PEMANFAATAN ALGORITMA ADASYN DAN SUPPORT VECTOR MACHINE DALAM MENINGKATKAN AKURASI PREDIKSI KANKER PARU-PARU

Mustika Tiara Triani Br Sirait, Nuraini Siti Fathonah, Mohamad Nurkamal Fauzan
Teknik Informatika, Universitas Logistik & Bisnis Internasional
Jalan Sariasih No.54, Sarijadi, Sukasari, Kota Bandung, Indonesia
1204027@std.ulbi.ac.id

ABSTRAK

Penelitian ini bertujuan untuk meningkatkan akurasi prediksi kanker paru-paru dengan memanfaatkan parameter ADASYN dan algoritma *Support Vector Machine* (SVM). Ketidakseimbangan data merupakan salah satu tantangan utama dalam klasifikasi medis, di mana kelas minoritas seringkali diabaikan oleh model prediksi konvensional. Hal ini menyebabkan rendahnya sensitivitas dan spesifisitas dalam mendeteksi kasus kanker paru-paru yang jarang terjadi. Untuk mengatasi masalah ini, ADASYN digunakan untuk menyeimbangkan dataset dengan menghasilkan sampel sintetis untuk kelas minoritas secara adaptif berdasarkan distribusi data, sehingga model dapat lebih responsif terhadap variasi dalam kelas tersebut. Algoritma SVM kemudian diterapkan pada dataset yang telah seimbang untuk memisahkan kelas dengan margin terbesar, yang bertujuan untuk meningkatkan akurasi prediksi. Hasil pengujian menunjukkan bahwa kombinasi penggunaan ADASYN dan SVM berhasil meningkatkan kinerja model dengan mencapai skor akurasi sebesar 98.95%. Temuan ini menunjukkan bahwa pemanfaatan parameter ADASYN dalam *pre-processing* data dan penerapan SVM sebagai algoritma klasifikasi dapat secara signifikan meningkatkan akurasi prediksi pada kasus kanker paru-paru. Dengan demikian, penelitian ini memberikan kontribusi penting bagi pengembangan sistem deteksi dini yang lebih efektif dan andal, yang pada akhirnya dapat meningkatkan hasil pengobatan dan keselamatan pasien.

Kata kunci : *ADASYN, Support Vector Machine, Kanker Paru-paru*

1. PENDAHULUAN

Kanker paru-paru tergolong penyakit mematikan dengan tingkat kematian tinggi di berbagai negara[1]. Deteksi dini dan prediksi akurat sangatlah krusial untuk meningkatkan peluang sembuh dan menekan angka kematian. Teknologi dan metode analisis data, khususnya machine learning, telah memberikan kontribusi penting dalam upaya mendeteksi dan memprediksi keberadaan kanker paru-paru.

Support Vector Machine (SVM) merupakan algoritma klasifikasi populer yang menghasilkan model prediksi kuat dengan margin antar kelas maksimal[2]. Namun, penerapan SVM dalam prediksi kanker paru-paru terhambat oleh ketidakseimbangan data. Hal ini terjadi karena jumlah sampel pada kelas minoritas (pasien kanker) jauh lebih sedikit dibandingkan kelas mayoritas (pasien non-kanker)[3]. Ketidakseimbangan data dapat menyebabkan model SVM cenderung mengabaikan kelas minoritas, sehingga mengurangi akurasi prediksi[4] untuk kelas tersebut.

Untuk menyeimbangkan data yang tidak seimbang, terdapat beberapa metode yang dapat digunakan, seperti resampling, teknik sampling hibrid, algoritma khusus, penambahan bobot, *cluster-based methods*, dan *synthetic* data *generation*[5]. Salah satu metode yang dapat digunakan adalah ADASYN (*Adaptive Synthetic Sampling*), yang menambah sampel sintetis secara adaptif berdasarkan distribusi kelas minoritas.

Penelitian yang dilakukan oleh (2023) membahas penggunaan algoritma pembelajaran terawasi seperti SVM, *Random Forest*, dan *Neural Networks* untuk prediksi kanker paru-paru, menyoroti berbagai teknik dan metrik evaluasi untuk meningkatkan akurasi prediksi [6]. Dari studi literatur yang telah dilakukan, salah satu artikel membahas mengenai SVM dan algoritma lainnya untuk membandingkan klasifikasi di antara algoritma seperti *Regresi* Logistik, KNN, *Decision Tree*, *Naive Bayes*, dan SVM. Akhirnya, SVM menunjukkan akurasi maksimum sebesar 98% dalam prediksi[6].

Penelitian ini meneliti penggunaan parameter ADASYN untuk meningkatkan akurasi prediksi kanker paru-paru dengan algoritma SVM. ADASYN diharapkan dapat menyeimbangkan distribusi kelas dalam dataset, sehingga SVM lebih efektif mengenali pola dan karakteristik kelas minoritas. Hasil penelitian diharapkan dapat berkontribusi pada deteksi dini kanker paru-paru dan meningkatkan kualitas hidup[7] pasien melalui pengambilan keputusan medis yang lebih akurat.

Langkah-langkah utama dalam penelitian ini meliputi persiapan data, penerapan ADASYN untuk oversampling, pelatihan model SVM, dan evaluasi kinerja model. Evaluasi akan dilakukan dengan mengukur akurasi, sensitivitas, spesifisitas, dan metrik lainnya untuk menilai sejauh mana ADASYN berhasil meningkatkan performa prediksi SVM dalam mendeteksi kanker paru-paru. Dengan demikian, penelitian ini diharapkan dapat memberikan wawasan baru tentang pemanfaatan teknik oversampling dalam meningkatkan akurasi prediksi pada kasus ketidakseimbangan data medis[8].

2. TINJAUAN PUSTAKA

Berikut adalah tinjauan pustaka yang bertujuan untuk mengkaji ulang berbagai literatur dari penelitian

2.1. Kanker Paru - Paru

Kanker paru-paru adalah jenis kanker yang dianggap sebagai "*monster* diam"[9]. Hal ini karena kanker paru-paru sering kali baru terdeteksi saat sudah berada pada tahap lanjut, mirip dengan kasus kanker *liver* atau pankreas. Merokok dapat menyebabkan 40% kasus kematian, dan juga untuk terpapar asap rokok orang lain (merokok pasif) juga dapat menyebabkan kanker paru-paru[10]. Faktor keturunan[11] menjadi salah satu faktor penyebab kanker paru-paru.Faktor lingkungan yakni paparan terhadap polutan seperti polusi kendaraan, emisi industri, dan gas radon adalah penyebab kedua terbanyak kematian akibat kanker paru-paru[12].

2.2. ADASYN

ADASYN (*Adaptive Synthetic Sampling Approach for Imbalanced Learning*) termasuk dalam kategori algoritma. Algoritma ini dirancang khusus untuk mengatasi masalah ketidakseimbangan kelas dalam dataset[13]. ADASYN bekerja dengan mensintesis sampel data baru untuk kelas minoritas berdasarkan distribusi lokal dari data, sehingga membantu model pembelajaran mesin dalam mengenali dan memprediksi pola dari kelas minoritas dengan lebih baik.

Rumus dasar yang digunakan dalam algoritma ADASYN[13] seperti formula 1 Hitung rasio ketidakseimbangan.

$$r_{\mathrm{i}} = \frac{d_i}{N_{\mathrm{i}}} \qquad (1)$$

Dimana r_{i} adalah rasio ketidakseimbangan untuk setiap sampel minoritas i, d_{i} adalah jumlah tangga mayoritas dari sampel minoritas i, dan N_{i} adalah jumlah total tetangga yang dipertimbangkan.

Rumus dasar yang digunakan dalam algoritma ADASYN seperti formula 2 Hitung bobot untuk setiap sampel minoritas.

$$g_{\mathrm{i}} = \frac{r_{\mathrm{i}}}{\sum_{j-1}^{N_{\mathrm{m}}} r_j} \qquad (2)$$

Dimana g_{i} adalah bobot yang menunjukkan seberapa banyak sampel sintetis yang harus dibuat untuk setiap sampel minoritas i, *N* adalah jumlah total sampel minoritas.

Rumus dasar yang digunakan dalam algoritma ADASYN seperti formula 3 Hitung jumlah total sampel sintetis yang dibutuhkan.

$$G = (N_{\mathrm{m}} - N_{maj}) \cdot \beta \qquad (3)$$

Dimana G adalah jumlah total sampel sintetis yang akan dibuat, N_{maj} adalah jumlah sampel mayoritas, dan β adalah parameter yang menentukan tingkat oversampling (biasanya antara 0 dan 1).

Rumus dasar yang digunakan dalam algoritma ADASYN seperti formula 4 Ciptakan sampel sintetis.

$$x_{synth} = N_{\mathrm{i}} + \delta \cdot (x_j - x_i) \qquad (4)$$

Dimana δ adalah nilai acak antara 0 dan 1.

2.3. Support Vector Machine

Support Vector Machine (SVM) adalah salah satu algoritma pembelajaran mesin yang paling efektif untuk klasifikasi dan prediksi[14]. SVM bekerja dengan menemukan hyperplane optimal yang memisahkan kelas-kelas dalam dataset[15]. Dalam konteks prediksi kanker, SVM telah terbukti sangat efektif karena kemampuannya untuk menangani data dengan dimensi tinggi dan menghasilkan margin klasifikasi yang jelas antara sel normal dan sel kanker. Vapnik dalam bukunya menjelaskan teori dasar SVM dan aplikasinya dalam berbagai bidang, termasuk prediksi penyakit[16].

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi[14]. Berikut adalah rumus dasar prediksi menggunakan SVM untuk klasifikasi:

$$f(\mathbf{X}) = \mathbf{W} \cdot \mathbf{X} + b \qquad (1)$$

Rumus untuk menghitung klasifikasi linear SVM

Keterangan:

W = vector bobot

X = vector fitur dari sampel yang akan diprediksi

b = bias atau intersep

$$f(\mathbf{X}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{X}) + b \qquad (2)$$

Rumus untuk menghitung klasifikasi non-linear SVM

Keterangan:

N = jumlah total sampel dalam dataset pelatihan.

α_i = koefisien Lagrange yang diperoleh dari pelatihan.

y_i = label kelas dari sampel pelatihan i.

$K(\mathbf{x}_i, \mathbf{X})$ = fungsi kernel yang mengukur kesamaan antara sampel pelatihan $\mathbf{x}_i$ dan sampel yang akan diprediksi **X.**

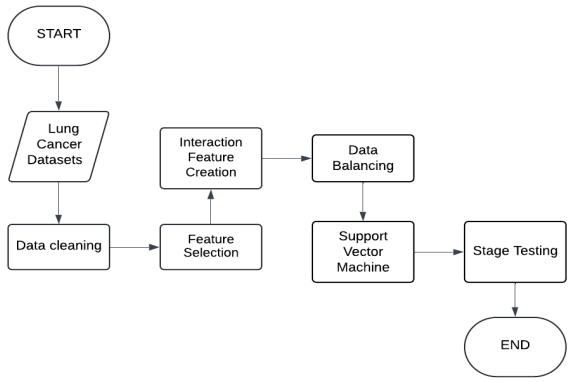
b = Bias atau intersep

2.4. Diagnosis Dini

Diagnosis dini merujuk pada identifikasi penyakit pada tahap awal, sebelum gejala menjadi parah. Pendekatan ini sangat penting dalam meningkatkan hasil pengobatan dan peluang kesembuhan. Diagnosis dini memungkinkan intervensi medis segera, yang sering kali lebih efektif dan kurang invasif dibandingkan dengan pengobatan penyakit pada tahap lanjut. Sebagai contoh, dalam kasus kanker, deteksi dini dapat dilakukan melalui pemeriksaan rutin dan pengenalan gejala awal.[17] Strategi ini terbukti meningkatkan keberhasilan pengobatan dan mengurangi angka kematian akibat kanker (Cancer Research UK). Menurut National Cancer Institute (2019), deteksi dini kanker paru-paru dapat meningkatkan angka harapan hidup lima tahun dari sekitar 18% menjadi lebih dari 55% jika kanker ditemukan pada tahap awal.

3. METODE PENELITIAN

Dalam penelitian ini, telah dilakukan pengujian prediksi kanker paru-paru menggunakan parameter dan tidak menggunakankan parameter. Perbandingan dari penggunaan parameter dan tidak menggunakan parameter dapat dilihat pada gambar 1.



Gambar 1. Alur Penelitian

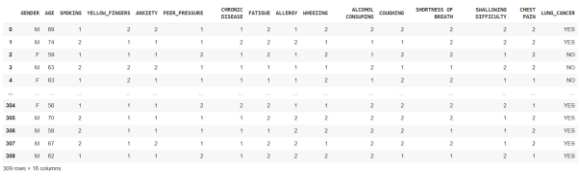
Dari gambar 1 diagram alur penelitian ini diadopsi dari metode Chrisp-dm. Proses dimulai dengan mengumpulkan dataset kanker paru-paru dan melakukan pembersihan data untuk menghilangkan *missing values*, duplikasi, dan noise. Setelah itu, fitur-fitur yang relevan dipilih untuk analisis, diikuti dengan penciptaan fitur interaksi untuk menangkap hubungan non-linear antara fitur-fitur yang ada. Selanjutnya, dilakukan penyeimbangan data untuk menghindari bias dalam model, menggunakan teknik seperti oversampling[18]. Teknik ini menambahkan jumlah sampel di kelas minoritas dengan mensintesis data baru, sehingga membantu mengatasi ketidakseimbangan kelas dalam dataset dan meningkatkan performa model machine learning dalam mengenali pola dari kelas minoritas. Teknik yang digunakan adalah ADASYN (*Adaptive Synthetic Sampling Approach for Imbalanced Learning*) Algoritma *Support Vector Machin*e (SVM) kemudian diterapkan untuk membangun model prediksi, yang selanjutnya diuji dengan data uji untuk mengevaluasi performa dan akurasinya. Proses ini berakhir setelah pengujian selesai.

.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Salah satu tahapan utama dalam penelitian adalah pengumpulan data yang memungkinkan peneliti menemukan jawaban atas pertanyaan penelitian. Pengumpulan data bertujuan untuk memperoleh wawasan mengenai topik penelitian.



	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
...																
304	F	56	1	1	1	2	2	2	1	1	2	2	2	2	1	YES
305	M	70	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
306	M	58	2	1	1	1	1	1	2	2	2	2	1	1	2	YES
307	M	67	2	1	2	1	1	2	2	1	2	2	2	1	2	YES
308	M	62	1	1	1	2	1	2	2	2	2	1	1	2	1	YES

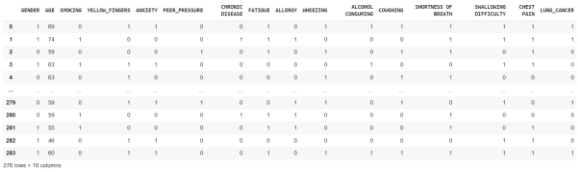
309 rows × 16 columns

Gambar 2. Dataset Mentah Kanker Paru-Paru

Dari penelitian ini data yang digunakan bersumber dari kaggle. Dataset kanker paru-paru merupakan Data mentah yang berisi informasi terkait kanker paru-paru. Adapun data historis yang digunakan memiliki total *record* sebanyak 309 record, dan memiliki jumlah kolom sebanyak 16 kolom.

4.2. Pembersihan data

Proses ini mencakup menghapus atau mengimputasi nilai yang hilang untuk memastikan tidak ada data yang kosong yang dapat mempengaruhi analisis. Langkah berikutnya adalah mengidentifikasi dan menangani pencilan (*outliers*) yang dapat mengganggu hasil analisis atau menyebabkan bias dalam model. Kesalahan dalam data seperti duplikasi atau inkonsistensi juga diperbaiki untuk memastikan data yang digunakan akurat dan konsisten. Selain itu, normalisasi atau standarisasi data dilakukan untuk memastikan bahwa semua fitur berada dalam skala yang sama, yang penting untuk algoritma pembelajaran mesin yang sensitif terhadap skala fitur.



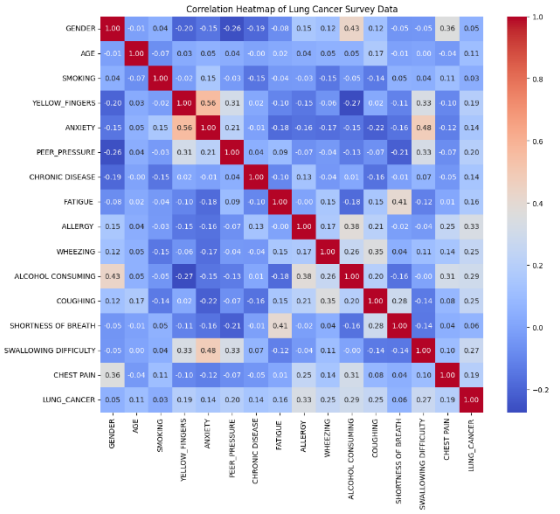
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	1	69	0	1	1	0	0	1	0	1	1	1	1	1	1	1
1	1	74	1	0	0	0	1	1	1	0	0	0	1	1	1	1
2	0	59	0	0	0	1	0	1	0	1	0	1	1	0	1	0
3	1	63	1	1	1	0	0	0	0	0	1	0	0	1	1	0
4	0	63	0	1	0	0	0	0	0	1	0	1	1	0	0	0
...																
279	0	59	0	1	1	0	0	1	1	0	1	1	0	1	0	1
280	0	59	1	0	0	0	1	1	1	0	0	0	1	0	0	0
281	1	55	1	0	0	0	1	1	0	0	0	1	0	1	0	0
282	1	46	0	1	1	0	0	0	0	0	0	0	1	1	0	0
283	1	60	0	1	1	0	0	1	1	1	1	1	1	1	1	1

276 rows × 16 columns

Gambar 3. Cleaning Data

Pembersihan data adalah langkah krusial untuk meningkatkan kualitas data dan memastikan bahwa analisis dan model yang dibangun berdasarkan data tersebut memiliki akurasi yang tinggi.

4.3. Pemilihan Fitur



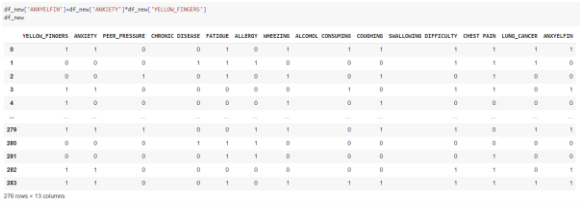
Gambar 4. Correlation Heatmap

Pemilihan fitur[1] dengan correlation heatmap melibatkan analisis korelasi antar fitur untuk mengidentifikasi dan memilih fitur yang paling relevan terhadap variabel target. Langkah-langkahnya meliputi menghitung matriks korelasi untuk mengukur

hubungan antar semua fitur dalam dataset, kemudian memvisualisasikan matriks tersebut dalam bentuk heatmap untuk memudahkan interpretasi. Fokus utama adalah pada fitur yang memiliki korelasi tinggi (positif atau negatif) dengan variabel target, karena fitur-fitur ini lebih mungkin memberikan informasi berguna untuk prediksi. Selain itu, fitur yang terlalu berkorelasi satu sama lain dapat dieliminasi untuk mengurangi redundansi dan risiko overfitting.

4.4. Pembuatan Fitur Interaksi

Pembuatan Fitur Interaksi (*Interaction Feature Creation*) adalah Membuat fitur baru dengan menginteraksikan fitur yang dipilih untuk meningkatkan kinerja model. Pembuatan fitur interaksi melibatkan penciptaan fitur baru berdasarkan interaksi antara fitur-fitur yang ada untuk menangkap hubungan non-linier yang mungkin tidak terlihat dalam fitur asli. Misalnya, mengalikan atau menggabungkan dua fitur yang ada untuk menciptakan variabel baru yang lebih informatif.



	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER	ANXYELFIN
0	1	1	0	0	1	0	1	1	1	1	1	1	1
1	0	0	0	1	1	1	0	0	0	1	1	1	0
2	0	0	1	0	1	0	1	0	1	0	1	0	0
3	1	1	0	0	0	0	0	1	0	1	1	0	1
4	1	0	0	0	0	0	1	0	1	0	0	0	0
...													
279	1	1	1	0	0	1	1	0	1	1	0	1	1
280	0	0	0	1	1	1	0	0	0	0	0	0	0
281	0	0	0	0	1	1	0	0	0	0	1	0	0
282	1	1	0	0	0	0	0	0	0	1	1	0	1
283	1	1	0	0	1	0	1	1	1	1	1	1	1

276 rows × 13 columns

Gambar 5. Penggabungan & Penambahan Kolom

Fitur interaksi dapat membantu model untuk memahami kompleksitas data dengan lebih baik dan meningkatkan kinerjanya. Proses ini biasanya dimulai dengan mengidentifikasi fitur yang berpotensi memiliki hubungan interaksi signifikan, kemudian menciptakan fitur baru melalui operasi matematika seperti perkalian, pembagian, penjumlahan, atau pengurangan antara fitur-fitur tersebut. Setelah fitur interaksi dibuat, penting untuk melakukan evaluasi untuk memastikan bahwa fitur-fitur baru tersebut benar-benar meningkatkan kinerja model. Dengan menciptakan fitur interaksi yang relevan, kita dapat memberikan lebih banyak informasi kepada model, yang pada akhirnya dapat meningkatkan kemampuan prediksi dan akurasi model dalam analisis data

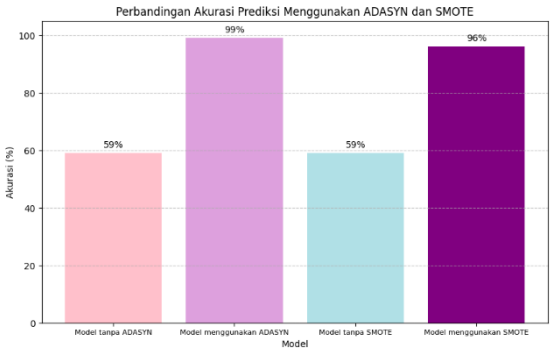
4.5. Data *Balancing*

Penyeimbangan data adalah proses penting untuk memastikan bahwa kelas dalam dataset seimbang, sehingga model pembelajaran mesin tidak bias terhadap kelas mayoritas. Salah satu teknik yang efektif untuk ini adalah ADASYN (*Adaptive Synthetic Sampling Approach*). ADASYN bekerja dengan membuat sampel sintetis dari kelas minoritas berdasarkan distribusi data, yang membantu memperbaiki ketidakseimbangan dengan lebih adaptif.

```
from imblearn.over_sampling import ADASYN
adasyn = ADASYN(random_state=42)
X, y = adasyn.fit_resample(X, y)
```

Gambar 6. implementasi ADASYN dengan SVM

Langkah-langkahnya meliputi: pertama, mengidentifikasi jumlah sampel yang dibutuhkan untuk mencapai keseimbangan antara kelas; kedua, menghasilkan sampel sintetis baru untuk kelas minoritas dengan mempertimbangkan kesulitan belajar yang berbeda dari sampel minoritas, di mana data yang lebih sulit (berdasarkan *Support Vector Machine*) akan dihasilkan lebih banyak sampel baru. Proses ini membantu model untuk belajar dari data minoritas yang lebih *representative*.



Gambar 7. Perbandingan Akurasi Prediksi

Bar chart perbandingan pada gambar 7, menunjukkan perbandingan akurasi prediksi model dalam mendeteksi Kanker Paru-Paru dengan dan tanpa menggunakan teknik penanganan ketidakseimbangan data, yaitu ADASYN dan SMOTE[19]. Model tanpa teknik penanganan ketidakseimbangan menunjukkan akurasi 59%, baik dengan maupun tanpa SMOTE.

Namun, ketika ADASYN diterapkan, akurasi model meningkat signifikan menjadi 99%, menunjukkan efektivitas ADASYN dalam mengatasi ketidakseimbangan kelas dan meningkatkan performa prediksi. Sementara itu, penggunaan SMOTE juga meningkatkan akurasi model menjadi 96%, meskipun sedikit kurang efektif dibandingkan ADASYN. Hasil ini menekankan pentingnya teknik penanganan ketidakseimbangan data untuk meningkatkan akurasi prediksi pada dataset yang tidak seimbang.

SMOTE (*Synthetic Minority Over-sampling Technique*) dan ADASYN (*Adaptive Synthetic Sampling*) adalah dua teknik oversampling yang dirancang untuk mengatasi. ketidakseimbangan kelas dalam dataset, tetapi mereka memiliki perbedaan dalam cara mereka menghasilkan contoh sintetis.

4.6. *Support Vector Machine Testing*

Tahap ini melibatkan penerapan algoritma *Support Vector Machine* untuk klasifikasi. Pemodelan SVM, Langkah pertama diambil adalah membuat prediksi kemudia akan dinilai untuk akurasi, confusion matrix[20], dan classification report. Semua proses

tersebut menggunakan Google Colab. Berikut adalah daftar hasil pengujian pada gambar 7,8 & 9.

```
#Model accuracy
from sklearn.metrics import classification_report, accuracy_score, f1_score
svc_cr=classification_report(y_test, y_svc_pred)
print(svc_cr)
```

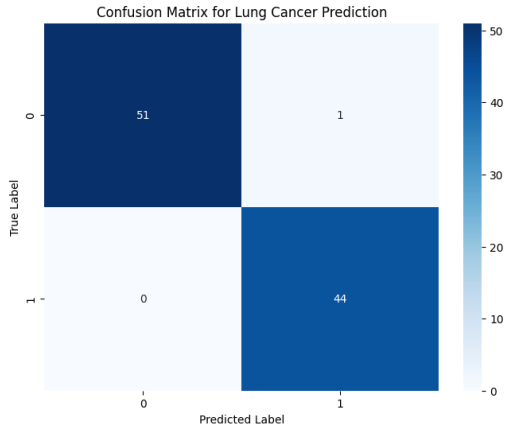
```
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        52
           1       0.98      1.00      0.99        44

    accuracy                           0.99        96
   macro avg       0.99      0.99      0.99        96
weighted avg       0.99      0.99      0.99        96
```

Gambar 8. Classification Report

```
from sklearn.metrics import accuracy_score,confusion_matrix,precision_score,f1_score,recall_score
print("accuracy score:", (accuracy_score(y_svc_pred, y_test))*100, "%")
```

```
accuracy score: 98.95833333333334 %
```

Gambar 9. Accuracy Score



Gambar 10. Confusion Matrix

5. KESIMPULAN DAN SARAN

Pengujian menunjukkan bahwa penggunaan parameter ADSYN untuk menyeimbangkan data kelas dan tuning parameter pada algoritma *Support Vector Machine* secara signifikan meningkatkan akurasi prediksi kanker paru-paru. ADASYN berhasil mengatasi ketidakseimbangan kelas dengan menciptakan sampel sintetis dari kelas minoritas, yang menghasilkan distribusi data yang lebih seimbang dan meningkatkan kinerja model. Evaluasi menunjukan metrik akurasi, precision, recall dan F1-score menunjukan penungkatan yang jelas dalam akurasi dan kemampuan klasifikasi model. Dengan demikian, kombinasi ADASYN untuk penyeimbangan data dan optimasi parameter SVM efektif dalam meningkatkan prediksi kanker paru-paru. Disarankan untuk memperluas ukuran dataset dengan mengumpulkan data dari berbagai rumah sakit atau institusi medis. Dataset yang lebih besar dan lebih beragam dapat membantu meningkatkan generalisasi model dan akurasi prediksi. Selain itu, penelitian lebih lanjut dapat mengeksplorasi teknik oversampling lainnya seperti SMOTE, Borderline-SMOTE, atau kombinasi teknik oversampling dan undersampling untuk melihat apakah ada peningkatan lebih lanjut dalam akurasi prediksi.

DAFTAR PUSTAKA

[1] E. Nemlander et al., “Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers,” PLoS One, vol. 17, no. 10 October, Oct. 2022, doi: 10.1371/journal.pone.0276703.

[2] Dr. M. Kasthuri and M. R. Jency, “Improving the Performance of Lung Cancer Prediction Using Machine Learning Techniques on Big Data,” International Journal of Computer Science and Mobile Computing, vol. 9, no. 10, pp. 64–72, Oct. 2020, doi: 10.47760/ijcsmc.2020.v09i10.008.

[3] Z. Wan, “Prediction and Visualization Analysis of Lung Cancer Risk by Machine Learning,” 2023.

[4] A. C. Pacurari et al., “Diagnostic Accuracy of Machine Learning AI Architectures in Detection and Classification of Lung Cancer: A Systematic Review,” Jul. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/diagnostics13132145.

[5] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, “Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,” in 2022 IEEE World AI IoT Congress, AIIoT 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.

[6] M. Elgohary, H. A. {2}, and A. El, “Prediction of Lung Cancer Using Supervised Machine Learning,” 2023. [Online]. Available: https://shorturl.at/vwzM4.This

[7] A. S. Ahmad and A. M. Mayya, “A new tool to predict lung cancer based on risk factors,” Heliyon, vol. 6, no. 2, Feb. 2020, doi: 10.1016/j.heliyon.2020.e03402.

[8] N. Zeinali, N. Youn, A. Albashayreh, W. Fan, and S. G. White, “Machine Learning Approaches to Predict Symptoms in People With Cancer: Systematic Review,” 2024, JMIR Publications Inc. doi: 10.2196/52322.

[9] H. Yu, Z. Zhou, and Q. Wang, “Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model,” IEEE Access, vol. 8, pp. 86400–86410, 2020, doi: 10.1109/ACCESS.2020.2992645.

[10] A. J. Didier, A. Nigro, Z. Noori, M. A. Omballi, S. M. Pappada, and D. M. Hamouda, “Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis,” 2024, Frontiers Media SA. doi: 10.3389/frai.2024.1365777.

[11] K. Wadowska, I. Bil-Lula, Ł. Trembecki, and M. Śliwińska-Mossoń, “Genetic markers in lung cancer diagnosis: A review,” Jul. 01, 2020, MDPI AG. doi: 10.3390/ijms21134569.

[12] M. B. Schabath and M. L. Cote, "Cancer progress and priorities: Lung cancer," Cancer Epidemiology Biomarkers and Prevention, vol. 28, no. 10, pp. 1563–1579, 2019, doi: 10.1158/1055-9965.EPI-19-0221.

[13] M. Rahman, Y. Zhou, S. Wang, and J. Rogers, "Wart Treatment Decision Support Using Support Vector Machine," International Journal of Intelligent Systems and Applications, vol. 12, no. 1, pp. 1–11, Feb. 2020, doi: 10.5815/ijisa.2020.01.01.

[14] R. S. Krishna, "Machine Learning Approaches in Early Lung Cancer Prediction: A Comprehensive Review," INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 07, no. 09, Sep. 2023, doi: 10.55041/ijsrem25584.

[15] H. Syahputra and A. Wibowo, "Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), vol. 9, no. 1, pp. 165–173, 2023, doi: 10.26555/jiteki.v9i1.25861.

[16] C. S. Anita, G. Vasukidevi, D. Rajalakshmi, K. Selvi, and T. Ramesh, "Lung cancer prediction model using machine learning techniques," Int J Health Sci (Qassim), pp. 12533–12539, Jun. 2022, doi: 10.53730/ijhs.v6ns2.8306.

[17] L. Z. Du, "Early diagnosis and management of neonatal sepsis: a perspective," Apr. 01, 2024, Zhejiang University School of Medicine Children's Hospital. doi: 10.1007/s12519-024-00803-4.

[18] Z. Ceylan, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Diagnosis of Breast Cancer Using Improved Machine Learning Algorithms Based on Bayesian Optimization," 2020. [Online]. Available: www.ijisae.org

[19] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," Scientific Journal of Informatics, vol. 8, no. 2, pp. 276–282, Nov. 2021, doi: 10.15294/sji.v8i2.32484.

[20] G. T., V. Y. M., U. M., R. D., and R. B. K., "Prediction of Lung Cancer Risk using Random Forest Algorithm Based on Kaggle Data Set," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, pp. 1623–1630, Mar. 2020, doi: 10.35940/ijrte.F7879.038620