JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Pola Dan Prediksi Churn: Hybrid Segmentasi SOM+K-Means Dan Klasifikasi Machine Learning

Rizka Rahmadhani, Ermatita*

Fakultas Ilmu Komputer, Program Studi Magister Ilmu Komputer, Universitas Sriwijaya, Palembang, Indonesia
Email: [1]rizka.rahmadhani02@gmail.com, [2,*]ermatita@unsri.ac.id
Email Penulis Korespondensi: ermatita@unsri.ac.id
Submitted **10-12-2025;** Accepted **27-01-2026;** Published **28-02-2026**

Abstrak

Churn pelanggan merupakan tantangan signifikan dalam industri perbankan, yang dapat berdampak besar terhadap profitabilitas dan keberlanjutan perusahaan. Penanganan churn pelanggan umumnya dilakukan dengan pendekatan klasifikasi biner yang seringkali tidak memberikan wawasan yang cukup mendalam mengenai karakteristik pelanggan. Penelitian ini mengusulkan untuk menganalisis churn dengan segmentasi pelanggan sebagai langkah awal sebelum klasifikasi, guna memberikan pemahaman yang lebih jelas mengenai karakteristik setiap segmen. Dengan itu, penelitian ini mengkombinasikan Self-Organizing Map (SOM) dan K-Means untuk pembentukan klaster yang lebih mudah diinterpretasi. Penerapan model SOM+K-Means ini dilakukan untuk segmentasi serta pemetaan visual, yang kemudian dianalisis untuk mengidentifikasi kelompok pelanggan yang berisiko churn serta fitur-fitur sebagai faktor dominan yang memengaruhinya. Kemudian hasil label klaster dipakai menjadi fitur untuk dilanjutkan dengan klasifikasi menggunakan tiga algoritma pembelajaran mesin yaitu Support Vector Machine (SVM), Random Forest (RF), dan XGBoost (XGB). Dalam tahap klasifikasi, Synthetic Minority Oversampling Technique (SMOTE) dan GridSearchCV digunakan untuk menyeimbangkan kelas dan memilih parameter yang memberikan hasil terbaik dari model. XGB menunjukkan kinerja terbaik dengan akurasi 85% dan AUC 85%. Temuan ini menunjukkan segmentasi pelanggan menggunakan SOM + K-Means memungkinkan perusahaan untuk merancang strategi retensi yang lebih tepat sasaran dan XGB sebagai model dengan kinerja yang baik untuk prediksi churn. Penelitian ini memberikan kontribusi penting dalam penerapan teknik clustering dan klasifikasi machine learning untuk analisis churn di sektor perbankan, yang dapat diterapkan untuk meningkatkan loyalitas pelanggan dan mengurangi tingkat churn.

Kata Kunci: Churn Bank; Segmentasi Pelanggan; Klasterisasi Hibrid; Prediksi Churn; Pembelajaran Mesin

Abstract

Customer churn is a significant challenge in the banking industry, which can have a substantial impact on the profitability and long-term sustainability. Customer churn management is typically addressed using binary classification approaches, which often fail to provide the depth needed to understand customer characteristics. This study proposes addressing churn through customer segmentation as an preliminary step before classification, offering a clearer and deeper understanding of each segment's characteristics. The research combines Self-Organizing Map (SOM) and K-Means clustering to create interpretable segments. The SOM+K-Means model is used for segmentation and visual mapping, which helps identify customer groups at risk of churn and the key features influencing these risks. Cluster labels are then used as features for classification using three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). In the classification phase, the Synthetic Minority Oversampling Technique (SMOTE) and GridSearchCV are applied to address class imbalance and optimize model parameters. XGB outperformed the other models with an accuracy of 85% and an AUC score of 85%. These results highlight that customer segmentation with SOM+K-Means enables more effective churn management strategies, while XGB proves to be a strong model for churn prediction. This research contributes to the application of clustering and machine learning classification techniques in churn analysis within the banking industry, offering a pathway to better customer retention strategies and lower churn rates.

Keywords: Bank Churn; Customer Segmentation; Hybrid Clustering; Churn Prediction; Machine Learning

1. PENDAHULUAN

Churn pelanggan atau berhentinya pelanggan dari layanan merupakan salah satu tantangan bisnis utama yang berdampak signifikan pada profitabilitas perusahaan [1]. Biaya untuk mendapatkan pelanggan baru jauh lebih tinggi dibandingkan mempertahankan pelanggan yang sudah ada, sehingga strategi retensi pelanggan menjadi prioritas utama [2]. Namun, pelanggan yang berisiko churn tidak selalu mudah diidentifikasi karena faktor yang memengaruhinya dapat bersifat kompleks dan beragam, seperti tipe kontrak, metode pembayaran, tingkat penggunaan layanan, dan pengalaman pelanggan secara keseluruhan [3]. Dalam sektor telekomunikasi, ditekankan pentingnya integrasi prediksi churn dengan segmentasi pelanggan agar perusahaan tidak hanya mengetahui siapa yang berpotensi churn, tetapi juga mampu memahami karakteristik dan faktor yang memengaruhinya sehingga strategi retensi dapat diarahkan secara lebih tepat sasaran [4]. Demikian pula pada sektor *business-to-business*, ditemukan bahwa churn meskipun terjadi dengan frekuensi lebih rendah, namun memiliki dampak finansial yang jauh lebih besar karena nilai transaksi yang tinggi, sehingga memprediksi dan mencegah churn menjadi sangat krusial [5].

Dalam menghadapi masalah churn, pada awalnya perusahaan lebih banyak mengandalkan strategi konvensional sebelum beralih pada pendekatan berbasis komputasi dimana manajemen churn tradisional lebih bersifat reaktif, menekankan pada kampanye *win-back* dan penawaran retensi umum [4]. Berbagai penelitian dengan pendekatan *machine learning* pun digunakan untuk memprediksi churn, mulai dari *Logistic Regression* [6], [7], *Decision Tree* [8], [9], *Neural Network* [10], [11], [12], hingga *Ensemble Methods* [1], [13], [14], dengan hasil yang cukup menjanjikan. Namun demikian, sebagian besar penelitian lebih berfokus pada prediksi churn sebagai klasifikasi biner churn vs non-churn, sementara aspek segmentasi pelanggan dan interpretasi faktor penyebab churn masih terbatas [6], [10], [11], [12], [13],

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

[14]. Padahal, perusahaan tidak dapat memperlakukan semua pelanggan dengan cara yang sama, misalnya pelanggan bernilai tinggi memerlukan strategi retensi yang lebih intensif dibandingkan pelanggan bernilai rendah [4]. Selain itu, pendekatan *supervised learning* sendiri sering kali terbatas pada ketepatan prediksi, namun kurang mampu memberikan visualisasi struktur data maupun segmentasi yang mudah diinterpretasikan oleh pengambil keputusan [15].

Untuk mengisi celah tersebut, penelitian ini mengusulkan pendekatan *hybrid clustering* menggunakan *Self-Organizing Maps* (SOM) dan K-Means untuk segmentasi sebelum melakukan klasifikasi. SOM berperan sebagai metode *unsupervised* yang mampu memproyeksikan data berdimensi tinggi ke dalam peta topologis yang lebih mudah divisualisasikan dan dipahami [16]. Hasil proyeksi ini kemudian diperhalus dengan K-Means yang berfungsi mempertegas batas antar klaster, menghasilkan struktur kelompok yang lebih stabil dan terdefinisi [17]. Beberapa penelitian di berbagai domain[17], [18], [19], [20]. telah menunjukkan bahwa kombinasi SOM dan K-Means mampu meningkatkan kinerja sekaligus mengurangi kompleksitas dibandingkan metode tunggal. Misalnya, dalam deteksi intrusi jaringan, *hybrid* SOM dan K-Means meningkatkan kualitas deteksi sekaligus mempercepat konvergensi [17]. Pada bidang kesehatan, dilaporkan bahwa hybrid SOM dan K-means menghasilkan kualitas *clustering* lebih tinggi dibanding algoritma klasik [18]. Dalam bidang geosains, memperlihatkan bahwa multi atribut SOM dan K-Means dapat menangkap fitur non-linear data seismik sehingga hasil clustering lebih presisi [20]. Selain itu, penelitian dalam bidang sistem tenaga listrik juga membuktikan bahwa metode *hybrid* SOM dan K-Means memiliki tingkat kinerja yang tinggi dan otomatisasi yang baik dalam *clustering* karakteristik beban pengguna, sehingga menegaskan keunggulannya dalam menangani data skala besar dan kompleks [19]. Secara umum, literatur menunjukkan bahwa integrasi SOM dengan metode clustering lain mampu meningkatkan efisiensi dan kualitas hasil. Kotyrba dkk. [21] membuktikan bahwa kombinasi SOM dengan algoritma *clustering* konvensional, termasuk K-Means, memberikan *average rand index* yang cukup tinggi (0,927), meskipun masih di bawah kombinasi SOM+CLARA (0,950), dan SOM+CURE (0,947). Temuan ini menegaskan bahwa SOM+K-Means menjadi pilihan yang layak karena keseimbangan antara kinerja dan kesederhanaan komputasi.

Untuk membentuk model penanganan churn yang interpretatif, hasil clustering juga dapat dimanfaatkan sebagai informasi perilaku ringkas melalui label *cluster* yang kemudian dimasukkan menjadi fitur saat dilakukan klasifikasi, sehingga berpotensi memperkaya proses prediksi churn karena merepresentasikan pola multi-dimensi pelanggan yang tidak selalu tertangkap secara eksplisit oleh fitur asli [22]. Mengingat data churn umumnya bersifat tidak seimbang atau *imbalanced* [23], penelitian ini menerapkan *Synthetic Minority Oversampling Technique* (SMOTE) pada data pelatihan agar meningkatkan sensitivitas model terhadap kelas churn [24]. Selanjutnya, tahap klasifikasi dilakukan dengan membandingkan beberapa algoritma yang populer digunakan karena performanya stabil, seperti *Support Vector Machine* [25], [26], *Random Forest* [27], [28], dan XGBoost [29], [30] yang akan dioptimasi menggunakan *grid search* dengan *GridSearchCV*, sehingga diperoleh model prediksi churn yang kompetitif sekaligus tetap memiliki interpretasi berbasis segmen.

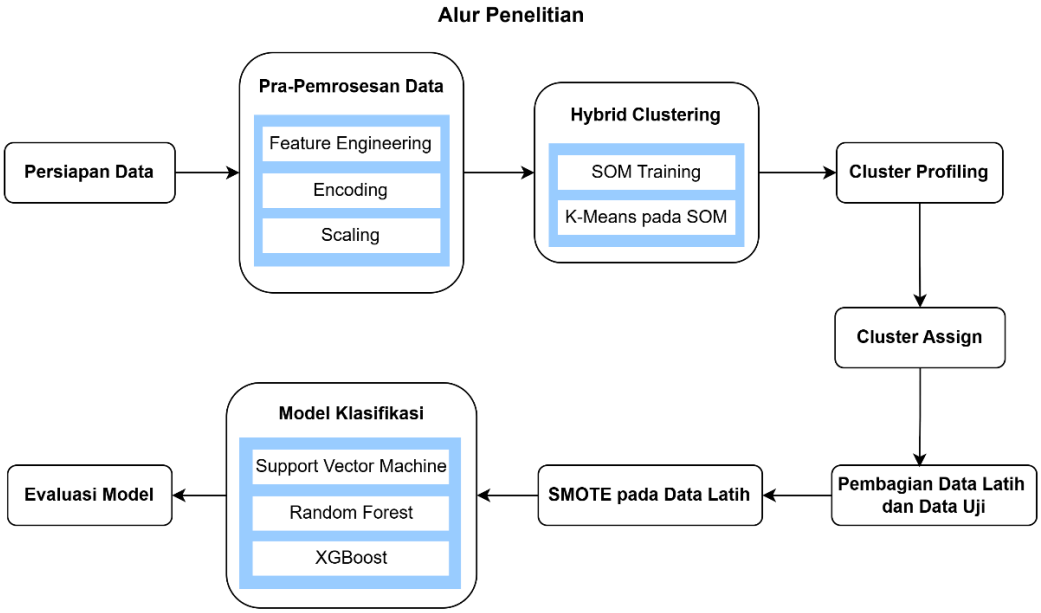
Dengan demikian metode ini diterapkan dengan tujuan untuk memahami struktur kelompok pelanggan melalui pendekatan *unsupervised*, yang tidak dapat dicapai oleh metode *supervised* sendiri meskipun memiliki akurasi tinggi. Tahap *clustering* dalam penelitian ini berfokus pada pembentukan segmentasi pelanggan yang interpretatif untuk menghasilkan *insight* bisnis dan dasar strategi yang tidak hanya berdasarkan probabilitas churn, tetapi disesuaikan dengan kebutuhan dan perilaku khas dari setiap segmen pelanggan. Penelitian ini mengusulkan pendekatan dua tahap, yaitu segmentasi pelanggan menggunakan SOM+K-Means, serta klasifikasi churn menggunakan beberapa *algoritma machine learning* dengan SMOTE dan *hyperparameter tuning*.

2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Penelitian ini menggunakan pendekatan kombinasi *hybrid clustering* dan klasifikasi *machine learning* (ML) untuk melakukan segmentasi dan menganalisis pola churn serta memprediksi churn pelanggan. Lebih jelasnya untuk alur penelitian dapat dilihat pada Gambar 1 di bawah. Gambar 1 ini menjelaskan tahapan proses yang dilakukan dalam penelitian ini, dimulai dari persiapan data, pra-pemrosesan data yaitu *feature engineering, encoding,* dan *scaling*, lalu melakukan *hybrid clustering* dengan SOM+K-Means*, cluster profiling* untuk menganalisis karakteristik churn dan visualisasi pola churn*, cluster assign* untuk menjadikan *cluster* sebagai fitur pada klasifikasi, membagi data menjadi data latih serta data uji, lalu menangani data tidak seimbang dengan *Synthetic Minority Over-sampling Technique* (SMOTE), kemudian melakukan pemodelan klasifikasi dengan *Support Vector Machine* (SVM), *Random Forest* (RF), dan *Extreme Gradient Boosting* (XGB), hingga terakhir mengevaluasi model klasifikasi.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 1. Alur Penelitian

2.2 Persiapan Data

Dataset yang digunakan dalam penelitian ini adalah *Bank Customer Churn Prediction Dataset* yang didapatkan dari *platform* Kaggle. Dataset ini terdiri dari 10.000 baris data pelanggan bank dengan 14 fitur yang mencakup karakteristik demografi, perilaku penggunaan layanan, dan status churn. Dataset ini tidak memiliki nilai hilang maupun duplikat. Namun merupakan dataset yang tidak seimbang pada label target *Exited*. Tabel 1 dibawah ini adalah keterangan fitur-fitur yang ada dari dataset yang dipakai.

Tabel 1. Dataset yang Digunakan

No	Fitur	Tipe Data	Deskripsi
1	*RowNumber*	*Integer*	Nomor urut pelanggan
2	*CustomerId*	*Integer*	ID unik pelanggan
3	*Surname*	*String*	Nama keluarga
4	*CreditScore*	*Numeric*	Skor kredit pelanggan
5	*Geography*	*Categorical*	Negara (*France, Spain, Germany*)
6	*Gender*	*Categorical*	Jenis kelamin
7	*Age*	*Numeric*	Usia pelanggan
8	*Tenure*	*Numeric*	Lamanya menjadi nasabah
9	*Balance*	*Numeric*	Saldo rekening
10	*NumOfProducts*	*Numeric*	Jumlah produk bank yang dimiliki
11	*HasCrCard*	*Binary*	Kepemilikan kartu kredit (1 = ya)
12	*IsActiveMember*	*Binary*	Status keaktifan
13	*EstimatedSalary*	*Numeric*	Estimasi pendapatan
14	*Exited (Target)*	*Binary*	Status churn (1 = churn, 0 = tidak churn)

Sebelum digunakan dataset ini melalui tahapan pra-pemrosesan terlebih dahulu, agar dapat menghasilkan analisis yang lebih baik dan dapat diandalkan.

2.3 Pra-pemrosesan Data

Pra-pemrosesan data dilakukan dalam upaya memastikan dataset berada dalam kondisi yang sesuai sebelum dianalisis dan dapat meningkatkan kualitas data [31]. Agar data lebih relevan untuk analisis dan klasifikasi, fitur seperti *RowNumber, CustomerId, Surname,* dan *Geography* tidak digunakan. Label *Exited* dipisahkan untuk model klasifikasi, dan tidak digunakan untuk *clustering*. Lalu dilakukan *feature engineering* atau membuat fitur tambahan untuk memperkaya informasi bisnis dari segmentasi. Beberapa fitur yang ditambahkan ada pada tabel 2 berikut.

Tabel 2. Tambahan *Feature Engineering*

Nama Fitur	Keterangan
BalanceSalaryRatio	*Balance / (EstimatedSalary + 1)*
BalanceAgeRatio	*Balance / (Age + 1)*
ProductHoldingIntensity	*NumOfProducts * IsActiveMember*

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

ProductsTenureRatio	*NumOfProducts / (Tenure + 1)*
ActiveSalary	*IsActiveMember * EstimatedSalary*
CreditScoreBucket	Kategori *CreditScore* kurang dari 500, 600, dan 700, atau lebih
HighBalanceBucket	Kategori *Balance* sama dengan 0, kurang dari 50000, 100000, atau lebih
AgeBucket	Kategori *Age* kurang dari 25, 60, atau lebih

Data ini tidak memiliki nilai hilang maupun duplikat, namun variabel kategori perlu diubah ke bentuk numerik agar dapat diproses oleh algoritma berbasis jarak seperti SOM dan K-Means [32]. Variabel ini dikonversi menggunakan o*ne-hot encoding* yang efisien dan sering digunakan [33]. *Encoding* dilakukan pada fitur *Gender* dimana *Male* akan menjadi 1, dan *Female* menjadi 0. Sementara variabel numerik dengan rentang yang sangat berbeda dapat menyebabkan bias terhadap variabel tertentu pada saat pembentukan klaster [34]. Sehingga *scaling* diperlukan untuk menjadikan nilai agar seragam diantara 0-1 [34]. Penelitian ini menerapkan *MinMax Scaling*, dan data telah siap untuk digunakan pada model.

2.4 Hybrid SOM+K-Means

SOM merupakan salah satu metode *unsupervised neural network* yang dimanfaatkan untuk memetakan data berdimensi tinggi ke dalam ruang dua dimensi, sambil mempertahankan hubungan topologis antar data [35]. SOM digunakan untuk *clustering* dan segmentasi pelanggan terutama ketika data memiliki berbagai fitur karena kemampuannya dalam mereduksi dimensi tanpa kehilangan karakteristik penting dan mempertahankan kemiripan antar objek berdasarkan jarak di ruang fitur [36]. SOM banyak dikombinasikan dengan metode pembelajaran mesin lainnya seperti *k-clustering* dan *neural network*, yang memungkinkan interpretasi data berdimensi tinggi dan memberikan visualisasi yang cepat dan interpretatif seperti menggunakan *u-matrix map*, *clustering map* dan *heatmap* [37]. Dengan demikian, penggunaan SOM dapat memberikan profil bisnis dari segmentasi pelanggan yang lebih tevisualisasi dan mudah diinterpretasi. Proses pelatihan SOM meliputi inisialisasi bobot, *memilih Best Matching Unit* (BMU) untuk tiap input, lalu memperbarui bobot BMU dan tetangganya berdasarkan radius *neighborhood* dan *learning rate* yang menurun seiring waktu untuk memastikan konvergensi [38].

Setelah mendapatkan fitur SOM, K-Means dihitung pada node SOM. K-Means adalah algoritma *clustering* partisional yang digunakan karena kesederhanaan dan efisiensi komputasinya dalam mengelompokkan data berdimensi tinggi menjadi *k cluster* berdasarkan kedekatan ke *centroid* [39] K-Means tetap menjadi salah satu algoritma clustering yang banyak dipakai dalam analisis segmentasi pelanggan, baik di sektor ritel [40], hingga *e-commerce* [41], untuk membagi pelanggan ke dalam segmen berdasarkan karakteristik demografi, perilaku, atau transaksi. Sebuah penelitian merekomendasikan penggunaan K-Means sebagai bagian dari *pipeline hybrid* seperti mengabungkan dengan metode reduksi dimensi, metode validasi *cluster*, atau algoritma *clustering* alternatif untuk meningkatkan kualitas segmentasi terutama saat data kompleks [42]. Dalam penerapan K-Means, pemilihan jumlah cluster k dan inisialisasi centroid menunjukkan pengaruh besar terhadap kualitas *cluster*, jika *k* terlalu kecil *cluster* dapat terlalu general, dan jika terlalu besar *cluster* bisa *overfitting* dan tidak berarti secara bisnis [39]. Sehingga di penelitian ini *k* yang dipakai adalah 3, untuk memetakan churn pelanggan yang "berisiko tinggi", "berisiko sedang", dan "berisiko rendah" agar tetap berarti secara bisnis.

2.5 Cluster Profiling

Cluster profiling dilakukan setelah pembentukan *cluster* oleh *hybrid clustering* SOM+K-Means. Pada tahap ini dilakukan visualisasi pada data yang didapatkan, dan dilakukan analisis terhadap karakteristik dari setiap *cluster* yang terbentuk. *Cluster profiling* memberikan wawasan yang lebih mendalam tentang perbedaan antara segmen-segmen pelanggan, serta memungkinkan perusahaan untuk merancang strategi retensi yang lebih tepat sasaran berdasarkan karakteristik unik dari setiap segmen [43]. Tujuan utamanya adalah untuk mengeksplorasi dan menggambarkan setiap klaster pelanggan berdasarkan atribut-atribut penting yang relevan sehingga penanganan pelanggan yang churn tidak hanya menggunakan prediksi biner tapi dapat berdasarkan dari karakteristik masing-masing *cluster* pelanggan.

2.6 Persiapan Klasifikasi

Setelah tahapan *clustering* dan *profiling cluster* selesai, langkah berikutnya adalah mempersiapkan data untuk proses klasifikasi churn. Tahap ini mencakup penugasan label klaster yang telah didapat pada tiap pelanggan (*cluster assign*), pembagian data ke dalam 70% subset pelatihan dan 30% data pengujian agar memastikan bahwa model bisa dilatih dengan data yang representatif dan diuji dengan data yang sebelumnya belum pernah dilihat, serta penanganan masalah ketidakseimbangan kelas yang sering terjadi pada data churn. Penanganan terhadap ketidakseimbangan jumlah kelas mayoritas dan minoritas sangat krusial agar model klasifikasi tidak bias ke kelas mayoritas [44]. Untuk itu penelitian ini menggunakan SMOTE pada data latih. SMOTE menghasilkan sampel sintetis untuk kelas minoritas dengan interpolasi antara contoh-contoh minoritas yang ada [45].

2.7 Model Klasifikasi

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Literatur menunjukkan bahwa metode ML telah menjadi salah satu alat utama dalam memprediksi churn pelanggan di berbagai industri [46]. Metode prediksi churn tradisional seperti sistem berbasis aturan dan pemodelan statistik seringkali gagal menangkap kompleksitas perilaku pelanggan secara memadai. Sebaliknya, pendekatan ML seperti SVM, RF, dan XGB telah menunjukkan kemampuan prediktif yang kuat dengan dataset terstruktur [46].

SVM merupakan algoritma klasifikasi populer karena kemampuan untuk menangani data dengan distribusi yang tidak jelas dan memisahkan kelas melalui *hyperplane* optimal, serta secara efektif menangani data non-linier via *kernel trick* [47]. *Kernel trick* memetakan data ke dalam ruang yang berdimensi lebih tinggi dan kemudian menemukan hyperplane linier di ruang baru tersebut, ini yang membuat SVM efektif untuk data non linier [47]. SVM banyak digunakan di banyak domain karena keandalan, fleksibilitas kernel, dan kemampuannya dalam generalisasi bahkan pada dataset berdimensi tinggi [48].

Selanjutnya metode RF adalah metode *ensemble* berbasis *bagging* yang menggabungkan banyak pohon keputusan independen untuk menghasilkan klasifikasi atau prediksi kolektif yang dapat membantu mengurangi *overfitting* yang sering terjadi pada *single decision tree*, serta meningkatkan stabilitas dan generalisasi model [49]. Dalam konteks klasifikasi, penelitian menunjukkan RF sering mengungguli metode tunggal (seperti *decision tree*) dalam hal akurasi, *F1-score*, *recall/precision* sehingga dapat menjadi pilihan yang andal [50], [51].

XGB merupakan algoritma boosting yang secara bertahap membangun model untuk memperbaiki kesalahan model sebelumnya sehingga mampu menangkap pola kompleks dan interaksi antar fitur lebih baik daripada model tunggal [52]. Dalam studi komparatif terbaru, XGB bersama RF dievaluasi pada berbagai tingkat *imbalance dataset* dan hasilnya menunjukkan performa tinggi dan stabil meskipun kelas minoritas sedikit, menunjukkan efektivitas XGB dalam menangani data churn [44]. Karena sifat *boosting*-nya, XGB bisa mencapai konvergensi dengan cepat dan membantu mencegah *overfitting* fitur penting [52].

Penelitian ini juga akan menerapkan *hyperparameter tuning* yang berguna untuk peningkatan performa model *machine learning* dengan cara menentukan nilai-nilai dari *hyperparameter* [53]. *GridSearchCV* digunakan untuk menguji model secara *bruteforce* menggunakan nilai-nilai *hyperparameter* yang telah ditentukan, metode ini telah disediakan pada *Sckit-learn* [53].

2.8 Evaluasi Model

Metode evaluasi model klasifikasi yang digunakan pada penelitian ini adalah perhitungan *Accuracy, Precision, Recall*, *F1-Score,* dan *Area Under ROC Curve* (AUC*)* yang merupakan metrik evaluasi klasifikasi biner yang sering dipakai [54]. Menurut penelitian Sathyanarayanan & Tantri [54], akurasi adalah persentase dari total prediksi yang benar dibandingkan dengan seluruh data yang ada; presisi mengukur proporsi prediksi positif yang benar dibandingkan dengan seluruh prediksi positif; *recall* mengukur proporsi dari seluruh kasus positif yang berhasil diprediksi dengan benar; dan *f1-score* adalah rata-rata harmonis dari presisi dan *recall* yang memberikan gambaran umum tentang keseimbangan antara keduanya; serta AUC adalah indikasi bahwa model dapat membedakan antara dua kelas dengan baik. AUC dapat menggunakan fungsi *roc_auc_score(y_true, y_scores)* yang akan menghitung AUC dari data probabilitas prediksi dengan label asli. Sedangkan berikut adalah rumus dari akurasi (1), presisi (2), *recall* (3), dan *F1-Score* (4) :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-score} = 2.\frac{Precision\ .Recall}{Precision+\ Recall} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Penerapan Hybrid Clustering SOM+K-Means

Penerapan SOM dilakukan dalam dua langkah yaitu inisialisasi parameter dan proses pelatihan. Dalam menentukan parameter yang dipakai dilakukan iterasi SOM+K-Means untuk mencari kombinasi parameter terbaik berbasis *silhouette score.* Dengan menggunakan *k* = 3, *random_seed* = 42, dan *neighborhood_function* = *gaussian* ditemukan kombinasi parameter SOM terbaik dari proses tersebut yaitu, *grid_size* = 8 x 8, *sigma* = 1.0, dan *learning rate* = 0.2. Pelatihan dilakukan dengan skema *train_random* selama 5.000 iterasi. Hasil pencarian parameter terbaik dapat dilihat di Tabel 4 berikut yang merupakan 5 kombinasi parameter terbaik dengan masing-masing parameter dan nilai *silhouette*-nya.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

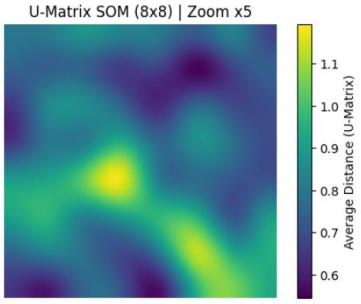
Tabel 3. Hasil Pencarian Parameter

Grid Size	*Sigma*	*Learning Rate*	*Silhouette*
8	1.0	0.2	0.202757
15	0.5	0.5	0.202747
8	0.5	0.5	0.202686
12	0.5	0.3	0.202637
10	1.5	0.5	0.202627

Dari parameter yang didapat, dilakukan pelatihan pada SOM. SOM belajar mengenali pola-pola data berdasarkan hubungan antar fitur pelanggan dengan *neighborhood_function* = *gaussian*. Setelah proses pelatihan SOM selesai, bobot node SOM yang dihasilkan disimpan dalam matriks yang memiliki bentuk (8,8,*n_features*). Matriks ini merepresentasikan bobot masing-masing node dalam *grid* SOM. Lalu diratakan menjadi dua dimensi untuk mempermudah analisis dengan ukuran (64, *n_features*), yang menunjukkan setiap node dalam *grid* 8x8 sebagai vektor fitur yang dapat digunakan untuk klasterisasi selanjutnya. Ini dapat membuat gambaran yang lebih mudah untuk diinterpretasikan mengenai struktur data yang telah dipetakan oleh SOM. Selanjutnya, dilakukan klasterisasi K-Means pada node SOM untuk mengelompokkan node menjadi beberapa klaster. Data asli dipetakan ke dalam klaster-klaster ini dengan mencari *Best Matching Unit* (BMU) untuk setiap data pelanggan, dan hasilnya digunakan untuk menentukan label klaster akhir bagi setiap pelanggan. Label klaster yang terbentuk kemudian digunakan untuk menganalisis pola churn dan merancang strategi retensi yang lebih terfokus pada setiap segmen pelanggan. Pendekatan ini berhasil mengidentifikasi segmen pelanggan berdasarkan pola perilaku, memungkinkan perusahaan untuk lebih memahami karakteristik masing-masing kelompok pelanggan dan memprediksi churn dengan lebih akurat.

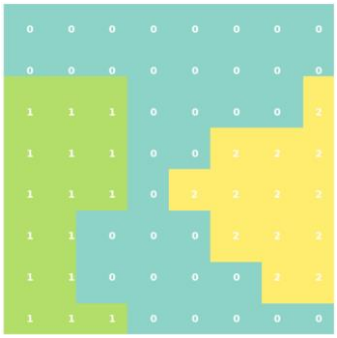
3.2 Hasil Cluster Profiling

Dari hasil pelatihan yang dilakukan, dapat dilihat hubungan antar node pada SOM yang divisualisasikan pada U-Matrix berikut pada Gambar 2. U-Matrix menunjukkan jarak antar node pada peta SOM, di mana jarak antar node menggambarkan perbedaan antar klaster. Warna pada U-Matrix menunjukkan tingkat jarak dari jarak tersebut. Di bagian warna kuning (terang) menunjukkan jarak yang lebih besar antara node-node yang berdekatan. Ini menandakan adanya batas yang jelas antara klaster pada peta SOM, yaitu perbedaan yang signifikan antara grup data yang satu dengan yang lain. Warna biru atau ungu (gelap) menunjukkan jarak yang lebih kecil, yang berarti node-node tersebut memiliki kesamaan tinggi yang mengindikasikan bahwa area tersebut lebih homogen dalam hal fitur, yang bisa berarti data dalam area tersebut sangat mirip atau berada dalam klaster yang sama.



Gambar 2. *U-Matrix*

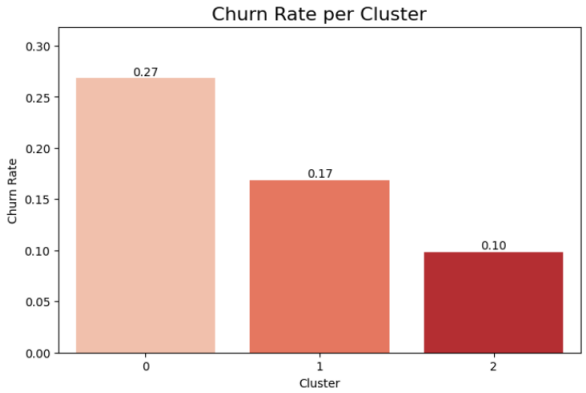
Selanjutnya dapat dilihat di Gambar 3 ini adalah *Cluster Map* K-Means pada SOM, di mana setiap angka mewakili label klaster yang dihasilkan oleh algoritma K-Means setelah melakukan clustering pada SOM. Berdasarkan gambar, terlihat bahwa klaster 0 memiliki bentuk yang paling besar, dan klaster 1 memiliki bentuk yang cukup besar serta terpisah dengan jelas dari klaster 2 yang terletak di sisi kanan. Hal ini menunjukkan bahwa K-Means telah berhasil mengidentifikasi bahwa dua segmen atau kelompok data memiliki fitur yang sangat berbeda untuk *clustering*.



Gambar 3. *Cluster Map*

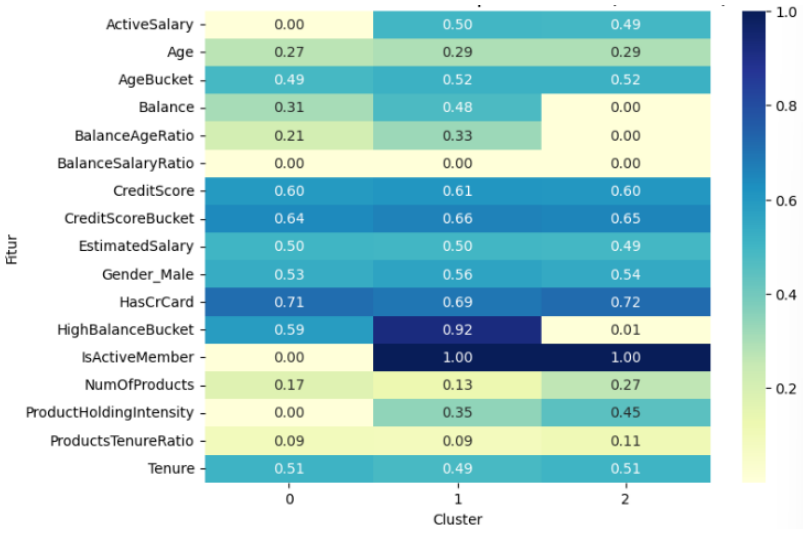
JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Didapatkan bahwa terdapat 3 klaster yang terbentuk dengan klaster 0 sebanyak 4850, klaster 1 sebanyak 3239, dan klaster 2 sebanyak 1911 pelanggan. Untuk memahami lebih lanjut mengenai distribusi churn pada setiap klaster, analisis dilakukan dengan mengaitkan status churn dengan setiap klaster yang terbentuk, sehingga mendapatkan hasil seperti pada Gambar 4 dibawah. Klaster 0 memiliki 27% pelanggan churn, klaster 1 memiliki 17% pelanggan churn, dan klaster 2 memiliki 10% pelanggan yang churn. Sehingga klaster dapat dibentuk menjadi klaster dengan risiko churn tinggi, risiko churn sedang, dan risiko churn rendah.



Gambar 4. *Churn Rate Cluster*

Setelah menganalisis distribusi churn pada setiap klaster, terlihat bahwa klaster-klaster tertentu memiliki tingkat churn yang lebih tinggi dibandingkan yang lainnya, yang menandakan adanya perbedaan signifikan dalam perilaku pelanggan di setiap segmen. Klaster dengan *churn rate* lebih tinggi menunjukkan bahwa pelanggan dalam segmen tersebut lebih rentan untuk berhenti menggunakan layanan. Karena itu, agar dapat memahami faktor-faktor yang memengaruhi churn lebih mendalam, penting untuk mengeksplorasi kontribusi fitur-fitur yang membedakan setiap klaster. Dengan memeriksa fitur dominan yang ada dalam setiap segmen pelanggan, kita dapat mengidentifikasi atribut-atribut kunci yang berperan besar dalam menentukan apakah pelanggan berisiko churn atau tidak. Analisis ini tidak hanya memberikan wawasan tentang karakteristik setiap klaster, tetapi juga membantu merancang strategi yang lebih tepat sasaran untuk mengurangi churn, dengan menargetkan fitur-fitur yang paling berpengaruh dalam proses pengambilan keputusan pelanggan untuk berhenti berlangganan. Berikut pada Gambar 5 merupakan *heatmap* dari kontribusi fitur di setiap klaster.



Gambar 5. *Heatmap* Kontribusi Fitur

Berdasarkan heatmap tersebut karakteristik setiap klaster dapat dianalisis sebagai berikut:

a. Klaster 0 - Risiko Churn Tinggi

Klaster ini memiliki risiko churn yang lebih tinggi, dengan nilai yang paling mempengaruhi pada fitur seperti *ActiveSalary* (0.00), *IsActiveMember* (0.00), dan *ProductHoldingIntensity* (0.00) yang merupakan nilai paling rendah untuk fitur tersebut dari setiap klaster. Ini menunjukkan bahwa klaster ini memiliki pelanggan yang paling tidak aktif, tidak terlibat dengan aktivitas gaji, serta tidak menggunakan produk dan layanan yang ditawarkan. Padahal *NumOfProducts* (0.17), dimana pelanggan memiliki sedikit produk atau layanan yang sedikit lebih banyak dari klaster 1. Fitur-fitur ini memberikan indikasi kuat bahwa mereka berisiko untuk churn. Nilai *AgeBucket* (0.49) menandakan bahwa pelanggan klaster ini relatif lebih muda dari klaster lainnya. *Balance* (0.31) dan *BalanceAgeRatio* (0.21), menunjukkan

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

saldo rendah, dengan rasio saldo terhadap usia yang juga rendah, menunjukkan kurangnya kesejahteraan finansial dan kestabilan, namun berada di tengah klaster lainnya. Dengan nilai *HighBalanceBucket* (0.51) rata-rata pelanggan termasuk pada kelompok dengan saldo menengah. *HasCrCard* (0.71) banyak pelanggan yang memiliki kartu kredit, dan memiliki *CreditScore* (0.60) terletak di *CreditScoreBucket* (0.64) menunjukkan bahwa skor kredit sedang namun merupakan yang paling rendah diantara klaster lain. Nilai *Tenure* (0.51) yang berarti masa langganan cukup lama, tetapi ini tidak cukup untuk membuat mereka terikat dengan layanan, karena dapat dilihat dari *ProductTenureRatio* (0.09), pelanggan cenderung memiliki masa langganan yang pendek untuk setiap produknya. Sedangkan *Gender_Male* (0.53) menunjukkan bahwa sebagian lebih pelanggan merupakan laki-laki namun hal ini tidak berbeda jauh dengan klaster-klaster lain.

Rekomendasi retensi untuk klaster 0 yang memiliki karakteristik pelanggan yang sangat tidak aktif, tidak menggunakan produk secara intensif, memiliki saldo cukup moderat, serta cenderung memiliki risiko churn tinggi karena minimnya keterlibatan dan stabilitas finansial. Perusahaan perlu menawarkan insentif khusus untuk mendorong pelanggan ini lebih terlibat, seperti diskon atau bonus untuk penggunaan produk lebih banyak, serta program loyalitas untuk memperbaiki hubungan dan meningkatkan penggunaan layanan. Menyediakan pemberitahuan khusus tentang manfaat produk yang dapat mendorong lebih banyak penggunaan layanan, pertimbangkan juga untuk menyediakan layanan edukasi finansial untuk meningkatkan pemahaman dalam mengelola keuangan.

b. Klaster 1 - Risiko Churn Sedang

Klaster ini memiliki nilai *HighBalanceBucket* yang sangat tinggi (0.92), dan nilai *IsActiveMember* yang sangat tinggi pula (1.00), mengindikasikan bahwa pelanggan di klaster ini termasuk pada kelompok yang memiliki saldo tinggi, dan aktif menggunakan layanan secara optimal. Nilai *Balance* (0.48) dan *BalanceAgeRatio* (0.33) menunjukkan tingkat kesejahteraan finansial yang lebih tinggi dibandingkan klaster 0 dan menunjukkan stabilitas finansial per usia yang relatif baik. *EstimatedSalary* dan *ActiveSalary* (0.50) menandakan penghasilan yang moderat dan adanya aktivitas dalam pengelolaan keuangan. *CreditScore* (0.61) dan *CreditScoreBucket* (0.66) yang lebih tinggi menunjukkan kebiasaan yang baik dalam mengelola utang. Nilai *Age* (0.29) dan *AgeBucket* (0.52) menunjukkan pelanggan di kelompok usia produktif dan sebagian lebih dari pelanggan merupakan laki-laki dengan *Gender_Male* (0.56). Untuk nilai *HasCrCard* (0.69) sebagian besar pelanggan memiliki kartu kredit, namun nilainya lebih kecil diantara klaster lain. *ProductHoldingIntensity* (0.35) memiliki keterlibatan pada produk yang ditawarkan namun berada di tengah antara nilai klaster lainnya dan nilai *NumOfProducts* (0.13) yang memiliki sedikit produk yang digunakan. Dan faktor yang memiliki nilai rendah lainnya dibanding klaster lain adalah *ProductTenureRatio* (0.09) dimana masa langganan produk lebih pendek, dan *Tenure* (0.49) yang berarti pelanggan memiliki masa langganan yang cukup stabil namun lebih pendek dibanding klaster lain sehingga memungkinkan untuk risiko churn.

Rekomendasi retensi untuk klaster 1 yang memiliki karakteristik pelanggan yang memiliki saldo tinggi dan tingkat aktivitas tinggi menunjukkan kesejahteraan finansial yang lebih baik dan keterlibatan yang lebih tinggi. Klaster ini juga memiliki skor kredit baik dan pengelolaan keuangan yang lebih baik. Namun, mereka memiliki sedikit produk yang digunakan dan masa langganan produk yang pendek. Dengan itu perlu peningkatan keterlibatan produk, misalnya dengan menawarkan *bundle* produk atau *cross-selling* untuk mendorong penggunaan produk lebih banyak, serta program loyalitas dengan penghargaan jangka waktu untuk mendorong penggunaan layanan jangka panjang.

c. Klaster 2 - Risiko Churn Rendah

Klaster ini memiliki nilai *IsActiveMember* yang sangat tinggi (1.00), yang mengindikasikan bahwa pelanggan di klaster ini sangat aktif dan terlibat dalam program perusahaan. *HasCrCard* (0.72) juga memiliki nilai tinggi sebagian besar memiliki kartu kredit yang dapat diartikan memiliki keterlibatan lebih besar dengan produk finansial. Nilai *NumOfProducts* (0.27) dan *ProductHoldingIntensity* (0.45) dimana pelanggan lebih banyak menggunakan produk daripada klaster lain. Serta dengan nilai *ProductTenureRatio* (0.11) dan *Tenure* (0.51) yang tinggi daripada klaster lain menunjukkan pelanggan memiliki masa langganan yang lebih panjang dan hubungan yang lebih kuat pada layanan. Nilai *Age* (0.29) dan *AgeBucket* (0.52) menunjukkan pelanggan di kelompok usia produktif dan sebagian lebih dari pelanggan merupakan laki-laki dengan *Gender_Male* (0.54). *CreditScoreBucket* (0.65) dan *CreditScore* (0.60) adalah nilai yang moderat menunjukkan stabilitas finansial yang baik walaupun masih dibawah klaster lain. *ActiveSalary* (0.49) dengan *EstimatedSalary* (0.49) menandakan keterlibatan moderat dalam aktivitas penghasilan. Nilai *Balance* (0.00), *BalanceAgeRatio* (0.00), dan *HighBalanceBucket* (0.01) menandakan tidak ada pelanggan dengan saldo tinggi, ini menunjukkan bahwa pelanggan klaster ini lebih terlibat dalam layanan meskipun tidak memiliki saldo tinggi.

Rekomendasi retensi untuk klaster 1 yang memiliki karakteristik pelanggan yang sangat aktif dan memiliki aktivitas tinggi pada penggunaan produk dan layanan. Sebagian besar pelanggan memiliki kartu kredit dan skor kredit yang cukup baik. Dengan masa langganan yang sedikit lebih lama daripada klaster lain. Meskipun risiko churn relatif rendah, perusahaan tetap perlu memperkuat loyalitas dengan mendorong penggunaan produk lebih banyak dan meningkatkan keterlibatan melalui penawaran eksklusif dan akses prioritas ke produk premium untuk pelanggan yang telah aktif menggunakan layanan. *Tenure* di klaster ini cukup stabil meskipun ada ruang untuk meningkatkan keterlibatan.

3.3 Hasil Model Klasifikasi

Dari hasil segmentasi yang telah terbentuk, label klaster yang dihasilkan dimasukkan sebagai fitur tambahan untuk proses klasifikasi, sehingga model dapat mempertimbangkan informasi terkait segmen pelanggan dalam memprediksi churn. Kemudian, membagi data menjadi 70% data latih dan 30% data uji. Untuk mengatasi ketidakseimbangan kelas yang terjadi pada data churn telah diterapkan teknik SMOTE pada data latih. Tabel 5 menunjukkan jumlah data sebelum dan

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9409
Hal 335-346
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

setelah penerapan SMOTE, yang menggambarkan perubahan jumlah sampel pada masing-masing kelas setelah teknik ini diterapkan. Kedua kelas yaitu kelas churn dan non-churn menjadi seimbang sebanyak 5574 data.

Tabel 4. Jumlah Data SMOTE

Jumlah Data	Sebelum SMOTE	Setelah SMOTE
0	5574	5574
1	1426	5574

Setelah didapatkan hasil dataset yang seimbang, penelitian ini telah menguji tiga algoritma klasifikasi ML untuk memprediksi churn pelanggan yaitu dengan SVM, RF, dan XGB menggunakan *GridSearchCv* untuk mendapatkan parameter terbaik. Dimana hasil terbaik masing-masing model didapatkan dari parameter-parameter berikut ini; SVM (*C*=10*, class_weight='balanced', gamma='scale', kernel='rbf' ,probability=True*), RF (*max_depth=None, min_samples_leaf*=1*, min_samples_split*=2*, n_estimators*=400), dan XGB (*colsample_bytree= 0.9, learning_rate*=0.1*, max_depth*=8*, n_estimators*=400*, subsample*= 0.8). Hasil dari setiap model dievaluasi menggunakan metrik *accuracy, precision, recall, F1-score,* dan *AUC* pada Tabel 6 di bawah.

Tabel 5. Perbandingan Model

Model	Accuracy	Precision	Recall	F1-Score	AUC
SVM	0.77	0.45	**0.72**	0.55	0.83
RF	0.83	0.58	0.60	0.59	0.85
XGB	**0.85**	**0.66**	0.54	**0.59**	**0.85**

Dari hasil tersebut dapat dilihat bahwa model SVM memberikan akurasi sebesar 77% yang menunjukkan bahwa model ini dapat melakukan klasifikasi dengan cukup baik, namun model RF terlebih lagi model XGB memberikan akurasi yang lebih tinggi sebesar 83% dan 85%. Pada metrik *precision* menunjukkan bahwa model SVM memiliki nilai 45% yang cenderung mengklasifikasikan banyak pelanggan non-churn sebagai churn. Sedangkan *precision* yang lebih baik dan tinggi ada pada model XGB sebesar 66%, lebih daripada model RF dengan 58%. Meski demikian, model SVM memiliki *recall* yang cukup tinggi dengan 72%, yang berarti model berhasil mendeteksi sebagian besar pelanggan churn sebenarnya meskipun masih ada yang terlewat. *Precision* pada RF dan XGB masih lebih rendah dari SVM dengan nilai 60% dan 54%. Untuk nilai F1-Score tertinggi ada pada model XGB dan RF dengan nilai 59%, begitu juga untuk nilai AUC dengan nilai 85%. Sedangkan SVM dengan nilai F1-Score dan AUC nya adalah 55% dan 83%.

3.4 Pembahasan

Penerapan *Hybrid Clustering* menggunakan SOM + K-Means memberikan wawasan yang lebih mendalam tentang karakteristik dari masing-masing segmen pelanggan, yang penting untuk penanganan churn dengan lebih akurat. SOM berfungsi untuk memetakan data yang berdimensi tinggi ke dalam grid dua dimensi yang lebih mudah dipahami, sementara K-Means digunakan untuk mengelompokkan node-node SOM menjadi klaster-klaster yang lebih spesifik. Model SOM+K-Means ini menghasilkan 3 klaster dengan *silhouette score* sebesar 0.2027. Nilai *silhouette* pada penelitian ini lebih besar dibandingkan dengan *silhouette* dari penelitian yang menggunakan model *K-Prototype* untuk segmentasi churn bank sebesar 0.1557 [55]. Dari hasil *profiling cluster* terlihat bahwa segmentasi ini memungkinkan identifikasi karakteristik pelanggan yang lebih spesifik, seperti tingkat keaktifan, pola penggunaan produk, dan saldo yang mempengaruhi churn. Dengan SOM + K-Means, dapat diidentifikasi pola yang lebih jelas antara segmen-segmen pelanggan yang berisiko churn tinggi, risiko churn sedang, dan risiko churn rendah yang tidak bisa diperoleh hanya dengan klasifikasi biner biasa. Sehingga dapat digunakan untuk merancang strategi retensi berbasis data dan lebih mudah diinterpretasikan. Dalam hal ini klaster dengan risiko churn tinggi dapat difokuskan pada program penawaran dan insentif khusus serta layanan edukasi finansial untuk meningkatkan keaktifan dan keterlibatan pada produk serta meningkatkan kemampuan pengelolaan keuangan. Klaster dengan risiko churn sedang difokuskan pada program keterlibatan produk lebih banyak dan penghargaan jangka waktu untuk penggunaan layanan jangka panjang. Serta klaster dengan risiko churn rendah difokuskan pada program penawaran eksklusif dan loyalitas agar pelanggan tetap aktif dan pemakaian layanan bahkan dapat lebih meningkat lagi.

Setelah model segmentasi berhasil diterapkan, model klasifikasi kemudian dipakai untuk memprediksi churn. Hasil evaluasi dari model klasifikasi menunjukkan bahwa XGB memberikan kinerja terbaik dengan akurasi 85% dan AUC yang tinggi sebesar 85%, diikuti oleh RF dan SVM. Meskipun XGB memiliki *precision* terbaik untuk churn 66%, *recall* 54% masih lebih rendah dibandingkan model lain, menunjukkan bahwa model ini masih belum dapat mendeteksi sejumlah pelanggan churn dengan sempurna. Model SVM meskipun memberikan *recall* tertinggi untuk churn 72%, memiliki *precision* yang sangat rendah yaitu 45%, yang artinya banyak pelanggan non-churn yang salah terklasifikasi sebagai churn. Hal ini menunjukkan bahwa SVM sangat sensitif terhadap churn, namun tidak cukup akurat dalam memprediksi pelanggan lain. Secara keseluruhan, XGBoost lebih unggul dalam memprediksi churn daripada model lain meskipun *recall* lebih rendah. XGBoost mampu memberikan keseimbangan yang lebih baik antara akurasi, AUC, *precision* dan *F*1-*score* dibandingkan dengan RF dan SVM. Pada penelitian sebelumnya, beberapa metode seperti *Random Forest* dan *Decision Tree* digunakan untuk memprediksi churn bank, dengan hasil akurasi yang lebih rendah dibandingkan penelitian ini, yaitu sebesar 78% dan 72% [27]. Meskipun *Random Forest* sering digunakan dalam konteks

prediksi churn, hasil yang diperoleh dalam studi tersebut menunjukkan bahwa model ini kurang optimal dalam menangani dataset churn. Di sisi lain, penelitian lain yang menggunakan XGBoost untuk prediksi churn bank berhasil mencapai akurasi 0.839 dan AUC 0.847 [56], yang menunjukkan performa yang cukup baik meskipun tidak mencapai nilai lebih tinggi dari yang diperoleh dalam penelitian ini. Hasil ini menunjukkan bahwa meskipun XGBoost sering digunakan dengan baik dalam pengklasifikasian churn, masih terdapat beberapa faktor yang mempengaruhi akurasi model dalam memprediksi churn pelanggan.

4. KESIMPULAN

Penelitian ini mengembangkan pendekatan untuk menganalisis pola churn dan memprediksi churn pelanggan dengan menggabungkan segmentasi pelanggan menggunakan teknik *hybrid* SOM+K-Means dan klasifikasi menggunakan beberapa algoritma pembelajaran mesin, yaitu *Random Forest*, *XGBoost*, dan *Support Vector Machine*. SOM+K-Means menjadikan klaster lebih mudah dibagi dan diinterpretasikan. Penerapan model segmentasi ini bertujuan agar penanganan churn tidak hanya dilakukan dengan prediksi biner, tetapi juga disesuaikan dengan karakteristik setiap segmen pelanggan agar perusahaan dapat memberikan strategi retensi yang lebih akurat dan mendalam. Dari hasil segmentasi ditemukan bahwa terdapat tiga klaster pelanggan dengan risiko churn yang berbeda-beda, yaitu klaster dengan risiko churn tinggi, sedang, dan rendah. Setiap klaster memiliki karakteristik yang unik, seperti tingkat aktivitas, saldo, dan penggunaan produk yang mempengaruhi potensi churn. Selain itu, dalam tahap klasifikasi, penelitian ini menggunakan SMOTE untuk menangani masalah kelas tidak seimbang pada data churn, yang terbukti efektif dalam meningkatkan representasi kelas churn dalam data latih. Dilakukan pula *hyperparameter tuning* dengan optimasi dari *GridSearchCV* sehingga diperoleh parameter yang memberikan hasil terbaik untuk setiap model. Meskipun hasil akurasi dan AUC model sudah tinggi, masih terdapat ruang untuk peningkatan terutama pada metrik *recall*. Keterbatasan dari penelitian ini adalah meskipun *feature engineering* dan *clustering* memberikan insight yang berguna, hasil prediksi churn masih dipengaruhi oleh distribusi data. Selain itu, penelitian ini masih terbatas pada penggunaan dataset tunggal dan belum mempertimbangkan variabel eksternal yang mungkin juga mempengaruhi churn pelanggan, seperti aspek layanan pelanggan, kepuasan pelanggan, dan faktor ekonomi makro. Untuk penelitian selanjutnya, disarankan agar bisa memperluas analisis menggunakan lebih banyak data dan mempertimbangkan penggunaan teknik lainnya yang dapat mengekspolarasi churn lebih jauh.

REFERENCES

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” J Big Data, vol. 6, no. 1, p. 28, 2019, doi: 10.1186/s40537-019-0191-6.
[2] Y. Ortakci and H. Seker, “Optimising customer retention: An AI-driven personalised pricing approach,” Comput Ind Eng, vol. 188, p. 109920, 2024, doi: https://doi.org/10.1016/j.cie.2024.109920.
[3] S. S. Poudel, S. Pokharel, and M. Timilsina, “Explaining customer churn prediction in telecom industry using tabular machine learning models,” Machine Learning with Applications, vol. 17, p. 100567, 2024, doi: https://doi.org/10.1016/j.mlwa.2024.100567.
[4] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business,” IEEE Access, vol. 9, pp. 62118–62136, 2021, doi: 10.1109/ACCESS.2021.3073776.
[5] A. A. Jamjoom, “The use of knowledge extraction in predicting customer churn in B2B,” J Big Data, vol. 8, no. 1, p. 110, 2021, doi: 10.1186/s40537-021-00500-3.
[6] S. Sundram, D. D. Poornima, D. Praveenkumar, M. C. Balakumar, D. D. Sasikala, and S. Omonov, “A Novel Stochastic Gradient Descent Based Logistic Regression (SGD-LR) Framework for Customer Churn Prediction,” International Journal of Intelligent Systems and Applications in Engineering IJISAE, vol. 12, no. 17s, pp. 754–765, Feb. 2024.
[7] T. Zhang, S. Moro, and R. F. Ramos, “A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation,” Future Internet, vol. 14, no. 3, 2022, doi: 10.3390/fi14030094.
[8] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, and T. Verdonck, “Profit driven decision trees for churn prediction,” Eur J Oper Res, vol. 284, no. 3, pp. 920–933, 2020, doi: https://doi.org/10.1016/j.ejor.2018.11.072.
[9] M. Kiguchi, W. Saeed, and I. Medi, “Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest,” Appl Soft Comput, vol. 118, p. 108491, 2022, doi: https://doi.org/10.1016/j.asoc.2022.108491.
[10] C. Wang, D. Han, W. Fan, and Q. Liu, “Customer Churn Prediction with Feature Embedded Convolutional Neural Network: An Empirical Study in the Internet Funds Industry,” Int J Comput Intell Appl, vol. 18, no. 01, p. 1950003, Mar. 2019, doi: 10.1142/S1469026819500032.
[11] A. Hermawan, W. Wijaya, and B. Daniawan, “Optimizing Artificial Neural Network for Customer Churn: Advanced Data Balancing and Feature Selection,” International Journal on Informatics Visualization, vol. 9, no. 3, pp. 1132–1141, May 2025, doi: 10.62527/joiv.9.3.3064.
[12] Seema and G. Gupta, “Development of fading channel patch based convolutional neural network models for customer churn prediction,” International Journal of System Assurance Engineering and Management, vol. 15, no. 1, pp. 391–411, 2024, doi: 10.1007/s13198-022-01759-2.
[13] S. Ouf, K. T. Mahmoud, and M. A. Abdel-Fattah, “A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry,” J Big Data, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00922-9.
[14] F. E. Usman-Hamza et al., “Sampling-based novel heterogeneous multi-layer stacking ensemble method for telecom customer churn prediction,” Sci Afr, vol. 24, Jun. 2024, doi: 10.1016/j.sciaf.2024.e02223.

[15] M. Szeląg and R. Słowiński, "Explaining and predicting customer churn by monotonic rules induced from ordinal data," Eur J Oper Res, vol. 317, no. 2, pp. 414–424, 2024, doi: https://doi.org/10.1016/j.ejor.2023.09.028.
[16] M. Troiano et al., "A Comparative Analysis of Machine Learning Algorithms for Identifying Cultural and Technological Groups in Archaeological Datasets through Clustering Analysis of Homogeneous Data," Electronics (Basel), vol. 13, no. 14, 2024, doi: 10.3390/electronics13142752.
[17] L. Tan, C. Li, J. Xia, and J. Cao, "Application of Self-Organizing Feature Map Neural Network Based on K-means Clustering in Network Intrusion Detection," Computers, Materials & Continua, vol. 61, no. 1, pp. 275–288, 2019, doi: 10.32604/cmc.2019.03735.
[18] D. D. S, T. S, and L. G, "Performance Evaluation Of Clustering-A Hybrid Approach [K-Mean-SOM NN] In Health Care," Journal for Re Attach Therapy and Developmental Diversities, vol. 6, no. 1, pp. 862–869, 2023, doi: https://doi.org/10.53555/jrtdd.v6i1.2309.
[19] J. Zhu and X. Han, "Big data clustering algorithm of power system user load characteristics based on K-means and SOM neural network," Multimed Tools Appl, vol. 84, no. 10, pp. 7477–7491, 2025, doi: 10.1007/s11042-024-19156-1.
[20] Z. Zhu et al., "Seismic Facies Analysis Using the Multiattribute SOM-K-Means Clustering," Comput Intell Neurosci, vol. 2022, no. 1, p. 1688233, 2022, doi: https://doi.org/10.1155/2022/1688233.
[21] M. Kotyrba, E. Volna, R. Jarusek, and P. Smolka, "The use of conventional clustering methods combined with SOM to increase the efficiency," Neural Comput Appl, vol. 33, no. 23, pp. 16519–16531, 2021, doi: 10.1007/s00521-021-06251-9.
[22] A. Saeed, Dr. A. A. Sanjrani, S. K. S. Bukhari, and S. Ahmad, "Improving The Accuracy Of Imbalanced Dataset Using KMeans Clustering," Spectrum of Engineering Sciences, vol. 3, no. 9, pp. 106–116, Sep. 2025.
[23] J.-X. Ong, G.-K. Tong, K.-C. Khor, and S.-C. Haw, "Enhancing Customer Churn Prediction With Resampling: A Comparative Study," TEM Journal, pp. 1927–1936, Aug. 2024, doi: 10.18421/TEM133-20.
[24] R. Suguna, J. Suriya Prakash, H. Aditya Pai, T. R. Mahesh, V. Vinoth Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," Sci Rep, vol. 15, no. 1, p. 16256, May 2025, doi: 10.1038/s41598-025-01031-0.
[25] N. N. Y., T. Van Ly, and D. V. T. Son, "Churn prediction in telecommunication industry using kernel Support Vector Machines," PLoS One, vol. 17, no. 5, May 2022, doi: 10.1371/journal.pone.0267935.
[26] A. Velia, T. R. Simamora, S. N. Suherman, A. A. Pravitasari, and F. Indrayatna, "Klasifikasi Customer Churn Pada Perusahaan Telekomunikasi Menggunakan Support Vector Machine," BIAStatistics Journal of Statistics Theory and Aplication, vol. 2022, no. 1, 2022.
[27] A. F. Azmi and A. Voutama, "Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Random Forest Dan Decision Tree Dengan Evaluasi Confusion Matrix," Komputa : Jurnal Ilmiah Komputer dan Informatika, vol. 13, no. 1, pp. 111–119, Apr. 2024, doi: 10.34010/komputa.v13i1.12639.
[28] L. N. Wakhidah, A. K. Zyen, and B. B. Wahono, "Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms," Journal of Applied Informatics and Computing, vol. 9, no. 1, pp. 89–95, Jan. 2025, doi: 10.30871/jaic.v9i1.8740.
[29] R. M. Daffaa, D. Santika, and F. Mahardika, "Perbandingan Xgboost dan Logistic Regression dalam Memprediksi Credit Card Customer Churn," Jupiter: Publikasi Ilmu Keteknikan Industri, Teknik Elektro dan Informatika, vol. 3, no. 3, pp. 07–21, Apr. 2025, doi: 10.61132/jupiter.v3i3.807.
[30] M. Z. Alotaibi and M. A. Haq, "Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods," Engineering, Technology & Applied Science Research, vol. 14, no. 3, pp. 14572–14578, Jun. 2024, doi: 10.48084/etasr.7480.
[31] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," Front Energy Res, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.
[32] A. Guérin, P. Chauvet, and F. Saubion, "A Survey on Recent Advances in Self-Organizing Maps," Dec. 2024, doi: https://doi.org/10.48550/arXiv.2501.08416.
[33] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," J Big Data, vol. 7, no. 1, p. 28, 2020, doi: 10.1186/s40537-020-00305-w.
[34] C. Wongoutong, "The impact of neglecting feature scaling in k-means clustering," PLoS One, vol. 19, no. 12, pp. 1–19, Oct. 2024, doi: 10.1371/journal.pone.0310839.
[35] M. Deetz, "K-Means Clustering of Self-Organizing Maps: An Empirical Study on the Information Content of Self-Classification of Hedge Fund Managers," INTERNATIONAL JOURNAL OF MANAGEMENT SCIENCE AND BUSINESS ADMINISTRATION, vol. 5, no. 3, pp. 43–57, 2019, doi: 10.18775/ijmsba.1849-5664-5419.2014.53.1006.
[36] J. Liao, A. Jantan, Y. Ruan, and C. Zhou, "Multi-Behavior RFM Model Based on Improved SOM Neural Network Algorithm for Customer Segmentation," IEEE Access, vol. 10, pp. 122501–122512, 2022, doi: 10.1109/ACCESS.2022.3223361.
[37] J. Qian et al., "Introducing self-organized maps (SOM) as a visualization tool for materials research and education," Results in Materials, vol. 4, p. 100020, 2019, doi: https://doi.org/10.1016/j.rinma.2019.100020.
[38] S. Licen, A. Astel, and S. Tsakovski, "Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review," Science of The Total Environment, vol. 878, p. 163084, 2023, doi: https://doi.org/10.1016/j.scitotenv.2023.163084.
[39] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Inf Sci (N Y), vol. 622, pp. 178–210, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.139.
[40] Y. Syahra, A. Fadlil, and H. Yuliansyah, "Customer Segmentation Using RFM and K-Means Clustering to Support CRM in Retail Industry," sinkron, vol. 9, no. 3, pp. 1120–1131, Jul. 2025, doi: 10.33395/sinkron.v9i3.14907.
[41] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," Sustainability, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/su14127243.
[42] F. H. Awad and M. M. Hamad, "Improved k-Means Clustering Algorithm for Big Data Based on Distributed SmartphoneNeural Engine Processor," Electronics (Basel), vol. 11, no. 6, p. 883, Mar. 2022, doi: 10.3390/electronics11060883.

[43] M. S. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," Neural Comput Appl, vol. 36, no. 9, pp. 4995–5005, Mar. 2024, doi: 10.1007/s00521-023-09339-6.
[44] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," Technologies (Basel), vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.
[45] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, "An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space," Sci Rep, vol. 15, no. 1, p. 23521, Jul. 2025, doi: 10.1038/s41598-025-09506-w.
[46] M. Imani, M. Joudaki, A. Beikmohammadi, and H. Arabnia, "Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning," Mach Learn Knowl Extr, vol. 7, no. 3, p. 105, Sep. 2025, doi: 10.3390/make7030105.
[47] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Information, vol. 15, no. 4, p. 235, Apr. 2024, doi: 10.3390/info15040235.
[48] A. Kar, N. Nath, U. Kemprai, and Aman, "Performance Analysis of Support Vector Machine (SVM) on Challenging Datasets for Forest Fire Detection," International Journal of Communications, Network and System Sciences, vol. 17, no. 02, pp. 11–29, 2024, doi: 10.4236/ijcns.2024.172002.
[49] E. Halabaku and E. Bytyçi, "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests," Intelligent Automation & Soft Computing, vol. 39, no. 6, pp. 987–1006, 2024, doi: 10.32604/iasc.2024.059429.
[50] A. N. S. Kinasih, A. N. Handayani, J. T. Ardiansah, and N. S. Damanhuri, "Comparative analysis of decision tree and random forest classifiers for structured data classification in machine learning," Science in Information Technology Letters, vol. 5, no. 2, pp. 13–24, Nov. 2024, doi: 10.31763/sitech.v5i2.1746.
[51] S. S. A. Larasati, E. N. K. Dewi, B. H. Farhansyah, F. A. Bachtiar, and F. Pradana, "Penerapan Decision Tree dan Random Forest dalam Deteksi Tingkat Stres Manusia Berdasarkan Kondisi Tidur," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 11, no. 5, pp. 1043–1050, Oct. 2024, doi: 10.25126/jtiik.2024117993.
[52] S. E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," Journal of Mathematics: Theory and Applications, pp. 21–26, Aug. 2022, doi: 10.31605/jomta.v4i1.1792.
[53] M. M. M. I, "Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," MULTINETICS, vol. 9, no. 1, pp. 43–50, May 2023, doi: 10.32722/multinetics.v9i1.5578.
[54] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.
[55] A. S. Wibowo and Rusindiyanto, "Analisis Churn Nasabah Bank Dengan Pendekatan Machine Learning dan Pengelompokan Profil Nasabah dengan Pendekatan Clustering," Konstruksi: Publikasi Ilmu Teknik, Perencanaan Tata Ruang dan Teknik Sipil, vol. 2, no. 1, pp. 30–41, Jan. 2024, doi: 10.61132/konstruksi.v2i1.43.
[56] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, and E. Hossain, "Investigating customer churn in banking: a machine learning approach and visualization app for data science and management," Data Science and Management, vol. 7, no. 1, pp. 7–16, 2024, doi: https://doi.org/10.1016/j.dsm.2023.09.002.