METODE HIBRIDA *OVERSAMPLING* UNTUK MENANGANI *IMBALANCED MULTI-LABEL*

Dara Tursina[1)], Sherly Rosa Anggraeni[2)], Chastine Fatichah[3)], Misbakhul Munir Irfan Subakti[4)]
[1, 2, 3, 4)] Departemen Teknik Informatika, Institut Teknologi Sepuluh Nopember, Indonesia
e-mail: daratursina20@gmail.com[1)], sherly.rosa@gmail.com[2)], chastine@if.its.ac.id[3)], irfan@if.its.ac.id[4)]

ABSTRAK

Data dan informasi terus mengalami pertambahan seiring dengan berkembangnya tekonologi digital. Ketersediaan data Data dan informasi terus mengalami pertambahan seiring dengan berkembangnya teknologi digital. Ketersediaan data menjadi semakin banyak dan kompleks. Kehadiran data yang tidak seimbang menyebabkan terjadinya kesalahan klasifikasi karena dominasi data kelas mayoritas terhadap kelas minoritas. Tidak hanya terbatas pada kelas biner, ketidakseimbangan data juga sering ditemui pada data multi-label yang semakin penting dalam beberapa tahun terakhir karena cakupan aplikasi yang luas. Namun, masalah ketidakseimbangan kelas telah menjadi ciri khas dari banyak himpunan data multi-label yang kompleks, menjadikannya fokus utama penelitian ini. Penanganan data multi-label yang tidak seimbang masih memiliki banyak potensi untuk dikembangkan. Salah satu pendekatan, Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution (MLSOL), dan Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data (UCLSO), telah dikembangkan. UCLSO hanya memusatkan perhatiannya pada kelas mayoritas, yang dapat menyebabkan ketidakseimbangan data dan overfitting yang berlebihan. Meskipun efektif dalam mencegah dominasi kelas mayoritas, pendekatan ini tidak dapat mengatasi kurangnya variasi dalam kelas minoritas. Sebaliknya, MLSOL berfokus pada kelas minoritas, memungkinkan terciptanya variasi dalam data multi-label dan meningkatkan kinerja klasifikasi secara signifikan. Penelitian ini bertujuan untuk mengatasi permasalahan ketidakseimbangan data dengan menggabungkan metode oversampling MLSOL dan UCLSO. Menggabungkan kedua pendekatan ini diharapkan dapat memanfaatkan kekuatan dan mengurangi kelemahan masing-masing, sehingga menghasilkan peningkatan kinerja yang signifikan. Hasil uji coba menunjukkan bahwa metode hibrida oversampling menghasilkan nilai tertinggi pada data biologi dengan nilai F-1 score sebesar 88% Sedangkan metode oversampling tunggal UCLSO dan MLSOL pada data biologi masing-masing memperoleh nilai F-1 score sebesar 67% dan 62%.

***Kata Kunci**: Ketidakseimbangan data, ensemble oversampling, ketidakseimbangan multi-label*

OVERSAMPLING HYBRID METHOD FOR HANDLING MULTI-LABEL IMBALANCED

Dara Tursina[1)], Sherly Rosa Anggraeni[2)], Chastine Fatichah[3)], Misbakhul Munir Irfan Subakti[4)]
[1, 2, 3, 4)] Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
e-mail: daratursina20@gmail.com[1)], sherly.rosa@gmail.com[2)], chastine@if.its.ac.id[3)], irfan@if.its.ac.id[4)]

ABSTRACT

Data and information continue to increase along with the development of digital technology. Data availability is becoming increasingly numerous and complex. The existence of unbalanced data causes classification errors due to the dominance of majority-class data over the minority class. Not only limited to the binary class, but data imbalance is also often encountered in multi-label data, which become increasingly important in recent years due to its vast application scope. However, the problem of class imbalance has been a characteristic of many complex multi-label datasets, making it the focus of this research. Handling unbalanced multi-label data still has a lot of potential for development. One approach, Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution (MLSOL) and Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data (UCLSO), has been developed. UCLSO's attention only focuses on the majority class, which can lead to data imbalance and excessive overfitting. Although effective in preventing majority class domination, this approach cannot overcome the lack of variation within the minority class. By contrast, MLSOL focuses on minority classes, allowing for variations in multi-label data and significantly improving classification performance. This research aims to overcome the problem of data imbalance by combining the MLSOL and UCLSO oversampling methods. Combining these two approaches is expected to exploit the strengths and reduce the weaknesses of each, resulting in significant performance improvements. The trial results show that the hybrid oversampling method produces the highest value on biological data with an F-1 score of 88%. Meanwhile, the single oversampling methods UCLSO and MLSOL on biological data produce an F-1 score of 67% and 62%, respectively.

***Keywords**: imbalanced data, ensemble oversampling, imbalanced multi-label*

I. PENDAHULUAN

Penelitian yang melibatkan data besar (*big data*) menjadi kajian yang banyak dipelajari oleh komunitas data mining dan machine learning. Skenario masalah yang ditemukan pun lebih beragam sehingga dapat memberikan berbagai pilihan topik, dimana salah satu yang mendapat perhatian yaitu mengenai kelas tidak seimbang. Berbagai penelitian telah dilakukan untuk menangani permasalahan klasifikasi ketika kelas tidak seimbang, namun banyak literatur yang lebih fokus pada kelas bersifat biner (memiliki dua kelas) dan *multi-class* (banyak kelas). Secara umum, kumpulan data yang tidak seimbang menjadi tantangan yang sangat signifikan dalam banyak aplikasi. Seperti deteksi penipuan, manajamen risiko, dan diagnosi medis. Ada tiga karakteristik yang dimiliki oleh imbalanced dataset yaitu *class overlap*, *small disjuncts* dan *data skew distribution* [1]. *Class overlap* berarti bahwa sampel data dari dua kelas memiliki atribut serupa dan adanya tumpang tindih dalam ruang fitur, yang dapat menyebabkan kesalahan dalam klasifikasi. *Small disjuncts* didefinisikan bahwa kelas minoritas dibagi menjadi beberapa sub-konsep, yang masing-masing hanya berisi beberapa sampel data dan mereka didistribusikan di berbagai sub-wilayah ruang fitur. *Data skew distribution* berarti bahwa jumlah sampel data sangat bervariasi antara mayoritas dan kelas minoritas. Rasio ketidakseimbangan atau *Imbalanced Ratio* (IR) antara kedua kelas bisa mencapai 1:100 atau bahkan lebih besar, sehingga akan membawa lebih banyak kesulitan dan tantangan untuk penelitian masalah klasifikasi. Oleh karena itu, rasio ketidakseimbangan merupakan faktor yang sangat penting yang mempengaruhi efek klasifikasi.

Pada penelitian sebelumnya permasalahan data dengan ketidakseimbangan pada *multi-class* mulai mendapatkan perhatian dari komunitas peneliti dalam beberapa tahun terakhir [2]. Literatur mengenai *pattern recognition* telah banyak membahas teknik klasifikasi pola pada *multi-class*, namun sebagian besar teknik tersebut menerapkannya pada kondisi data yang seimbang, seperti klasifikasi dokumen tertulis dan pengenalan suara, meskipun dalam banyak penerapan di bidang lain data yang ditemukan lebih sering memiliki kelas yang tidak seimbang. Dalam menangani permasalahan ketidakseimbangan *multi-class*, kita akan dihadapkan pada gugus data dengan konfigurasi yang berbeda, seperti kemungkinan untuk memiliki tidak hanya satu kelas mayoritas tetapi beberapa dan sebaliknya dimana satu kelas mayoritas dan beberapa kelas minoritas. Beberapa metode standar pun tidak dapat diterapkan secara langsung pada kasus ketidakseimbangan *multi-class*. Hal ini lah yang membuat permasalahan klasifikasi pada kasus imbalanced *multi-class* menjadi lebih rumit. Metode yang digunakan harus mencoba untuk menangkap dan menggali lebih jauh hubungan antara beberapa kelas yang tidak seimbang serta siap untuk bekerja dengan ketidakseimbangan yang lebih ekstrim karena mereka lebih mungkin terjadi dalam gugus data *multi-class*. Dalam mengatasi permasalahan klasifikasi *multi-class*, metode dekomposisi merupakan salah satu yang populer, sedangkan metode *ensemble* dapat digunakan untuk menangani permasalahan ketidakseimbangan kelasnya [3]. Pendekatan yang diperkenalkan untuk mengatasi permasalahan data dengan kelas tidak seimbang hampir sebagian besar hanya dirancang untuk skenario kelas biner. Beberapa metode tidak dapat diterapkan secara langsung pada kasus ketidakseimbangan *multi-class*.

Tidak hanya pada kasus imbalanced *multi-class*, data multi-label terkenal akan kesulitannya dibandingkan dengan dua data sebelumnya (*binary class* dan *multi-class*). Beberapa penelitian terkini mengajukan metode terbaru untuk mengatasi data-data multi-label. Dalam beberapa tahun terakhir data multi-label menjadi salah satu bagian data terpenting karena cakupan domain apliaksi yang begitu luas. Namun, masalah ketidakseimbangan kelas telah menjadi karakteristik yang melekat pada banyak kumpulan data multi-label, dimana sampel dan label terkait tidak terdistribusi secara merata di seluruh ruang data. Masalah ketidakseimbangan dalam data multi-label menimbulkan tantangan terhadap analisis data multi-label yaitu dapat dilihat dari tiga perspektif yakni, ketidakseimbangan dalam label, antar label, dan kumpulan label [4].

Beberapa solusi yang pernah dilakukan peneliti adalah fokus pada data binary class dan *multi-class* dengan berbagai macam metode *oversampling* dan undersampling. Potensi mengembangkan penanganan imbalanced data masih banyak peluang. Banyak peneliti yang mengembangkan dengan menggunakan data multi-label. Pada penelitian sebelumnya, data multi-label hanya menggunakan metode *Imbalanced Multi-Label Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data (UCLSO*) dan *Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution* (MLSOL). UCLSO hanya fokus pada kelas mayoritas. Sehingga semkin besar perbandingan yang diperoleh akan mengakibatkan ketidakseimbangan data dan *overfitting* yang berlebihan. Meskipun hal ini efektif dalam mencegah dominasi yang berlebihan dari kelas mayoritas, namun tidak dapat mengatasi kurangnya keberagaman dalam kelas minoritas. Sedangkan pada kasus data multi-label keberagaman minoritas lebih bervariasi. Lalu pada MLSOL hanya fokus pada kelas minoritas dan dapat menciptakan keberagaman atau variasi pada data multi-label. Tidak hanya itu MLSOL secara signifikan dapat meningkatkan kinerja klasfikasi dalam skenario multi-label. Hanya saja, karena berfokus pada satu kelas minoritas maka tidak sepenuhnya MLSOL ini memanfaaatkan informasi kontekstual dari instance kelas mayoritas, sehingga berpotensi menyebabkan hilangnya informasi penting [3]. Oleh karena itu penelitian ini berfokus pada penanganan data yang tidak seimbang (imbalanced) dengan memadukan metode

Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution (MLSOL) dengan Integrating *Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data* (UCLSO). Menggabungkan dua pendekatan metode ini akan memanfaatkan kekuatan dan mengurangi kelemahan masing-masing metode atau saling melengkapi satu sama lain. Dengan menggabungkan dua metode ini, diharapkan mendapatkan hasil yang lebih efektif pada klasifikasi data multi-label yang tidak seimbang. Selain itu juga memperoleh hasil evaluasi dan generalisasi model yang lebih baik.

II. TINJAUAN PUSTAKA

A. *Imbalanced* Data

Data memicu algoritme pembelajaran mesin. Dengan tidak adanya kumpulan data berkualitas baik, bahkan algoritme terbaik pun kesulitan untuk menghasilkan hasil yang baik. Kumpulan data yang tidak seimbang ditentukan oleh perbedaan besar dalam distribusi kelas dalam kumpulan data. Imbalanced classification adalah salah satu masalah terpenting dari machine learning dan data mining dalam membantu dataset yang memiliki kelas yang tidak seimbang (imbalanced data) [5].

Masalah klasifikasi yang tidak seimbang mengacu pada klasifikasi masalah di mana jumlah sampel dalam satu atau beberapa kategori jauh lebih besar daripada kategori lainnya. Masalah ini umum dalam tugas klasifikasi aktual (misalnya, deteksi penipuan, diagnosis penyakit, deteksi intrusi jaringan, dan sebagainya) dan adalah salah satu masalah paling menantang dalam pembelajaran mesin [6]. Klasifikasi yang tidak seimbang menimbulkan tantangan untuk pemodelan prediktif karena sebagian besar algoritme pembelajaran mesin yang digunakan untuk klasifikasi dirancang dengan asumsi jumlah contoh yang sama untuk setiap kelas. Ini menghasilkan model yang memiliki kinerja prediktif yang buruk, khususnya untuk kelas minoritas. Ini menjadi masalah karena biasanya, kelas minoritas lebih penting dan karena itu masalahnya lebih sensitif terhadap kesalahan klasifikasi untuk kelas minoritas daripada kelas mayoritas.

Model klasifikasi mencoba mengkategorikan data ke dalam kelompok yang berbeda. Dalam kumpulan data yang tidak seimbang, satu keranjang merupakan bagian besar dari kumpulan data pelatihan (kelas mayoritas), sedangkan keranjang lainnya kurang terwakili dalam kumpulan data (kelas minoritas). Solusi untuk masalah ketidakseimbangan kelas pada tingkat data bertujuan untuk memodifikasi distribusi kelas. Praktik umum adalah melakukan resample data dengan undersampling atau *oversampling*, yang masing-masing mengurangi instance kelas mayoritas dan meningkatkan instance kelas minoritas. Pada level algoritmik, masalah ditangani dengan membuat algoritme pembelajaran baru atau dengan memodifikasi algoritme yang sudah ada. Solusi tingkat algoritme memiliki keuntungan dengan memasukkan secara langsung preferensi pengguna ke dalam model. Ketidakseimbangan data tetap menjadi salah satu masalah paling luas yang memengaruhi pembelajaran mesin kontemporer [7]. Efek negatif ketidakseimbangan data pada algoritme pembelajaran tradisional paling parah jika digabungkan dengan faktor kesulitan kumpulan data lainnya, seperti disjungsi kecil, adanya outlier, dan jumlah observasi pelatihan yang tidak mencukupi. Faktor kesulitan yang disebutkan di atas juga dapat membatasi penerapan beberapa metode untuk menangani ketidakseimbangan data, khususnya algoritma *oversampling* berbasis lingkungan berdasarkan SMOTE. Analisis kompleksitas komputasi yang dilakukan menunjukkan kompleksitas waktu yang berkurang secara signifikan dari algoritma Undersampling berbasis radial (*NearMiss*) yang diusulkan, dan hasil studi eksperimental yang dilakukan menunjukkan kegunaannya, terutama pada kumpulan data yang sulit [8]. Peluang pada penelitian selanjutnya adalah dengan melakukan percobaan mengkombinasi algoritma terbaru dengan metode *classifier ensemble* dataset.

B. Ketidakseimbangan *(Imbalanced)*

Terdapat dua jenis pendekatan untuk menangani kasus imbalanced, diantaranya pendekatan pada level data dan level algoritma [8]. Pendekatan pada level data dilakukan dengan melakukan proses sampling pada data mayoritas ataupun data minoritas sehingga jumlah data menjadi lebih seimbang. Sedangkan pendekatan pada level algoritma yaitu melakukan improvisasi pada metode-metode *classifier* tanpa memproses atau merubah data awal [9]. Kedua metode tersebut memiliki kelebihan dan kekurangan masing-masing, diantaranya untuk pendekatan level data cenderung tangguh dan stabil terhadap hampir seluruh *classifier* [10]. Namun, kelemahan pada solusi ini yaitu memungkinkan terjadinya *overfitting* atau missing information pada data yang telah dilakukan sampling. Sedangkan pada solusi level algoritma, data yang diolah merupakan data asli tanpa perubahan apapun namun data tersebut akan sangat bergantung pada *classifier* tertentu. Dengan kata lain, algoritma yang diusulkan belum tentu mampu mencapai performa yang sama baik apabila diimplementasikan pada data lain yang memiliki karakteristik berbeda. Dari kedua jenis solusi ini, riset singkat berdasarkan jumlah penelitian yang ada menyatakan bahwa ternyata solusi pada level data lebih banyak dikembangkan untuk menangani kasus imbalanced dibandingkan solusi dengan modifikasi algoritma [5]. Secara umum, solusi pada level data menggunakan teknik sampling dibagi menjadi tiga jenis yaitu *oversampling*, undersampling, dan gabungan dari keduanya (*hybrid*) [9].

Metode *oversampling* merupakan metode yang bertujuan untuk menambahkan jumlah data pada kelas minoritas dengan memanfaatkan teknik sampling pada data training kelas minoritas sehingga diharapkan rasio antar kelas minoritas dan kelas mayoritas dapat lebih berimbang [10]. Sedangkan sebaliknya, teknik undersampling justru mengeliminasi sebagian data yang dianggap kurang relevan pada kelas mayoritas. Sedangkan metode hybrid merupakan kombinasi dari kedua teknik sampling tersebut sesuai dengan kebutuhan dan karakteristik data [11]. Beberapa permasalahan yang sering muncul pada kasus imbalanced yaitu:

1. *Outlier* yaitu ketika data yang bernilai ekstrim atau beda sangat jauh dengan mayoritas kelompoknya. Pada kondisi ini seringkali terjadi misklasifikasi sehingga pada beberapa penelitian, outlier tersebut dihapus.
2. Banyaknya data antar kelas yang overlap. Apabila terdapat overlapping, maka discriminative rule akan sulit untuk diproses. Hal tersebut dapat berdampak pada semakin besarnya kemungkinan terjadi mis-klasifikasi pad akelas minoritas dikarenakan jumlah data yang lebih minim [12]. Apabila overlap yang terjadi dapat diminimalisir, maka metode klasifikasi sederhana apapun akan dapat menghasilkan distribusi kelas dengan sangat baik.
3. Terdapat beberapa data pada *sub-cluster* yang memiliki jarak terlalu rapat antar kedua kelas (*small disjunction*). Adanya sub-cluster yang saling berdekatan tersebut dapat menambah kompleksitas dari suatu data dikarenakan pada umumnya jumlah data pada setiap sub-cluster tersebut tidak berimbang.

C. *Oversampling*

Oversampling justru merupakan metode sampling dengan menambahkan jumlah data pada kelas minoritas sehingga dapat mengimbangi atau mendekati jumlah data pada kelas mayoritas [14].

Permasalahan yang umum pada *oversampling* adalah terjadinya *overfitting* dikarenakan penambahan data secara berulang menyebabkan decision boundary menjadi lebih ketat. Oleh karena itu, pada perkembangannya, metode *oversampling* bukan lagi mengopi data yang sama tetapi membuat data baru yang mirip. Hal ini bertujuan untuk memperhalus boundary decision [12].

D. *Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution* (MLSOL)

Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution (MLSOL) adalah sebuah metode atau teknik yang digunakan dalam masalah klasifikasi multi-label untuk mengatasi ketidakseimbangan kelas dalam dataset [15]. Metode ini berfokus pada pembuatan sampel sintetis untuk menjaga distribusi label yang seimbang di dalam dataset, terutama ketika beberapa label jarang muncul atau kurang representatif. Dalam konteks MLSOL:

1. Masalah Ketidakseimbangan Kelas: Klasifikasi multi-label melibatkan pemberian beberapa label pada setiap data. Dalam banyak kasus dunia nyata, beberapa label mungkin jarang muncul dalam dataset, yang mengakibatkan masalah ketidakseimbangan kelas [16].
2. *Oversampling* Sintetis: Teknik *oversampling* sintetis bertujuan untuk mengatasi masalah ketidakseimbangan kelas dengan menghasilkan sampel-sampel buatan (sintetis) untuk kelas minoritas [17]. Teknik-teknik ini dapat membantu menjaga distribusi label agar lebih seimbang, membuat klasifikasi lebih kuat dan akurat, terutama untuk kelas-kelas minoritas [11].
3. Distribusi Label Lokal: MLSOL, seperti yang diindikasikan oleh namanya, mungkin berfokus pada mempertimbangkan distribusi label secara lokal di dalam dataset. Ini berarti bahwa MLSOL dapat mempertimbangkan label-label dari data-data tetangga atau serupa ketika menghasilkan sampel-sampel sintetis. Tujuannya adalah agar sampel-sampel sintetis tidak hanya beragam tetapi juga relevan dengan distribusi label lokal [18].
4. Manfaat MLSOL: MLSOL, atau teknik serupa, dapat memiliki beberapa keuntungan dalam klasifikasi multi-label, seperti meningkatkan kinerja prediksi secara keseluruhan, mengurangi bias terhadap kelas mayoritas, dan memastikan bahwa klasifikasi dapat membuat prediksi yang lebih baik untuk kelas-kelas minoritas.

Berdasarkan penelitian sebelumnya, metode *Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution* (MLSOL) dipadukan dengan metode resampling dengan cara yang sederhana namun tetap fleksibel [19]. Secara ringkas pemeliharaan label pada MSOL yaitu mempertahankan hubungan label yang ada dalam kumpulan data, memastikan bahwa contoh yang disintesis selaras dengan korelasi kelas asli. Representasi kelas yang bertujuan untuk meningkatkan keterwakilan kelas-kelas yang kurang terwakili, mengatasi masalah ketidakseimbangan kelas secara efektif. Informed *oversampling* pada MSOL memanfaatkan informasi label dari data awal dan referensi, menjadikan *oversampling* lebih tepat sasaran dan tepat sasaran [20]. MSOL bergantung pada data awal dan referensi, yang harus dipilih dengan tepat. Kualitas sumber data ini dapat memengaruhi efektivitas pengambilan sampel yang berlebihan. Dalam kumpulan data dengan hubungan label yang kompleks, MLSOL mungkin kesulitan untuk menangkap ketergantungan yang rumit secara memadai.

E. *Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data* (UCLSO)

Integrating Unsupervised *Clustering* and Label-specific *Oversampling* to Tackle Imbalanced Multi-Label Data (UCLSO) adalah sebuah metode atau teknik yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam masalah klasifikasi multi-label [21]. Metode ini menggabungkan dua komponen utama: pengelompokan tanpa supervisi (unsupervised *clustering*) dan *oversampling* yang khusus untuk setiap label (label-specific *oversampling*) guna menghasilkan dataset yang lebih seimbang. Berikut adalah penjelasan lebih lanjut tentang UCLSO. Berikut adalah penjelasan lebih lanjut tentang UCLSO:

1. Ketidakseimbangan Data Multi-Label: Dalam masalah klasifikasi multi-label, setiap sampel data dapat memiliki beberapa label yang berbeda. Terkadang, beberapa label mungkin jauh lebih umum atau sering muncul daripada yang lain, menghasilkan ketidakseimbangan kelas.
2. Pengelompokan Tanpa Supervisi (Unsupervised *Clustering*): Komponen pertama dari UCLSO adalah pengelompokan tanpa supervisi. Ini adalah teknik yang digunakan untuk mengelompokkan sampel-sampel data menjadi kelompok-kelompok yang serupa berdasarkan ciri-ciri atau atribut yang ada dalam data tersebut. Tujuannya adalah untuk menemukan struktur internal dalam data tanpa memerlukan label atau anotasi tambahan [22].
3. *Oversampling* Label Khusus (Label-specific *Oversampling*): Komponen kedua dari UCLSO adalah *oversampling* yang khusus untuk setiap label. Ini berarti bahwa metode ini mempertimbangkan ketidakseimbangan label secara individual. Label-label yang kurang representatif akan diambil perhatian khusus, dan sampel-sampel sintetis akan dihasilkan untuk meningkatkan jumlah sampel dengan label-label tersebut.
4. Integrasi: UCLSO mengintegrasikan kedua komponen berdasarkan pemahaman terhadap karakteristik, struktur, dan pola data. Hasil pengelompokan tanpa supervisi digunakan untuk membantu mengidentifikasi hubungan antara sampel-sampel data dan label-label. Dengan memahami hubungan ini, *oversampling* label khusus dapat dilakukan dengan lebih cerdas, sehingga menghasilkan sampel-sampel sintetis yang lebih relevan dan berguna.
5. Cara yang lebih akurat dan cerdas. Dengan memanfaatkan pengelompokan tanpa supervisi, metode ini dapat mengidentifikasi pola-pola dalam data yang tidak selalu terlihat dalam klasifikasi berbasis label tunggal.

III. METODOLOGI PENELITIAN

A. *Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data (UCLSO)*

Pada tahap ini, diterapkan *Unsupervised Clustering with Label-Specific Oversampling* (UCLSO), suatu pendekatan yang memanfaatkan algoritma KMeans untuk menjelajahi dataset multilabel yang tidak seimbang[24]. Parameter *k*, yang menentukan jumlah kluster yang dihasilkan oleh algoritma KMeans, memainkan peran penting dalam penyesuaian analisis. Dalam konteks eksperimen, pemilihan nilai *k* yang sesuai menjadi krusial untuk memastikan hasil yang informatif. Algoritma KMeans diinisialisasi dengan parameter yang tepat dan mulai membentuk kluster, berfungsi sebagai representasi pola tersembunyi dalam dataset. Tiap instans diberikan tugas untuk menjadi bagian klusternya sendiri, memberikan gambaran yang lebih jelas tentang struktur dataset yang kompleks. Keunggulan utama dari algoritma KMeans terletak pada kemampuannya mengelompokkan instans dengan karakteristik serupa. Ini menjadi kunci dalam menyusun strategi komprehensif untuk mengatasi ketidakseimbangan kelas, karena KMeans dapat membantu mengidentifikasi dan memahami hubungan serta pola yang tersembunyi dalam dataset multidimensional. Dalam eksplorasi ini, KMeans bukan hanya menjadi alat analisis, tetapi juga mitra yang dapat diandalkan, membantu mengungkap kompleksitas dan hubungan dalam dataset multilabel. Mengambil inspirasi dari wawasan KMeans, algoritma UCLSO mengeksplorasi struktur berkelompok untuk melakukan *Label-Specific Oversampling*. Untuk setiap label yang kurang diwakili dalam suatu kluster, instans sintetis dihasilkan dengan menginterpolasi titik data yang sudah ada. Proses ini menangani ketidakseimbangan lokal, memastikan strategi *oversampling* yang lebih nuansa dengan mempertimbangkan distribusi label dan pola kluster [2]. UCLSO terintegrasi secara mulus dengan *Synthetic Minority Over-sampling Technique* (SMOTE), yang diterapkan secara terpisah untuk setiap label. Hal ini menghasilkan dataset yang tidak hanya seimbang secara global, tetapi juga menunjukkan keseimbangan lokal dalam kluster. Pendekatan terpadu ini bertujuan menciptakan set pelatihan yang lebih representatif dan terdiversifikasi untuk model machine learning selanjutnya.

Langkah-langkah implementasi UCLSO sebagai berikut:

1. Pertama, identifikasi kluster data menggunakan algoritma KMeans, di mana entitas-data dengan atribut serupa dikelompokkan bersama, membentuk kelompok yang mencerminkan pola tertentu dalam data. Pengamatan terhadap beberapa kluster menunjukkan adanya ketidakseimbangan dalam distribusi label, dengan beberapa label muncul lebih jarang dibandingkan yang lain.

2. Selanjutnya, UCLSO diterapkan secara spesifik pada klusternya dengan melakukan *oversampling* pada label yang kurang umum, menciptakan variasi sintetis yang diperlukan untuk meningkatkan representasi label minoritas. Data hasil *oversampling* diintegrasikan dengan dataset asli, bertujuan membentuk dataset yang tidak hanya seimbang secara global tetapi juga mencerminkan variasi yang lebih baik dalam distribusi label, memberikan gambaran yang lebih komprehensif.
3. Selanjutnya, model machine learning dievaluasi pada dataset hasil *oversampling*, dan parameter UCLSO disesuaikan sesuai kebutuhan.

B. *Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution* (MLSOL)

Dalam melanjutkan pendekatan penanganan ketidakseimbangan kelas pada dataset multilabel, langkah berikutnya adalah menerapkan *Modified Multi-Label Synthetic Oversampling* (MLSOL) setelah tahap *Unsupervised Clustering with Label-Specific Oversampling* (UCLSO). Pendekatan ini dirancang untuk mengatasi ketidakseimbangan kelas pada tingkat kluster dengan fokus pada *oversampling* instance spesifik. MLSOL dimulai dengan inisialisasi parameter khusus, termasuk jumlah generasi sintetis (*GenNum*) yang dihitung sebagai konversi bilangan bulat dari panjang dataset (*D*) dikalikan dengan parameter *P*. Langkah ini mengukur seberapa banyak sintetis yang akan dihasilkan untuk memperkaya dataset. Selanjutnya, MLSOL menggunakan algoritma *K-Nearest Neighbors* (kNN) untuk menemukan tetangga terdekat dari setiap contoh dalam dataset. Hal ini penting untuk menilai kedekatan relatif antar contoh, menjadi dasar untuk pengembangan contoh sintetis.

$$C = \frac{distances[:, -1]}{distances[:, 1]} \tag{1}$$

$$w = \frac{1}{1 + c} \tag{2}$$

Untuk setiap contoh, MLSOL menghitung nilai *C* dan *w*. Nilai *C* menunjukkan seberapa dekat suatu contoh dengan tetangga terdekatnya dibandingkan dengan tetangga terjauhnya. Nilai *w* dihitung berdasarkan nilai *C*, menentukan bobot untuk contoh tersebut. Semakin tinggi nilai *C*, semakin rendah nilai w, dan sebaliknya. Tujuan dari nilai *w* adalah memengaruhi pemilihan contoh *seed* selama pembentukan contoh sintetis. MLSOL memasuki iterasi pembentukan contoh sintetis, di mana contoh *seed* dipilih dari dataset berdasarkan bobot *w*. Selanjutnya, contoh referensi dipilih secara acak dari tetangga terdekat (kNN).

Dataset yang telah diperluas kemudian diindeks ulang untuk memastikan kontinuitas dan integritas dataset. Terakhir, bobot sampel diperbarui untuk memasukkan sampel-sampel baru yang dihasilkan oleh MLSOL. Penerapan MLSOL bersama UCLSO menghasilkan dataset yang lebih seimbang dan informatif. Model yang dikembangkan menggunakan dataset yang telah diperluas mampu mengatasi ketidakseimbangan kelas pada tingkat kluster dan instance secara spesifik. Dengan kombinasi kekuatan analisis kluster dari UCLSO dan presisi MLSOL dalam *oversampling* pada tingkat kluster, pendekatan ini memberikan kontribusi positif terhadap peningkatan kinerja model dalam menangani distribusi kelas yang tidak merata pada tugas klasifikasi multilabel. Dengan memanfaatkan kekayaan informasi dari kedua pendekatan ini, model dapat lebih efektif menangani kasus multilabel dengan ketidakseimbangan yang signifikan, membawa dampak positif dalam berbagai aplikasi mulai dari kesehatan hingga pengenalan pola kompleks.

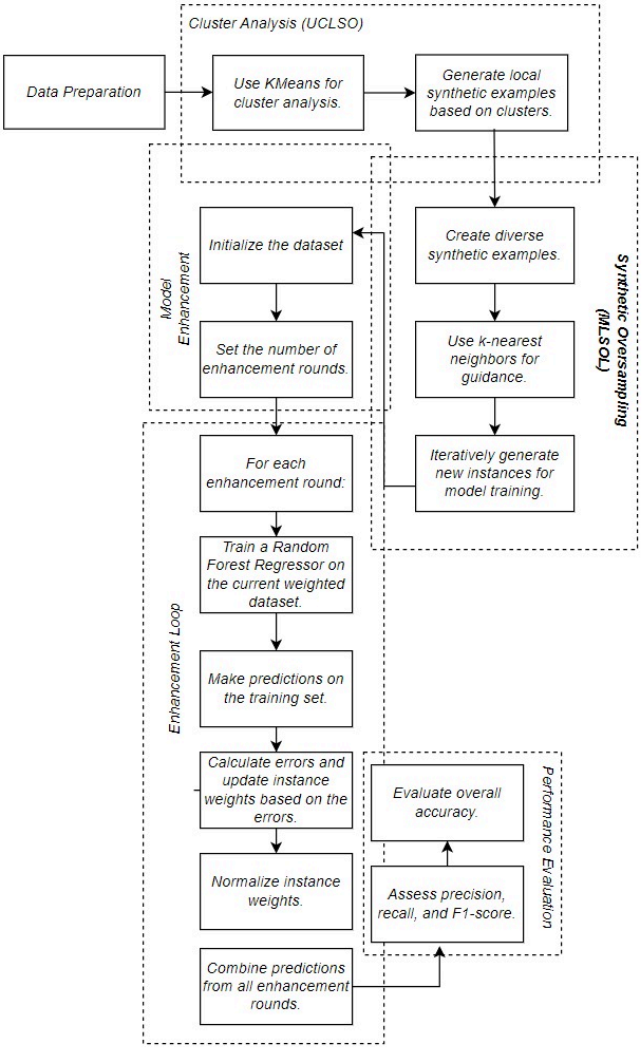
C. *Penggabungan metode MLSOL + UCLSO (Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution (MLSOL) dengan Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data (UCLSO)*

Pelestarian Label dan Keanekaragaman Data, MLSOL unggul dalam menjaga hubungan label, sementara UCLSO berfokus pada pembuatan sampel yang beragam dalam kluster. Menggabungkan kedua metode memastkan bahwa dalam pengambilan sampel data yang berlebihan, informasi spesifik dari setiap label tetap dipertahankan, sambil juga menciptakan variasi dalam sintetis [2]. Kombinasi ini dapat secara efektif mengatasi ketidakseimbangan kelas dengan me-manfaatkan pelestarian label MLSOL dan kemampuan UCLSO untuk membuat sampel yang beragam. Pendekatan khusus label UCLSO yang berfokus pada cara meningkatkan representasi label minoritas sambil mempertahankan informasi penting dari setiap label dalam dataset melengkapi MLSOL dengan menambahkan strategi pengambilan sampel berlebih yang dapat menangani ketergantungan label yang kompleks, seperti ketergantungan label yang rumit antara label-label dalam dataset. Kombinasi ini dapat meningkatkan kestabilan pengambilan sampel berlebih tanpa terlalu dipengaruhi oleh fluktuasi dalam data, karena kekuatan masing-masing metode seimbang dengan kelemahan metode lainnya. Dalam konteks UCLSO dan MLSOL, biasanya tidak terjadi *overfitting* karena algoritma-algoritma ini dirancang untuk menghasilkan sampel sintetis yang relevan dengan

struktur kluster data. Data sintetis yang dihasilkan mencerminkan distribusi label dalam kluster yang ada, sehingga tidak ada penambahan data yang sama atau sama yang dapat menyebabkan *overfitting* [23].

Bagian dari proses undersampling juga dilakukan untuk mengurangi ketidakseimbangan data sebelum tahap *oversampling*. Ini membantu menjaga keseimbangan antara label dalam kluster dan memungkinkan lebih banyak sampel sintetis yang relevan dihasilkan. Jadi, UCLSO dan MLSOL melakukan kombinasi dari undersampling dan *oversampling* untuk mengatasi masalah ketidakseimbangan dalam data multi-label. Meskipun ada proses *oversampling*, UCLSO dan MLSOL mencoba untuk mempertahankan struktur data yang ada dan menghindari penambahan data. Gambar 1 menjelaskan tahapan penggabungan metode UCLSO dan MLSOL dalam bentuk diagram.

D. Penggabungan kluster



Gambar 1. Desain Diagram Hibrida UCLSO dan MLSOL

Dalam pengembangan pendekatan penanganan ketidakseimbangan kelas pada dataset multilabel, langkah selanjutnya adalah menggabungkan kluster yang telah di *oversample* melalui proses yang dikenal sebagai *"Combine Over-Sampled Clusters*." Langkah ini merupakan kelanjutan dari penerapan Multi-Label *Synthetic Oversampling* (MLSOL) dalam metodologi *Unsupervised Clustering with Label-Specific Oversampling* (UCLSO). UCLSO merupakan metode yang memiliki distribusi dan struktur intrinsik dalam dataset yang tidak seimbang dengan menggunakan algoritma KMeans sebagai alat analisis dasar. Setelah menerapkan MLSOL yang merinci *oversampling* pada tingkat kluster, langkah *Combine Over-Sampled Clusters* menjadi krusial untuk menghasilkan dataset yang lebih seimbang dan representatif.

Proses *Combine Over-Sampled Clusters* dimulai dengan menggabungkan dataset yang telah diperluas oleh MLSOL (*X_augmented*) dengan dataset asli (*D*). Hal ini dilakukan setelah mereset indeks dari dataset asli untuk memastikan kontinuitas. Fungsi MLSOL kemudian dipanggil, mengambil dataset asli, jumlah tetangga terdekat (*k*), rasio *oversampling* (*P*), dan berat sampel awal opsional sebagai input. Hasilnya adalah sebuah dataset yang telah diisi dengan sampel sintetis yang mencerminkan karakteristik setiap kluster. Dalam konteks ini, setiap kluster sekarang memiliki representasi instan yang lebih seimbang berkat strategi *oversampling* yang cermat yang diterapkan oleh MLSOL.

Proses penggabungan kluster ini membentuk dasar dari dataset baru yang lebih seimbang dan representatif. Ini menjadi langkah penting untuk memastikan bahwa strategi *oversampling* yang diimplementasikan pada tingkat kluster dapat terintegrasi secara efektif dengan dataset asli. Penggabungan ini menciptakan keseimbangan harmonis antara struktur inheren yang terungkap oleh UCLSO dan *oversampling* yang ditargetkan oleh MLSOL. Bobot sampel yang diperbarui (*sample_weights*) memainkan peran kunci dalam mempertahankan signifikansi setiap instan selama proses penggabungan ini. Dengan demikian, langkah *Combine Over-Sampled Clusters* memiliki manfaat besar dalam meningkatkan kualitas dataset, memfasilitasi pelatihan model yang lebih baik untuk tugas klasifikasi multilabel berikutnya.

Secara keseluruhan, integrasi yang cermat antara *oversampling* pada kedua tingkat kluster dan instan berkontribusi pada efektivitas keseluruhan pendekatan ini dalam mengatasi tantangan ketidakseimbangan kelas dalam dataset multilabel. Dataset yang dihasilkan dari langkah ini siap digunakan untuk melatih model yang kokoh, dengan representasi distribusi kelas yang lebih komprehensif.

E. Random Forest Regressor

Pada tahap awal, dilakukan pembersihan data dengan menerapkan *Modified Multi-Label Synthetic Oversampling* (MLSOL) untuk mengatasi ketidakseimbangan kelas pada dataset multilabel. MLSOL menghasilkan contoh sintetis dengan mempertimbangkan bobot sampel yang diperbarui, terutama fokus pada kelas minoritas, sehingga menciptakan distribusi kelas yang lebih seimbang. Setelah berhasil menggabungkan kluster hasil dari MLSOL, langkah selanjutnya adalah evaluasi dan persiapan dataset untuk pelatihan model. Dataset diperkaya dengan kluster yang dihasilkan, dan dimensi serta informasi label dan bobot sampel yang diperbarui ditampilkan. Dataset yang telah diperkaya kemudian dibagi menjadi data latih dan data uji menggunakan fungsi train_test_split, dengan proporsi 80% dan 20% secara berturut-turut. Pembagian ini penting untuk memastikan representasi yang seimbang dalam melatih dan menguji model.

Langkah selanjutnya melibatkan inisialisasi model dasar, yaitu *Random Forest Regressor*, dan bobot sampel. Model dasar dan bobot sampel diinisialisasi sebelum memulai iterasi peningkatan. Setiap iterasi dimulai dengan tahap *Unsupervised Clustering with Label-Specific Oversampling* (UCLSO), di mana dataset diperkaya dengan memahami pola kluster pada data berdasarkan fitur-fitur yang relevan. Setelah UCLSO, dilanjutkan dengan *Modified Multi-Label Synthetic Oversampling* (MLSOL) untuk menghasilkan contoh sintetis yang lebih sesuai dengan bobot sampel yang diperbarui. Selama iterasi, model dasar diperbarui berdasarkan informasi dari kedua tahap sebelumnya. Kesalahan klasifikasi dievaluasi, dengan penekanan pada kelas minoritas, dan pembaruan bobot sampel memberikan penyesuaian dinamis.

Pada akhir iterasi, model yang telah ditingkatkan diuji pada dataset pengujian, dan hasil prediksi dievaluasi. Pendekatan ini bertujuan meningkatkan akurasi klasifikasi pada kelas fungsional yang kurang umum serta menangani ketidakseimbangan kelas. Rangkaian langkah ini menciptakan mekanisme pembelajaran adaptif, memastikan bahwa model dapat fokus pada contoh yang sulit dan kelas minoritas, sehingga menghasilkan hasil yang lebih kuat pada dataset.

Dalam konteks pengolahan dataset tidak seimbang, tahapan *Random Forest Regressor* menjadi esensial. Model awal diinisialisasi dengan *Random Forest Regressor*, memanfaatkan 50 pohon keputusan dengan seed acak 42. Pilihan ini didasarkan pada keandalan algoritma tersebut dalam menangani kompleksitas klasifikasi dan regresi, terutama dalam konteks fitur-fitur nonlinear dan interaksi yang kompleks. Inisialisasi dengan *Random Forest Regressor* memberikan dasar yang kuat sebelum peningkatan lebih lanjut. Pada setiap iterasi, *Random Forest Regressor* dilakukan dengan penekanan khusus pada kelas minoritas, yang seringkali kurang umum atau terwakili secara minimal dalam dataset biologis. *Random Forest Regressor* secara dinamis menyesuaikan bobot sampel, memastikan responsivitas terhadap variasi dalam fungsionalitas protein yang mungkin kurang umum.

Proses ini memberikan model kemampuan adaptasi terhadap variasi dalam dataset, sesuai dengan evolusi alamiah dalam konteks biologi. Tujuan utamanya adalah meningkatkan akurasi klasifikasi secara keseluruhan, terutama pada kelas fungsional yang jarang terjadi. *Random Forest Regressor* cocok untuk dataset dengan ketidakseimbangan kelas dan distribusi kelas yang kompleks, seperti dalam kasus dataset fungsionalitas protein[24]. Hasil yang diperoleh tidak hanya meningkatkan akurasi model, tetapi juga meningkatkan pemahaman terhadap fungsionalitas data secara umum, mendukung pengembangan lebih lanjut dalam bidang komputasional.

IV. HASIL DAN PEMBAHASAN

A. Dataset

Analisis ini melibatkan penelitian terhadap dataset *Yeast, Genbase, Enron, Scene Image, Flag*, dan dataset medis. Pemilihan Yeast dan Genbase sebagai subjek penelitian bertujuan memberikan wawasan mendalam tentang fungsi sel ragi dan klasifikasi fungsi protein, menghadapi tantangan ketidakseimbangan distribusi label. Oleh karena itu, pendekatan *oversampling* seperti UCLSO, MLSOL, dan *Random Forest Regressor* diimplementasikan untuk meningkatkan adaptasi model terhadap karakteristik kompleks kedua dataset biologis ini. Dataset Enron,

yang berfokus pada data teks dari perusahaan Enron Corporation, menjadi subjek penelitian terpisah dalam konteks analisis data teks dan jaringan sosial. UCLSO digunakan untuk mengeksplorasi dependensi antar label pada data teks Enron, sementara MLSOL dan *Random Forest Regressor* membantu mengatasi ketidakseimbangan distribusi label.

Analisis juga melibatkan dataset medis, yang mencerminkan representasi tekstual dari catatan medis dengan label yang mencakup beragam kategori diagnosa, prosedur, dan kondisi kesehatan. Pendekatan *oversampling* seperti UCLSO, MLSOL, dan *Random Forest Regressor* diterapkan untuk mengatasi ketidakseimbangan distribusi label pada data teks medis. Penentuan parameter khusus, eksperimen dengan nilai k yang berbeda, dan penyesuaian parameter pada metode *oversampling* dan *Random Forest Regressor* menjadi langkah kritis dalam mengoptimalkan performa model pada dataset medis. Analisis ini mencakup dataset *Scene Image* dan *Flag*, yang berfokus pada klasifikasi gambar dan pengenalan bendera. Pendekatan *oversampling* menggunakan UCLSO, MLSOL, dan *Random Forest Regressor* diimplementasikan untuk mengatasi ketidakseimbangan distribusi label pada kedua dataset ini. Pengaturan parameter yang cermat dan adaptasi pendekatan terhadap karakteristik khusus setiap dataset memberikan kontribusi signifikan terhadap hasil analisis secara keseluruhan. Jurnal ini memberikan pemahaman yang mendalam tentang strategi *oversampling* dalam mengatasi tantangan ketidakseimbangan label di berbagai domain dataset, termasuk data medis.

B. Perbandingan Metode

Dalam bab ini, merinci analisis perbandingan teknis antara empat metode yang berbeda untuk mengatasi ketidakseimbangan label pada dataset multilabel, dengan fokus utama pada tiga jenis data: data teks, gambar, dan dataset biologis. Metode pertama yang dievaluasi adalah kombinasi UCLSO + MLSOL + Random Forest Regressor. Implementasi teknisnya melibatkan pemanfaatan UCLSO untuk mengeksplorasi dependensi antar label pada data. MLSOL kemudian diterapkan untuk *oversampling* sintetis dengan mempertimbangkan hasil *clustering*, sementara proses *Random Forest Regressor* diaplikasikan secara iteratif untuk meningkatkan adaptasi model terhadap kompleksitas dataset. Kelebihan utama metode ini terletak pada penanganan holistik terhadap ketidakseimbangan label, mencapai representasi yang seimbang pada semua jenis dataset. Namun, perlu dicatat bahwa metode ini melibatkan komputasi intensif akibat proses *Random Forest Regressor*, dan penyetelan parameter jumlah kluster pada UCLSO memerlukan pertimbangan yang hati-hati.

Metode kedua yang dievaluasi adalah UCLSO + MLSOL + KNN, dengan fokus utama pada penanganan ketidakseimbangan label melalui eksplorasi dependensi label menggunakan UCLSO dan *oversampling* sintetis dengan MLSOL, tanpa melibatkan proses *Random Forest Regressor*. Kelebihannya meliputi implementasi yang lebih efisien dibandingkan dengan metode yang melibatkan *Random Forest Regressor*. Meskipun demikian, metode ini mungkin kurang efektif pada dataset yang sangat tidak seimbang dengan menambahkan KNN (K-Nearest Neighbors) sebagai klasifikasi dasar untuk memperkuat model. Metode ketiga yang dianalisis adalah UCLSO, di mana implementasinya melibatkan penggunaan UCLSO untuk mengeksplorasi dependensi label pada data. Fokusnya adalah pada pengelompokan dan pengurangan sampel mayoritas tanpa melibatkan *oversampling* sintetis diikuti oleh pemanfaatan KNN sebagai model klasifikasi. Kelebihannya melibatkan implementasi yang lebih cepat dengan kompromi yang baik terhadap keseimbangan label, terutama efektif pada dataset dengan jumlah sampel mayoritas yang besar. Keterbatasan metode ini mungkin terletak pada kurangnya efektivitas pada dataset yang sangat tidak seimbang dan kurangnya *oversampling* sintetis yang dapat mengurangi representasi label minoritas.

Metode keempat yang dibandingkan adalah MLSOL. Metode ini difokuskan pada *oversampling* sintetis dengan mempertimbangkan hasil *clustering* pada data dan KNN digunakan sebagai model klasifikasi untuk menggolongkan contoh sintetis yang dihasilkan ke dalam kelas yang sesuai. Kelebihannya terletak pada efektivitas meningkatkan distribusi label pada dataset yang tidak seimbang dan implementasi yang lebih cepat dibandingkan dengan metode yang melibatkan *Random Forest Regressor*. Namun, keterbatasannya mungkin terletak pada tidak langsungnya penanganan dependensi label dan potensial kurang efektif pada dataset dengan struktur dependensi yang kompleks.

C. Hasil dan Pembahasan

Namun, setelah menerapkan pendekatan UCLSO + MLSOL + Random Forest Regressor, distribusi label mengalami perubahan yang drastis. Distribusi akhir menunjukkan peningkatan yang substansial dalam proporsi setiap label, mencapai nilai yang lebih seimbang pada setiap kelas. Sebagai contoh, distribusi label *NUC* dan *MIT* meningkat secara signifikan, mencapai nilai 0.32935, yang mendekati distribusi label kelas mayoritas seperti *CYT* dan *VAC*. Peningkatan distribusi label ini mencerminkan keberhasilan pendekatan UCLSO + MLSOL + *Random Forest Regressor* dalam mengatasi ketidakseimbangan, terutama pada kelas-kelas minoritas. Metode ini secara efektif menghasilkan variasi dan representasi yang lebih baik pada label-label yang sebelumnya kurang diwakili dalam dataset. Hasil ini memberikan gambaran positif tentang efektivitas pendekatan *oversampling* yang diterapkan dalam meningkatkan keseimbangan distribusi label pada dataset biologis ini.

Dengan demikian, melalui perubahan distribusi label yang signifikan dari sebelum ke sesudah *oversampling*, penelitian ini memberikan bukti kuat bahwa pendekatan UCLSO + MLSOL + *Random Forest Regressor* dapat menjadi solusi yang efektif untuk mengatasi ketidakseimbangan distribusi label pada dataset biologis. Perubahan ini memberikan kontribusi positif terhadap peningkatan kinerja model pada kelas-kelas minoritas, yang pada gilirannya, dapat meningkatkan keakuratan dan generalisasi model pada analisis dataset tersebut.

Analisis hasil evaluasi pada dataset Yeast dan Genbase dengan menggunakan empat metode *oversampling* (UCLSO + MLSOL + *Random Forest Regressor*, UCLSO + MLSOL + KNN, UCLSO + KNN, dan MLSOL+ KNN) memberikan gambaran kinerja relatif masing-masing metode. Metode terbaik adalah UCLSO + MLSOL + *Random Forest Regressor* dengan mencapai presisi, *recall*, *F1-score*, dan akurasi tertinggi pada kedua dataset. Keberhasilan ini dapat *distributing* pada kekuatan kombinasi UCLSO untuk menangani kelas mayoritas, MLSOL untuk menghasilkan variasi pada kelas minoritas, dan proses *Random Forest Regressor* untuk meningkatkan adaptasi model secara iteratif. Namun, UCLSO + MLSOL + KNN menunjukkan kinerja baik pada Genbase tetapi mengalami penurunan signifikan pada Yeast, terutama dalam akurasi. Hal ini mungkin menunjukkan bahwa kombinasi UCLSO dan MLSOL tanpa *Random Forest Regressor* kurang efektif dalam menangani kompleksitas dependensi antar label pada Yeast, mengakibatkan ketidakseimbangan yang lebih besar dan *overfitting* yang berlebihan.

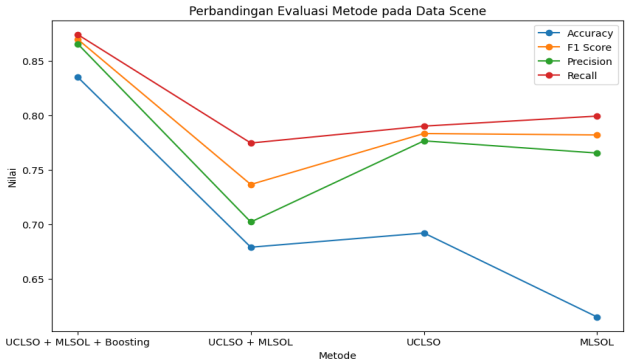
TABEL I
DISTRIBUSI LABEL DATASET SEBELUM DAN SESUDAH OVERSAMPLING

Label	Distribusi Awal	Distribusi Akhir
CYT	0.309183	0.32935
NUC	0.004212	0.32935
MIT	0.022746	0.32935
ME3	0.030329	0.32935
ME2	0.037911	0.32935
ME1	0.108677	0.32935
EXC	0.161752	0.32935
VAC	0.289806	0.32935
POX	0.015164	0.32935
ERL	0.020219	0.32935

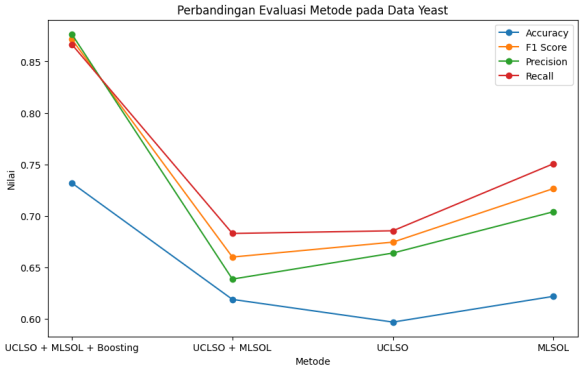
TABEL II
PERBANDINGAN HASIL OVERSAMPLING DATA

Metode	Evaluasi	Data Biologi		Data Teks		Data Image	
		Yeast	Genbase	Enron	Medical	Scene	Flags
UCLSO+MLSOL+Random Forest Regressor	Precision	0.8764	0.8666	0.8721	0.8769	0.8655	0.8577
	Recall	0.8665	0.8908	0.8234	0.8544	0.8740	0.8995
	F1-Score	0.8714	0.8785	0.8471	0.8655	0.8697	0.8781
	Accuracy	0.8630	0.8796	0.8342	0.8640	0.8628	0.8693
UCLSO+MLSOL+KNN	Precision	0.6387	0.7712	0.7618	0.8362	0.7021	0.8228
	Recall	0.6830	0.7671	0.7946	0.8382	0.7745	0.8133
	F1-Score	0.6601	0.7692	0.7779	0.8372	0.7365	0.8181
	Accuracy	0.6190	0.5680	0.5830	0.6940	0.6790	0.8030
UCLSO+KNN	Precision	0.6640	0.6843	0.6173	0.7928	0.7765	0.7510
	Recall	0.6856	0.6466	0.6971	0.8341	0.7900	0.7946
	F1-Score	0.6746	0.6649	0.6548	0.8129	0.7832	0.7722
	Accuracy	0.5970	0.5280	0.5150	0.6870	0.6920	0.8050
MLSOL+KNN	Precision	0.7041	0.6010	0.6005	0.6948	0.7653	0.7295
	Recall	0.7507	0.6430	0.6397	0.6455	0.7992	0.7083
	F1-Score	0.7266	0.6213	0.6195	0.6692	0.7819	0.7188
	Accuracy	0.6220	0.4910	0.5340	0.6950	0.6150	0.6580

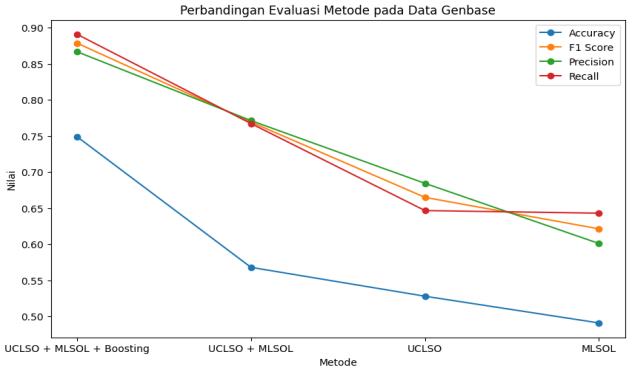
Selanjutnya, pada dataset Enron dan Medical, metode UCLSO + MLSOL + Random Forest Regressor kembali menunjukkan performa tertinggi, memanfaatkan kekuatan kombinasi UCLSO untuk menangani kelas mayoritas, MLSOL untuk variasi pada kelas minoritas, dan proses *Random Forest Regressor* untuk meningkatkan adaptasi model. Meskipun UCLSO + MLSOL tanpa *Random Forest Regressor* KNN tetap memberikan kinerja yang baik dan lebih efektif dengan melibatkan peran *Random Forest Regressor*. UCLSO, yang fokus pada kelas mayoritas, menunjukkan performa baik pada dataset Medical tetapi mengalami penurunan signifikan pada dataset Enron, menandakan ketidakefektifan metode ini dalam menangani kompleksitas dependensi antar label pada data teks Enron. MLSOL, fokus pada *oversampling* sintetis, menunjukkan hasil yang baik pada dataset Medical tetapi kurang efektif pada dataset Enron, kemungkinan karena kurangnya eksplorasi dependensi antar label pada data teks Enron.



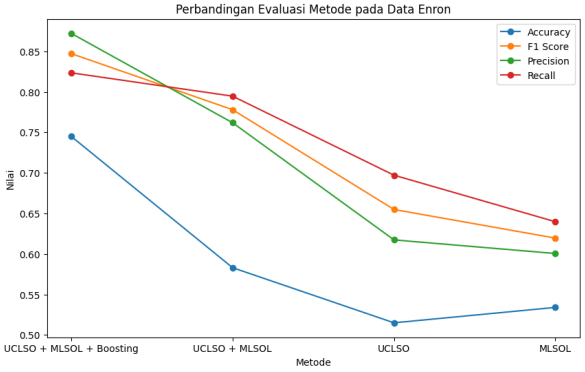
Gambar 1. Grafik Perbandingan Evaluasi Metode pada Data Scene



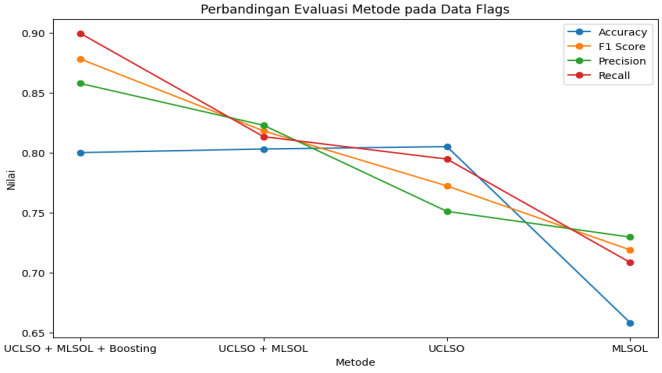
Gambar 2. Grafik Perbandingan Evaluasi Metode pada Data Yeast



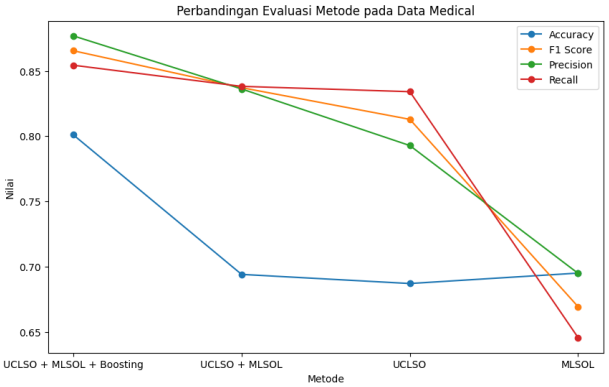
Gambar 3. Grafik Perbandingan Evaluasi Metode pada Data Genbase



Gambar 4. Grafik Perbandingan Evaluasi Metode pada Data Enron



Gambar 5. Grafik Perbandingan Evaluasi Metode pada Data Flags



Gambar 6. Grafik Perbandingan Evaluasi Metode pada Data Medical

Terakhir, pada dataset Scene dan Flags, UCLSO + MLSOL + Random Forest Regressor kembali menonjol sebagai metode terbaik dengan presisi, *recall*, *F1-score*, dan akurasi tertinggi pada kedua dataset. UCLSO + MLSOL + *Random Forest Regressor* tetap kinerja baik, terutama pada dataset Flags, namun terdapat penurunan akurasi pada dataset Scene. Ini mungkin menunjukkan bahwa dependensi antar label pada data gambar Scene lebih efektif diatasi dengan melibatkan proses *Random Forest Regressor*. UCLSO, dengan fokus pada kelas mayoritas,

menunjukkan performa baik pada dataset Flags tetapi mengalami penurunan yang signifikan pada dataset Scene, menandakan ketidakseimbangan dalam menangani kompleksitas dependensi antar label pada data gambar Scene. MLSOL, yang fokus pada *oversampling* sintetis, menunjukkan hasil yang baik pada dataset Flags tetapi kurang efektif pada dataset Scene, kemungkinan karena kurangnya eksplorasi dependensi antar label yang rumit pada data gambar Scene.

Secara keseluruhan, analisis ini memberikan wawasan mendalam tentang kinerja relatif metode *oversampling* dalam konteks dataset yang berbeda. Penggunaan gabungan UCLSO, MLSOL, dan *Random Forest Regressor* memberikan solusi yang efektif dalam menangani kompleksitas dan ketidakseimbangan yang signifikan pada berbagai jenis dataset. Analisis ini memberikan kontribusi yang berharga untuk memahami strategi *oversampling* dalam mengatasi tantangan ketidakseimbangan label di berbagai domain dataset.

V. KESIMPULAN

Penelitian ini mengeksplorasi strategi penanganan ketidakseimbangan label pada dataset multi-label, yang semakin krusial seiring pertumbuhan data dan kompleksitasnya. Ketersediaan data yang melimpah namun tidak seimbang mengakibatkan tantangan dalam klasifikasi, terutama ketika data kelas mayoritas mendominasi data kelas minoritas. Fenomena ini tidak hanya relevan untuk klasifikasi biner atau multi-kelas, tetapi juga pada data multi-label, yang kini semakin penting dalam berbagai aplikasi. Dua pendekatan utama yang dieksplorasi dalam penelitian ini adalah *Synthetic Oversampling of Multi-Label Data Based on Local Label Distribution* (MLSOL) dan *Integrating Unsupervised Clustering and Label-specific Oversampling to Tackle Imbalanced Multi-Label Data* (UCLSO). MLSOL memusatkan perhatian pada kelas minoritas untuk meningkatkan variasi label, sementara UCLSO fokus pada kelas mayoritas untuk mencegah dominasi yang berlebihan. Namun, penelitian ini menyimpulkan bahwa kombinasi kedua pendekatan ini dengan menambahkan proses *Random Forest Regressor* menghasilkan keseimbangan yang optimal dalam menangani ketidakseimbangan dan kompleksitas label pada dataset multi-label. Hasilnya terlihat jelas dalam berbagai jenis dataset, termasuk biologis, teks, dan gambar. UCLSO + MLSOL + Random Forest Regressor menunjukkan kinerja tertinggi dengan presisi, *recall*, *F1-score*, dan akurasi yang konsisten tinggi. Meskipun terdapat metode yang menunjukkan kinerja baik pada jenis dataset tertentu, seperti UCLSO + MLSOL + KNN pada dataset biologis, pendekatan ini tidak selalu efektif pada jenis dataset lainnya. Oleh karena itu, kesimpulan utama adalah bahwa kombinasi holistik UCLSO + MLSOL + Random Forest Regressor memberikan solusi terbaik untuk menangani ketidakseimbangan label pada dataset multi-label, memanfaatkan kelebihan masing-masing pendekatan untuk hasil yang optimal secara konsisten. Kesimpulan ini diharapkan dapat memberikan landasan untuk pengembangan strategi *oversampling* yang lebih canggih dan efektif dalam mengatasi ketidakseimbangan dataset multi-label di masa depan.

DAFTAR PUSTAKA

[1] Q. Meidianingsih dan D. E. W. Meganingtyas, "Analisis Perbandingan Performa Metode Ensemble Dalam Menangani Imbalanced *Multi-class* Classification," *J. Apl. Stat. Komputasi Stat.*, vol. 14, no. 2, hal. 13–21, 2022, doi: 10.34123/jurnalasks.v14i2.335.

[2] H. Duan, Y. Wei, P. Liu, dan H. Yin, "A novel ensemble framework based on K-means and resampling for imbalanced data," *Appl. Sci.*, vol. 10, no. 5, 2020, doi: 10.3390/app10051684.

[3] M. A. Tahir, J. Kittler, dan A. Bouridane, "Multilabel classification using heterogeneous ensemble of multi-label classifiers," *Pattern Recognit. Lett.*, vol. 33, no. 5, hal. 513–523, 2012, doi: 10.1016/j.patrec.2011.10.019.

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

[5] M. Błaszczyk dan J. Jedrzejowicz, "Framework for imbalanced data classification," *Procedia Comput. Sci.*, vol. 192, hal. 3477–3486, 2021, doi: 10.1016/j.procs.2021.09.121.

[6] R. Rastogi dan S. Mortaza, "Imbalance multi-label data learning with label specific features," *Neurocomputing*, vol. 513, hal. 395–408, 2022, doi: 10.1016/j.neucom.2022.09.085.

[7] J. J. Rodríguez, J. F. Díez-Pastor, Á. Arnaiz-González, dan L. I. Kuncheva, "Random Balance ensembles for multiclass imbalance learning," *Knowledge-Based Syst.*, vol. 193, hal. 105434, 2020, doi: 10.1016/j.knosys.2019.105434.

[8] A. Fernández, V. López, M. Galar, M. J. Del Jesus, dan F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Syst.*, vol. 42, hal. 97–110, 2013, doi: 10.1016/j.knosys.2013.01.018.

[9] F. Marbun, A. Baizal, dan M. A. Bijaksana, "Perpaduan Combined Sampling Dan Ensemble of Support Vector Machine (Ensvm) Untuk Menangani Kasus Churn Prediction Perusahaan Telekomunikasi," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 8, no. 2. hal. 43, 2010, doi: 10.12962/j24068535.v8i2.a316.

[10] S. Chen, R. Wang, J. Lu, dan X. Wang, "Stable matching-based two-way selection in multi-label active learning with imbalanced data," *Inf. Sci. (Ny).*, vol. 610, hal. 281–299, 2022, doi: 10.1016/j.ins.2022.07.182.

[11] B. Liu dan G. Tsoumakas, "Synthetic *Oversampling* of Multi-label Data Based on Local Label Distribution," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11907 LNAI, hal. 180–193, 2020, doi: 10.1007/978-3-030-46147-8_11.

[12] T. Zhu, C. Luo, Z. Zhang, J. Li, S. Ren, dan Y. Zeng, "Minority *oversampling* for imbalanced time series classification," *Knowledge-Based Syst.*, vol. 247, hal. 108764, 2022, doi: 10.1016/j.knosys.2022.108764.

[13] A. N. Tarekegn, M. Giacobini, dan K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognit.*, vol. 118, hal. 107965, 2021, doi: 10.1016/j.patcog.2021.107965.

[14] M. Koziarski, "Radial-Based Undersampling for imbalanced data classification," *Pattern Recognit.*, vol. 102, 2020, doi: 10.1016/j.patcog.2020.107262.

[15] L. Cai, H. Wang, F. Jiang, Y. Zhang, dan Y. Peng, "A new *clustering* mining algorithm for multi-source imbalanced location data," *Inf. Sci. (Ny).*, vol. 584, hal. 50–64, 2022, doi: 10.1016/j.ins.2021.10.029.

[16] E. K. Y. Yapp, X. Li, W. F. Lu, dan P. S. Tan, "Comparison of base classifiers for multi-label learning," *Neurocomputing*, vol. 394, hal. 51–60, 2020, doi: 10.1016/j.neucom.2020.01.102.

[17] G. Wei, W. Mu, Y. Song, dan J. Dou, "An improved and random synthetic minority *oversampling* technique for imbalanced data," *Knowledge-Based*

Syst., vol. 248, hal. 108839, 2022, doi: 10.1016/j.knosys.2022.108839.
[18] J. Dou, Z. Gao, G. Wei, Y. Song, dan M. Li, "Switching synthesizing-incorporated and cluster-based synthetic *oversampling* for imbalanced binary classification," *Eng. Appl. Artif. Intell.*, vol. 123, no. April 2022, hal. 106193, 2023, doi: 10.1016/j.engappai.2023.106193.
[19] T. G.S., Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, dan S. Chennupati, "An extension of Synthetic Minority *Oversampling* Technique based on Kalman filter for imbalanced datasets," *Mach. Learn. with Appl.*, vol. 8, no. January, hal. 100267, 2022, doi: 10.1016/j.mlwa.2022.100267.
[20] F. Charte, A. J. Rivera, M. J. del Jesus, dan F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, hal. 3–16, 2015, doi: 10.1016/j.neucom.2014.08.091.
[21] P. Sadhukhan, A. Pakrashi, S. Palit, dan B. Namee, "Integrating Unsupervised *Clustering* and Label-Specific *Oversampling* to Tackle Imbalanced Multi-Label Data," hal. 489–498, 2023, doi: 10.5220/0011901200003393.
[22] J. Ren, Y. Wang, M. Mao, dan Y. ming Cheung, "Equalization ensemble for large scale highly imbalanced data classification," *Knowledge-Based Syst.*, vol. 242, hal. 108295, 2022, doi: 10.1016/j.knosys.2022.108295.
[23] D. C. Li, S. Y. Wang, K. C. Huang, dan T. I. Tsai, "Learning class-imbalanced data with region-impurity synthetic minority *oversampling* technique," *Inf. Sci. (Ny).*, vol. 607, hal. 1391–1407, 2022, doi: 10.1016/j.ins.2022.06.067.
[24] J. H. J. Einmahl dan Y. He, "Pr ep rin ot pe er r Pr ep er ed," vol. 4, no. 2, hal. 0–3, 2021.
[25] W. Lu, Z. Li, dan J. Chu, "Adaptive Ensemble Undersampling-Boost: A novel learning framework for imbalanced data," *J. Syst. Softw.*, vol. 132, hal. 272–282, 2017, doi: 10.1016/j.jss.2017.07.006.
[26] M. Zheng *et al.*, "UFFDFR: Undersampling framework with denoising, fuzzy c-means *clustering*, and representative sample selection for imbalanced data classification," *Inf. Sci. (Ny).*, vol. 576, hal. 658–680, 2021, doi: 10.1016/j.ins.2021.07.053.