, doi: 10.1016/j.imu.2021.100538.

[29] K. Alkharabsheh, S. Alawadi, V. R. Kebande, Y. Crespo, M. Fernández-Delgado, and J. A. Taboada, "A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of God class," *Inf Softw Technol*, vol. 143, Mar. 2022, doi: 10.1016/j.infsof.2021.106736.

[30] G. Sambasivam and G. D. Opiyo, "A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks," *Egyptian Informatics Journal*, vol. 22, no. 1, pp. 27–34, Mar. 2021, doi: 10.1016/j.eij.2020.02.007.

[31] S. Kaisar and A. Chowdhury, "Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests," *ICT Express*, vol. 8, no. 4, pp. 563–568, Dec. 2022, doi: 10.1016/j.icte.2022.02.011.

[32] J. Ahmed and R. C. Green II, "Predicting severely imbalanced data disk drive failures with machine learning models," *Machine Learning with Applications*, vol. 9, p. 100361, Sep. 2022, doi: 10.1016/j.mlwa.2022.100361.

[33] B. Thiyam and S. Dey, "Efficient Feature Evaluation Approach for a class-imbalanced dataset using Machine learning," *Procedia Comput Sci*, vol. 218, pp. 2520–2532, 2023, doi: 10.1016/j.procs.2023.01.226.

[34] B. Thiyam and S. Dey, "Efficient Feature Evaluation Approach for a class-imbalanced dataset using Machine learning," *Procedia Comput Sci*, vol. 218, pp. 2520–2532, 2023, doi: 10.1016/j.procs.2023.01.226.

[35] G. Rekha, A. K. Tyagi, and V. K. Reddy, "A wide scale classification of class imbalance problem and its solutions: A systematic literature review," *Journal of Computer Science*, vol. 15, no. 7. Science Publications, pp. 886–929, 2019. doi: 10.3844/jcssp.2019.886.929.