JURNAL BUFFER INFORMATIKA
Volume 11 Nomor 2, Oktober 2025
p-ISSN : 2527-4856, e-ISSN : 2614-5413
https://journal.fkom.uniku.ac.id/index.php/buffer

Klasifikasi Kecelakaan Lalu Lintas Menggunakan Kombinasi Forward Selection, ADASYN, dan Random Forest

Aspianur[1)] **, Taghfirul Azhima Yoga Siswa**[2)], **Naufal Azmi Verdikha**[3)]

[1,2,3)] *Sains Dan Teknologi, Universitas Muhammadiyah Kalimantan Timur*
Jl. Ir. H. Juanda No. 15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75124
Email :2111102441104@umkt.ac.id[1)], tay758@umkt.ac.id[2)], nav651@umkt.ac.id[3)]

Abstrak

Kecelakaan lalu lintas merupakan penyebab utama kematian bagi kelompok usia 15–29 tahun, dengan lebih dari 1,3 juta kematian setiap tahunnya. Di Indonesia, pada tahun 2023 tercatat lebih dari 100 ribu kasus kecelakaan, dengan korban jiwa mencapai 25 ribu orang. Penelitian ini bertujuan mengembangkan model klasifikasi tingkat keparahan kecelakaan lalu lintas dengan mengintegrasikan metode *ADASYN* dan *Forward Selection* ke dalam algoritma *Random Forest*. Data yang digunakan berasal dari Polresta Samarinda periode 2020–2024 dengan total 2.385 data dan 35 fitur. Proses integrasi dilakukan dengan menerapkan *ADASYN* terlebih dahulu untuk mengatasi ketidakseimbangan kelas, diikuti oleh pemilihan fitur terbaik melalui metode *Forward Selection* sebelum pemodelan menggunakan *Random Forest*. Evaluasi model dilakukan menggunakan *10-fold cross-validation* dan *confusion matrix* untuk mengukur akurasi, *precision*, *recall*, dan *f1-score*. Hasil penelitian menunjukkan bahwa fitur paling signifikan adalah cuaca, jumlah luka ringan, luka berat, dan meninggal dunia. Model awal tanpa penyeimbangan dan seleksi fitur hanya mencapai akurasi 79,26%. Setelah penerapan *ADASYN*, akurasi meningkat menjadi 84,26%, dan setelah integrasi *Forward Selection*, akurasi akhir mencapai 95,28%, disertai *precision* 94,23%, *recall* 96,15%, dan *f1-score* 95,18%. Temuan ini membuktikan bahwa integrasi *ADASYN* dan *Forward Selection* secara bertahap meningkatkan kinerja klasifikasi *Random Forest* secara signifikan pada data kecelakaan tidak seimbang.

Kata Kunci : Kecelakaan Lalu Lintas; Random Forest; Forward Selection; ADASYN; Klasifikasi.

Abstract

Traffic accidents are a leading cause of death among individuals aged 15–29, with over 1.3 million fatalities recorded annually. In Indonesia, more than 100,000 accident cases were reported in 2023, resulting in approximately 25,000 deaths. This study aims to develop a classification model for the severity level of traffic accidents by integrating Adaptive Synthetic Sampling (ADASYN) and Forward Selection methods into the Random Forest algorithm. The dataset, sourced from the Samarinda Police Department (Polresta Samarinda) for the period 2020–2024, consists of 2,385 records and 35 features. The integration process begins with the application of ADASYN to address class imbalance, followed by feature selection using the Forward Selection method before modeling with the Random Forest algorithm. Model evaluation was conducted using 10-fold cross-validation and a confusion matrix to assess accuracy, precision, recall, and F1-score. The results indicate that the most significant features are weather conditions, the number of minor injuries, serious injuries, and fatalities. The initial model, without balancing or feature selection, achieved an accuracy of only 79.26%. After applying ADASYN, accuracy improved to 84.26%, and further increased to 95.28% following the integration of Forward Selection, along with a precision of 94.23%, recall of 96.15%, and F1-score of 95.18%. These findings demonstrate that the integration of ADASYN and Forward Selection significantly enhances the classification performance of the Random Forest algorithm on imbalanced traffic accident data.

Keywords: Traffic Accidents; Random Forest; Forward Selection; ADASYN; Classification.

1. PENDAHULUAN

Kecelakaan lalu lintas merupakan isu serius baik secara global maupun nasional. WHO (2023) mencatat bahwa insiden ini menjadi penyebab utama kematian usia 15–29 tahun dengan lebih dari 1,3 juta korban setiap tahun. Di Indonesia, Korps Lalu Lintas Polri mencatat lebih dari 100.000 kasus pada 2023 dengan 25.000 korban jiwa. Di Samarinda, jumlah kecelakaan meningkat dari 97 kasus (2021) menjadi 173 kasus (2023) [1], mayoritas melibatkan sepeda motor dan kendaraan pribadi. Faktor penyebabnya antara lain kepadatan lalu lintas, infrastruktur tidak merata [2], serta pelanggaran rambu lalu lintas [3].

Dalam konteks ini, pendekatan berbasis data, khususnya machine learning, diperlukan untuk klasifikasi dan prediksi tingkat keparahan kecelakaan. *Random Forest* banyak digunakan karena kemampuannya mengelola data besar dan kompleks, namun rentan bias pada data tidak seimbang. [4] menunjukkan bahwa kombinasi algoritma klasifikasi seperti *Random Forest* dengan teknik resampling (*SMOTE-ENN*) mampu meningkatkan akurasi prediksi risiko kecelakaan lalu lintas. Temuan ini diperkuat oleh [5] mengungkapkan bahwa tanpa penyeimbangan data, performa *Random Forest* dapat turun drastis, bahkan di bawah 83%.

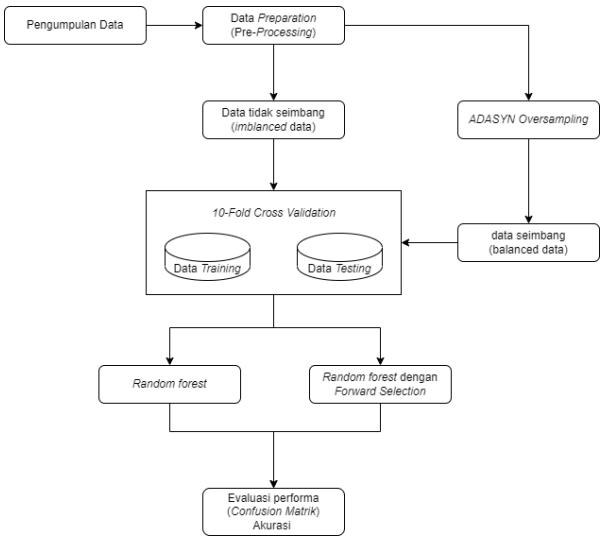
Studi lain [6], [7] menunjukkan pentingnya penerapan klasifikasi kecelakaan untuk mendukung pengambilan keputusan strategis dalam manajemen keselamatan transportasi. Penelitian oleh [8] dan [9] menunjukkan bahwa pemilihan fitur relevan dapat meningkatkan akurasi model, meskipun belum mengintegrasikan seleksi fitur dan penyeimbangan data secara bersamaan.

Selain ketidakseimbangan data, tingginya dimensi fitur juga menjadi tantangan dalam klasifikasi. Pengurangan fitur dengan teknik seperti *Forward Selection* terbukti meningkatkan akurasi dan interpretabilitas model [10], [11]. Sementara itu, *ADASYN* dinilai lebih adaptif dibanding *SMOTE* dan *Random Under Sampling* dalam menangani data tidak seimbang [12].

Berdasarkan hal tersebut, penelitian ini mengangkat dua fokus utama pertama, untuk mengidentifikasi fitur-fitur yang paling berpengaruh dalam menentukan tingkat keparahan kecelakaan lalu lintas di Kota Samarinda dan kedua, untuk mengevaluasi efektivitas integrasi algoritma *Random Forest* dengan metode *ADASYN* dan *Forward Selection* dalam meningkatkan performa klasifikasi. Penelitian ini bertujuan mengembangkan model klasifikasi yang tidak hanya akurat, tetapi juga seimbang dan efisien. Secara khusus, kontribusi dari penelitian ini adalah menghadirkan pendekatan terpadu yang menggabungkan teknik penyeimbangan data dan seleksi fitur dalam satu kerangka pemodelan, sehingga menghasilkan model klasifikasi *Random Forest* yang lebih optimal dalam menangani data kecelakaan lalu lintas dengan distribusi kelas yang tidak merata. Diharapkan hasil penelitian ini dapat memberikan dasar pengambilan keputusan berbasis data bagi pihak berwenang dalam merancang kebijakan keselamatan lalu lintas yang lebih tepat sasaran.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk mengembang kan model klasifikasi yang dapat mengidentifikasi faktor penyebab kecelakaan lalu lintas di kota Samarinda menggunakan algoritma *Random Forest*, serta mengatasi masalah ketidakseimbangan data melalui metode *ADASYN* dan melakukan seleksi fitur menggunakan metode *Forward Selection*. Berikut ini adalah prosedur penelitian yang akan diterapkan dalam penelitian ini:



Gambar 1. Alur Penelitian

Pada gambar 1. ini menggambarkan tahapan penelitian klasifikasi kecelakaan lalu lintas. Data kecelakaan dari Polresta Samarinda diproses melalui tahap pembersihan (*pre-processing*), lalu diseimbangkan menggunakan metode *ADASYN* untuk mengatasi ketidakseimbangan kelas. Selanjutnya, data dibagi dengan teknik *10-Fold Cross Validation*, lalu dimodelkan menggunakan *Random Forest*, baik tanpa maupun dengan *Forward Selection* untuk pemilihan fitur. Evaluasi dilakukan menggunakan *confusion matrix* dengan metrik akurasi, *precision*, *recall*, dan *f1-score*.

2.2 Pengumpulan Data

Pada penelitian ini menggunakan data laka lantas yang bersifat privasi yang didapatkan secara langsung di Polresta Kota Samarinda. Polresta Kota Samarinda dengan izin akses yang sudah disetujui. Pada periode tahun 2020-2024 . dari instansi tersebut mendapatkan fitur-fitur yang di gunakan meliputi tanggal kejadian, waktu kejadian, tingkat kecelakaan, jumlah meninggal dunia, jumlah luka ringan, jumlah luka berat, koordinat GPS, cuaca, fungsi jalan, kondisi permukaan jalan, batas kecepatan di lokasi, penyebab. Lokasi penelitian di Polresta Kota Samarinda Jl. Slamet Riyadi No.1, Karang Asam Ulu, Kec. Sungai Kunjang, Kota Samarinda, Kalimantan Timur 75126.

2.3 Data Pre-processing

Empat tahapan utama dilakukan:

a. Data *Selection*
   Menghilangkan atribut administratif yang tidak relevan.
b. Data *Cleaning*
   Menghapus duplikat dan nilai kosong menggunakan *dropna()* serta konversi format waktu.
c. Data *Transformation*
   Label encoding dan normalisasi dilakukan agar data kategori dan numerik dapat diproses oleh algoritma.
d. Data *Balancing*
   ADASYN digunakan untuk menyeimbangkan distribusi antara kecelakaan ringan/sedang dan berat, sebagaimana direkomendasikan oleh [12].

2.4 Pembagian Data

Dataset dibagi menggunakan metode *10fold cross-validation* [13] untuk menghindari bias dan memberikan estimasi performa yang stabil.



Gambar 2. Diagram Alur *10k-fold validation*

Pada gambar 2. menunjukkan skema *10-Fold Cross Validation*, yaitu proses validasi model di mana data dibagi menjadi 10 bagian. Setiap bagian (kuning) secara bergiliran digunakan sebagai data uji, sementara 9 bagian lainnya (biru) menjadi data latih. Proses ini diulang 10 kali hingga seluruh data pernah digunakan sebagai data uji. Tujuannya adalah untuk memperoleh evaluasi model yang lebih akurat, stabil, dan tidak bias, karena seluruh data dimanfaatkan secara bergantian untuk pelatihan dan pengujian. Skema ini umum digunakan untuk meningkatkan keandalan hasil klasifikasi.

2.5 Pemodelan *Random Forest*

RandomForest digunakan karena kemampuannya dalam menangani data yang kompleks dan beragam [4], [14]. Model ini dibangun dengan *RandomForestClasifier* dan evaluasi awal dilakukan menggunakan metrik akurasi, *presisi*, *recall*, dan skor *f1-score*.

```
# Buat dan latih model Random Forest
clf = RandomForestClassifier(
    n_estimators=100,
    max_depth=5,
    class_weight='balanced',
    random_state=42
)
clf.fit(X_train, y_train)

# Prediksi
y_pred = clf.predict(X_test)
```

Gambar 3. *hyperparameter Random Forest*

Pada gambar 3. menunjukkan proses pelatihan model *Random Forest Classifier* dengan beberapa pengaturan *hyperparameter*. Model dikonfigurasi dengan *n_estimators=100* (jumlah pohon dalam hutan), *max_depth=5* (kedalaman maksimum setiap pohon), dan *class_weight='balanced'* yang digunakan untuk menangani ketidakseimbangan kelas dengan memberikan bobot proporsional terhadap frekuensi masing-masing kelas. Selain itu, *random_state=42* ditetapkan agar hasil eksperimen dapat direplikasi. Setelah model dilatih menggunakan *clf.fit(X_train, y_train)*, dilakukan prediksi terhadap data uji dengan *clf.predict(X_test)* untuk mengevaluasi performa model.

2.6 *Random Forest* dengan *Forward Selection*

Untuk mengurangi dimensi fitur dan meningkatkan efisiensi, diterapkan metode *Sequential Forward Selection* [15].

```
# Gunakan Forward Selection
sfs = SequentialFeatureSelector(
    rf,
    n_features_to_select="auto",  # bisa diganti angka jika ingin fix jumlah fitur
    direction="forward",
    scoring="accuracy",
    cv=StratifiedKFold(n_splits=5),
    n_jobs=-1
)

# Fit selector ke data
sfs.fit(x, y)

# Ambil fitur terpilih
selected_features = x.columns[sfs.get_support()]

# Buat DataFrame hasil
df_selected = pd.DataFrame({'Selected Features': selected_features})
```

Gambar 4. Pengunaan seleksi fitur pada *Random Forest*

Pada gambar 4. di atas menerapkan *Forward Selection* menggunakan algoritma *Random Forest* untuk memilih fitur terbaik dalam klasifikasi data kecelakaan lalu lintas. Proses dimulai dengan membangun model *Random Forest*, lalu menerapkan *Sequential Forward Selection (SFS)* dengan *5-fold cross-validation* untuk memilih kombinasi fitur yang memberikan akurasi terbaik. Setelah fitur optimal dipilih, model *Random Forest* dilatih ulang hanya dengan fitur terpilih untuk meningkatkan performa klasifikasi.

2.7 Evaluasi Model

Model dievaluasi menggunakan *confusion matrix* yang mengukur TP, TN, FP, dan FN [16], [17]. Evaluasi dilakukan pada model sebelum dan sesudah penerapan *ADASYN* dan *Forward Selection* guna melihat peningkatan performa klasifikasi.

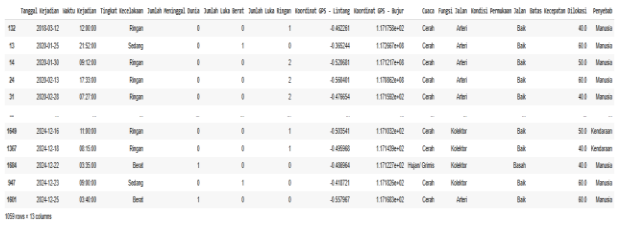
$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \; X \; 100\%$$

Keterangan :
TP (*True Positive*): Prediksi benar untuk kelas positif (minoritas).
TN (*True Negative*): Prediksi benar untuk kelas negatif (mayoritas).
FP (*False Positive):* Kelas negatif yang salah diprediksi sebagai positif.
FN (*False Negative*): Kelas positif yang salah diprediksi sebagai negatif.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pra-pemrosesan Data

Penelitian ini menggunakan data kecelakaan lalu lintas dari Polres Samarinda (2020–2024) yang meliputi 35 fitur. Melalui pemilihan atribut, dipilih 13 fitur relevan.



Gambar 5. Data *Selection*

Pada gambar 5. ini menunjukkan data kecelakaan yang telah dipilih dan diproses. Data terdiri dari 2.385 baris dan 35 fitur, dengan target klasifikasi yaitu Tingkat Kecelakaan (Ringan/Sedang = 0, Berat = 1). Beberapa fitur kategorikal seperti cuaca dan penyebab dikodekan menggunakan *Label Encoding* agar bisa digunakan dalam pemodelan *machine learning.*

Setelah data dibersihkan dari missing value dan duplikasi, maka tersisa data sebanyak 1.167 entri. Data kategorikal seperti cuaca, fungsi jalan, dan tingkat keparahan kemudian ditransformasikan menjadi *numerik* menggunakan teknik *label encoding*. Untuk mengatasi ketidakseimbangan kelas (911 data ringan/sedang dan 148 berat), digunakan metode *ADASYN* sehingga distribusinya menjadi seimbang (911 vs. 912).

3.2 Hasil Permodelan *Random Forest*

Pengujian awal dilakukan menggunakan model *Random Forest* tanpa *balancing* dan seleksi fitur. Evaluasi dengan *10-fold cross-validation* menunjukkan akurasi rata-rata sebesar 79.26%, dengan *precision* 27.86%, *Recall* 30.08%, dan *F1Score* 28.74%.

Tabel 1. Hasil Pengujian Model *Random Forest*

Fold	Akurasi	*Precision*	*Recall*	*F1 - Score*
1	78.69%	29.03%	34.62%	31.58%
2	75.41%	21.21%	26.92%	23.73%
3	78.57%	26.67%	32.00%	29.09%
4	79.67%	28.57%	32.00%	30.19%
5	81.87%	27.78%	20.00%	23.26%
6	79.67%	30.00%	36.00%	32.73%
7	81.32%	36.67%	42.31%	39.29%
8	80.22%	30.77%	30.77%	30.77%
9	77.47%	24.14%	26.92%	25.45%
10	79.67%	23.81%	19.23%	21.28%
Rata-rata	79.26%	27.86%	30.08%	28.74%

Pada table 1. menunjukkan bahwa meskipun akurasi *Random Forest* tanpa penyeimbangan data atau seleksi fitur cukup tinggi untuk akurasi (79.26%), nilai *precision* (27.86%), *recall* (30.08%), dan *F1-score* (28.74%) masih rendah. Ini menandakan model bias ke kelas mayoritas dan kurang mampu mendeteksi kecelakaan berat secara akurat. Akurasi saja tidak cukup pada data tidak seimbang, sehingga diperlukan teknik *ADASYN* untuk penyeimbangan data dan *Forward Selection* untuk optimasi fitur agar model lebih adil dan akurat dalam memprediksi semua kelas.

3.3 Hasil Pemodelan *Random Forest* Setelah *ADASYN*

Pengujian awal dilakukan menggunakan model *Random Forest* tanpa seleksi fitur. Evaluasi dengan *10-fold cross-validation* menunjukkan akurasi rata-rata sebesar 84,26%, dengan *precision* dan *recall* di atas 84%.

Tabel 2. Hasil pengujian Model *RF*

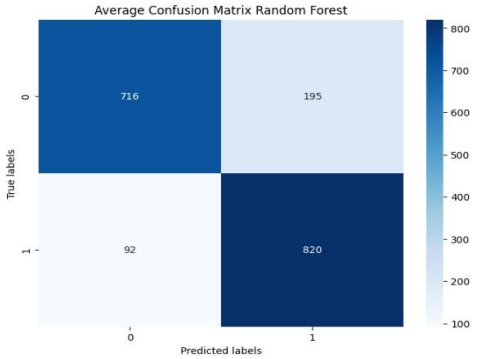
Fold	Akurasi	*Precision*	*Recall*	*F1 - Score*
1	88.52%	88.76%	88.52%	88.51%
2	81.42%	81.65%	81.42%	81.38%
3	85.25%	86.21%	85.25%	85.14%
4	85.71%	85.73%	85.71%	85.71%
5	77.47%	78.72%	77.47%	77.22%
6	86.81%	87.26%	86.81%	86.77%
7	86.26%	86.30%	86.26%	86.26%
8	86.26%	87.58%	86.26%	86.14%
9	82.97%	84.16%	82.97%	82.82%
10	81.87%	81.96%	81.87%	81.85%
Rata-rata	84.26%	84.83%	84.26%	84.18%

Pada table 2. Menunjukkan hasil evaluasi model *Random Forest* setelah penerapan *ADASYN*. Terlihat bahwa akurasi dan metrik lainnya seperti *precision, recall,* dan *F1-score* meningkat secara signifikan dibandingkan model sebelumnya, dengan nilai konsisten di atas 77% hingga 88%. Hal ini menunjukkan bahwa penyeimbangan data dengan *ADASYN* berhasil meningkatkan kemampuan model dalam mengklasifikasikan data secara lebih adil terhadap kedua kelas, termasuk kelas minoritas (kecelakaan berat).

Hasil ini menunjukkan performa yang cukup baik, meskipun masih terdapat bias terhadap kelas mayoritas.

Tabel 3. *Confusion Matrix*

	Predicted Positive (0)	*Predicted Negative (1)*
Actual Positive (0)	716	195
Actual Negative (1)	92	820



Gambar 6. *Confusion Matrix Random Forest*

Confusion matrix berperan penting dalam mengevaluasi performa model klasifikasi, khususnya dalam konteks data yang tidak seimbang seperti klasifikasi kecelakaan lalu lintas. Table 3 dan Gambar 6. menggambarkan hasil eksperimen menggunakan *Random Forest* tanpa penerapan teknik penyeimbangan (*ADASYN*) dan tanpa seleksi fitur (*Forward Selection*). Meskipun terdapat jumlah prediksi benar yang cukup tinggi (716 dan 820 untuk masing-masing kelas), jumlah kesalahan masih cukup besar (195 *false positive* dan 92 *false negative*), menunjukkan bahwa model masih bias terhadap kelas mayoritas dan belum mampu mengenali kelas minoritas dengan baik.

$$Accuracy = \frac{716 + 820}{716 + 820 + 195 + 92} = 0.8426 = 84\%$$

3.4 Hasil Pemodelan Kombinasi dengan *Forward Selection dan ADASYN*

Selanjutnya, dilakukan pemilihan fitur menggunakan metode *Forward Selection*. Fitur yang paling berkontribusi terhadap akurasi adalah cuaca, jumlah luka ringan, dan jumlah luka berat. Setelah model dilatih ulang hanya dengan fitur terpilih, akurasi meningkat signifikan menjadi 95.28%.

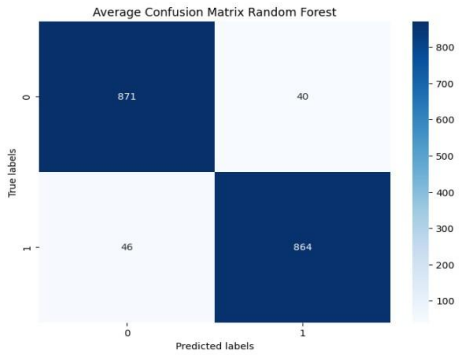
Tabel 4. Hasil Pengujian *RF* Dengan *Forward Selection*

Fold	Akurasi	*Precision*	*Recall*	*F1 - Score*
1	97.27%	97.27%	97.27%	97.27%
2	95.60%	95.69%	95.60%	95.60%
3	93.41%	93.41%	93.41%	93.41%
4	95.05%	95.06%	95.05%	95.05%
5	95.60%	95.60%	95.60%	95.60%
6	95.05%	95.10%	95.05%	95.05%
7	95.05%	95.19%	95.05%	95.05%
8	93.96%	94.00%	93.96%	93.95%
9	96.70%	96.73%	96.70%	96.70%
10	95.05%	95.06%	95.05%	95.05%
Rata-rata	95.28%	95.31%	95.28%	95.28%

Pada table 4. menunjukkan hasil evaluasi model *Random Forest* setelah penerapan metode seleksi fitur *Forward Selection* pada *10-fold cross-validation*. Secara konsisten, model menghasilkan akurasi, *precision*, *recall*, dan *f1-score* yang tinggi pada setiap *fold*, dengan rata-rata mencapai 95,28% untuk akurasi dan *recall*, serta 95,31% untuk precision. Hal ini menandakan bahwa pemilihan fitur yang relevan secara signifikan meningkatkan performa model dalam mengenali kedua kelas (ringan/sedang dan berat), serta menjaga keseimbangan antara prediksi positif yang tepat dan kelengkapan identifikasi kasus. Dengan kata lain, model menjadi lebih akurat, seimbang, dan andal dalam klasifikasi tingkat keparahan kecelakaan.

Table 5. *Confusion Matrix*

	Predicted P sitive (0)	*Predicted Negative (1)*
Actual Positive (0)	871	40
Actual Negative (1)	46	864



Gambar 7. *Confusion Matrix Random Forest*

Pada table 5 dan gambar 7. menunjukkan hasil eksperimen setelah penerapan *ADASYN* dan *Forward Selection*. Hasilnya jauh lebih baik: kesalahan klasifikasi berkurang signifikan dan prediksi benar meningkat (871 dan 864). Hal ini membuktikan bahwa *ADASYN* efektif menyeimbangkan distribusi kelas dengan cara mensintesis data minoritas, sehingga model dapat belajar lebih baik dari kedua kelas. Forward Selection juga berkontribusi dengan menyaring fitur-fitur yang paling relevan, meningkatkan fokus model terhadap variabel yang berpengaruh signifikan.

$$Accuracy = \frac{864 + 871}{864 + 871 + 40 + 46} = 0.9528 = 95\%$$

3.5 Evaluasi dan Perbandingan Model

Dibandingkan dengan model tanpa penyeimbangan dan seleksi fitur (79.26%), penerapan *ADASYN* saja meningkatkan akurasi menjadi 84.26%. Ketika *Forward Selection* diterapkan setelah *ADASYN*, akurasi melonjak ke 95.28%.

Tabel 6. Perbandingan Hasil Akurasi Pengujian *Random Forest*

Fold	*Random Forest*	*Setelah ADASYN*	*Kombinasi Forward Selection* dan *ADASYN*	Perbandingan Hasil (*RF* + Setelah *ADASYN)*	Perbandingan Hasil (RF + *ADASYN) + (RF + FS & ADASYN)*
1	78.69%	88.52%	97.27%	+9.83%	+8.75%
2	75.41%	81.42%	95.60%	+6.01%	+14.18%
3	78.57%	85.25%	93.41%	+6.68%	+8.16%
4	79.67%	85.71%	95.05%	+6.04%	+9.34%
5	81.87%	77.47%	95.60%	-4.04%	+18.13%
6	79.67%	86.81%	95.05%	+7.14%	+8.24%
7	81.87%	86.26%	95.05%	+4.39%	+8.79%
8	80.22%	86.26%	93.96%	+6.04%	+7.70%
9	77.47%	82.97%	96.70%	+5.50%	+13.73%
10	79.67%	81.87%	95.05%	+2.20%	+13.18%
Rata-rata	79.26%	84.26%	95.28%	+6.31%	+13.08%

Perbandingan pada table 6. ini mengindikasikan bahwa kombinasi kedua teknik tersebut secara signifikan meningkatkan kemampuan klasifikasi model. *Confusion matrix* menunjukkan bahwa model mampu mengklasifikasikan kelas minoritas (kecelakaan berat) dengan lebih akurat setelah penerapan strategi ini.

3.6 Diskusi

Berdasarkan Penelitian ini dilakukan untuk menjawab dua fokus utama yang telah dirumuskan sebelumnya, yaitu mengidentifikasi fitur- fitur yang paling mempengaruhi tingkat keparahan kecelakaan

lalu lintas di Kota Samarinda dan mengevaluasi efektivitas kombinasi algoritma *Random Forest* dengan metode *Forward Selection* dan *ADASYN* dalam meningkatkan kinerja klasifikasi pada data tidak seimbang.

a. Hasil penelitian menunjukkan bahwa fitur jumlah meninggal dunia, jumlah luka berat, jumlah luka ringan, dan cuaca merupakan variabel paling berpengaruh yang teridentifikasi melalui *Forward Selection* dan terbukti meningkatkan akurasi model secara signifikan. Temuan ini sejalan dengan penelitian [18] yang menyebutkan bahwa tipe tabrakan, kondisi cahaya, cuaca, dan fungsi jalan turut memengaruhi klasifikasi kecelakaan berat.
b. Model awal hanya menghasilkan akurasi 79.26%, *precision* 29.03%, *recall* 34.62%, dan *f1-score* 31.58%. Setelah penerapan *ADASYN*, terjadi peningkatan metrik evaluasi menjadi akurasi 84.26%, *precision* 81.82%, *recall* 84.62%, dan *f1-score* 83.20%. Peningkatan lebih besar tercapai setelah seleksi fitur *Forward Selection*, menghasilkan akurasi 95.28%, *precision* 94.23%, *recall* 96.15%, dan *f1-score* 95.18%. Penelitian lain [19] yang menggunakan *Random Forest* untuk klasifikasi jenis kecelakaan lalu lintas juga mencatat akurasi tinggi sebesar 92,93% (*10K-Fold CV*) dan *f1-score* 95%, dengan akurasi 100% pada kelas kecelakaan berat. Penelitian [20] membandingkan *Forward Selection* dan *Backward Elimination* pada *SVM* dan memperoleh akurasi 93% dan 97%. Sementara itu, studi [21] pada klasifikasi penyakit stroke menunjukkan bahwa *ADASYN* mampu meningkatkan *recall* dari 0% menjadi 78,3% dan *f1-score* dari 0% menjadi 21,81% meskipun akurasi menurun. Hasil-hasil ini menegaskan bahwa integrasi *ADASYN* dan *Forward Selection* efektif meningkatkan kinerja model, terutama dalam mengidentifikasi kasus kecelakaan berat pada data yang tidak seimbang.

4. KESIMPULAN

Hasil penelitian menunjukkan bahwa metode seleksi fitur menggunakan *Forward Selection* pada dataset kecelakaan lalu lintas berhasil mendapatkan fitur terpilih yaitu jumlah meninggal dunia, jumlah luka berat, jumlah luka ringan, dan cuaca.

Selain itu, hasil evaluasi model menunjukkan bahwa penerapan tiga metode sekaligus, yaitu *Random Forest*, *ADASYN*, dan *Forward Selection*, terbukti sangat efektif meningkatkan kinerja model. Awalnya, *Random Forest* tanpa penanganan ketidakseimbangan dan tanpa seleksi fitur hanya menghasilkan akurasi 79,26%, *precision* 29,03%, *recall* 34,62%, dan *f1-score* 31,58%. Setelah diterapkan *ADASYN*, metrik evaluasi meningkat signifikan menjadi akurasi 84,26%, *precision* 81,82%, *recall* 84,62%, dan *f1-score* 83,20%. Peningkatan lebih besar tercapai setelah seleksi fitur *Forward Selection*, menghasilkan akurasi akhir 95,28%, *precision* 94,23%, *recall* 96,15%, dan *f1-score* 95,18%. Hasil ini menegaskan bahwa kombinasi *ADASYN* dan *Forward Selection* dapat meningkatkan kinerja model, terutama dalam mendeteksi kasus kecelakaan berat yang sebelumnya sulit teridentifikasi akibat ketidakseimbangan data.

5. SARAN

Untuk mengatasi keterbatasan dalam penelitian ini, beberapa saran dapat diajukan bagi penelitian selanjutnya:

1. Penggunaan Metode Seleksi Fitur Alternatif Penelitian ini hanya menggunakan metode *Forward Selection* dalam proses pemilihan fitur. Studi lanjutan disarankan untuk mengevaluasi efektivitas metode lain seperti *Recursive Feature Elimination (RFE)*, *Mutual Information*, atau model berbasis pembobotan seperti *XGBoost* untuk memperoleh fitur yang lebih optimal.
2. Perbandingan Model dengan Algoritma Lain Penggunaan algoritma *Random Forest* dalam studi ini dapat dilengkapi dengan eksperimen menggunakan algoritma lain seperti *LightGBM*, *Gradient Boosting*, atau *Deep Neural Network* untuk menilai keandalan pendekatan yang digunakan serta meningkatkan performa klasifikasi.
3. Validasi dengan Dataset Eksternal Mengingat data yang digunakan hanya berasal dari Kota Samarinda, diperlukan validasi eksternal menggunakan dataset dari wilayah lain untuk menguji generalisasi model terhadap distribusi data yang berbeda.
4. Analisis Lanjutan terhadap Fitur Signifikan Beberapa fitur seperti cuaca dan jumlah korban menunjukkan kontribusi besar terhadap hasil klasifikasi. Penelitian berikutnya dapat melakukan analisis mendalam terhadap interaksi antar fitur ini untuk memperoleh pemahaman yang lebih komprehensif terkait faktor-faktor risiko kecelakaan.

DAFTAR PUSTAKA

[1] BPS, "Jumlah kecelakaan lalu lintas Menurut Bulan di Kota Samarinda, 2021-2023," Badan Pusat Statistika Kota Samarinda. Accessed: Mar. 20, 2025. [Online]. Available: https://samarindakota.bps.go.id/id/statistics-table/2/MjMyIzI=/jumlah-kecelakaan-lalu-lintas-menurut-bulan-di-kota-samarinda.html

[2] D. Rika Widianita, "Pertanggungjawaban Penyelenggara Jalan Atas Kerusakan Jalan Yang Mengakibatkan Kecelakaan Lalu Lintas," *AT-TAWASSUTH J. Ekon. Islam*, vol. VIII, no. I, pp. 1–19, 2023.

[3] M. Behzadi Goodari, H. Sharifi, P. Dehesh, M. A. Mosleh-Shirazi, and T. Dehesh, "Factors affecting the number of road traffic accidents in Kerman province, southeastern Iran (2015–2021)," *Sci. Rep.*, vol. 13, no. 1, pp. 1–9, 2023, doi: 10.1038/s41598-023-33571-8.

[4] B. Wang, C. Zhang, Y. D. Wong, L. Hou, M. Zhang, and Y. Xiang, "Comparing Resampling Algorithms and Classifiers for Modeling Traffic Risk Prediction," *Int. J. Environ. Res. Public Health*, vol. 19, no. 20, 2022, doi: 10.3390/ijerph192013693.

[5] S. M. Mostafa, S. A. Salem, and S. M. Habashyis, "Predictive Model for Accident Severity," *IAENG Int. J. Comput. Sci.*, vol. 49, no. 1, pp. 110–124, 2022.

[6] S. Khan *et al.*, "Comprehensive risk assessment of Pakistan railway network: A semi-quantitative risk matrix approach," *Heliyon*, vol. 10, no. 12, p. e32682, 2024, doi: 10.1016/j.heliyon.2024.e32682.

[7] M. K. Islam, I. Reza, U. Gazder, R. Akter, M. Arifuzzaman, and M. M. Rahman, "Predicting Road Crash Severity Using Classifier Models and Crash Hotspots," *Appl. Sci.*, vol. 12, no. 22, 2022, doi: 10.3390/app122211354.

[8] Y. Priantama and T. A. Yoga Siswa, "Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, p. 251, 2022, doi: 10.26798/jiko.v6i2.651.

[9] N. Behboudi, S. Moosavi, and R. Ramnath, "Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques," 2024, [Online]. Available: http://arxiv.org/abs/2406.13968

[10] S. I. Mohammadpour, M. Khedmati, and M. J. H. Zada, "Classification of truck-involved crash severity: Dealing with missing, imbalanced, and high dimensional safety data," *PLoS One*, vol. 18, no. 3 March, pp. 1–22, 2023, doi: 10.1371/journal.pone.0281901.

[11] Ali CELIK dan Onur SEVLI, "Predicting Traffic Accident Severity Using Machine Learning Techniques," *Türk Doğa ve Fen Derg.*, vol. 11, no. 3, pp. 79–83, 2022, doi: 10.46810/tdfd.1136432.

[12] A. V. Vitianingsih, Z. Othman, S. S. K. Baharin, A. Suraji, and A. L. Maukar, "Application of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 5, pp. 676–690, 2022, doi: 10.22266/ijies2022.1031.58.

[13] H. Park *et al.*, "Traffic Crash Countermeasure Recommendations Using Deep Neural Network: A Decision Support Tool for Traffic Safety Engineers," *IEEE Access*, vol. 13, no. April, pp. 53718–53730, 2025, doi: 10.1109/ACCESS.2025.3543641.

[14] N. Sakib, T. Paul, N. Anwari, and M. Hadiuzzaman, "Ensemble-based model to investigate factors influencing road crash fatality for imbalanced data," *Transp. Eng.*, vol. 18, no. October, p. 100284, 2024, doi: 10.1016/j.treng.2024.100284.

[15] N. Izzah and W. H. Chan, "Identification of Important Risk Factors of Diabetes and Heart Disease using Feature Selection on Classification Algorithms," pp. 125–132, 2024.

[16] K. D. Prasetya, "Multitemporal Spatial Analysis for Monitoring and Classification of Coal Mining and Reclamation Using Satellite Imagery," pp. 1–23, 2025.

[17] A. Hadi, M. Qamal, and Y. Afrillia, "Comparison of Random Forest Algorithm Classifier and Naïve Bayes Algorithm in Whatsapp Message Type Classification," pp. 9–17, 2025.

[18] B. L. Fauzan, T. Agustin, and A. M. H. Mahmudah, "Prediksi Klasifikasi Kecelakaan Lalu Lintas di Kota Surakarta dengan Menggunakan Metode Regresi Logistik Multinomial," *Sustain. Civ. Build. Manag. Eng.*, vol. 1, no. 4, p. 9, 2024, doi:

10.47134/scbmej.v1i4.3159.
[19] N. Betris *et al.*, “Identifikasi jenis kecelakaan lalu lintas di kota samarinda menggunakan metode random forest,” vol. 9, no. 2, pp. 3153–3160, 2025.
[20] N. I. Y. Wandayana Nur’Amanah Suharmin, Isran K. Hasan, “The Indonesian Journal of Computer Science,” vol. 14, no. 2, pp. 3374–3384, 2025.
[21] S. K. Gusti, I. Afrianty, and F. Syafria, “BULLETIN OF COMPUTER SCIENCE RESEARCH Penerapan Metode ADASYN Dalam Mengatasi Imbalanced Data Untuk Klasifikasi Penyakit Stroke Menggunakan Support Vector Machine,” vol. 5, no. 4, pp. 532–541, 2025, doi: 10.47065/bulletincsr.v5i4.612.