KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research
Vol.3, No.6 Desember 2025
e-ISSN: 3025-9851; p-ISSN: 3025-986X, Hal 416-422
DOI : https://doi.org/10.61722/jssr.v3i6.6433

KLASIFIKASI TINGKAT KEPARAHAN KECELAKAAN KERJA MENGGUNAKAN ALGORITMA RANDOM FOREST

Izzah Afkarinah
Teknik Informatika, Universitas Yudharta Pasuruan
Arif Faizin
Teknik Informatika, Universitas Yudharta Pasuruan
Ahmad Zulham Fahamsyah Havy
Teknik Informatika, Universitas Yudharta Pasuruan
Alamat: Jl. Yudharta No. 07 (Pesantren Ngalah)
Korespondensi penulis: izzahafkarinah23@gmail.com[1], arifusan@yudharta.ac.id[2], zulham92@yudharta.ac.id[3]

Abstrak. *This study aims to develop a classification model for the severity of workplace accidents using the Random Forest algorithm with OSHA–ITA data from January 2015 to September 2024. From 95,194 initial incident data, cleaning, missing value handling, removal of irrelevant variables, encoding, and the formation of Severity targets (Severe, Moderate, Mild) were carried out. The dataset was divided into 70% training data and 30% testing data with a total of 28,559 samples. The results showed that the initial model was biased towards the majority class, while the minority class was difficult to recognize. After applying SMOTE, the model's performance improved with more balanced predictions. The most influential features included Nature Title, Part of Body Title, Event Title, Source Title, Hospitalized, and Amputation. These findings emphasize the importance of selecting relevant features and data balancing techniques to improve the performance of Random Forest classification in occupational accidents.*
Keywords: *Random Forest, SMOTE, classification, workplace accidents, OSHA ITA.*

Abstrak. Penelitian ini bertujuan membangun model klasifikasi tingkat keparahan kecelakaan kerja menggunakan algoritma Random Forest dengan data OSHA–ITA periode Januari 2015–September 2024. Dari 95.194 data insiden awal, dilakukan pembersihan, penanganan *missing values*, penghapusan variabel tidak relevan, encoding, serta pembentukan target *Severity* (Berat, Sedang, Ringan). Dataset dibagi menjadi 70% data training dan 30% data testing dengan total 28.559 sampel. Hasil menunjukkan bahwa model awal bias terhadap kelas mayoritas, sementara kelas minoritas sulit dikenali. Setelah penerapan SMOTE, performa model meningkat dengan prediksi lebih seimbang. Fitur paling berpengaruh meliputi *Nature Title*, *Part of Body Title*, *Event Title*, *Source Title*, *Hospitalized*, dan *Amputation*. Temuan ini menegaskan pentingnya pemilihan fitur relevan dan teknik penyeimbangan data untuk meningkatkan kinerja klasifikasi Random Forest pada kecelakaan kerja.
Kata Kunci: *Random Forest, SMOTE, klasifikasi, kecelakaan kerja, OSHA ITA*

PENDAHULUAN

Keselamatan dan kesehatan kerja (K3) merupakan aspek krusial dalam dunia industri yang bertujuan untuk melindungi tenaga kerja dari risiko kecelakaan dan penyakit akibat kerja. Meskipun berbagai regulasi dan sistem manajemen K3 telah diterapkan, kecelakaan kerja tetap menjadi isu serius yang berdampak pada produktivitas, biaya perusahaan, dan kesejahteraan pekerja. Menurut UU No. 1 Tahun 1970 tentang Keselamatan Kerja, perusahaan wajib menciptakan lingkungan kerja yang aman.

Received Oktober 20, 2025; Revised November 30, 2025;Desember 01, 2025
** Izzah Afkarinah, izzahafkarinah23@gmail.com*

Setiap kecelakaan memiliki tingkat keparahan yang berbeda, mulai dari ringan hingga fatal. Klasifikasi tingkat keparahan ini penting untuk evaluasi risiko dan pengambilan keputusan preventife. Namun, proses klasifikasi sering dilakukan secara manual atau subjektif, yang dapat menyebabkan kesalahan analisis.

Kecelakaan kerja adalah suatu kejadian tidak terduga dan tidak direncanakan yang mengganggu proses kerja dan dapat menyebabkan cedera atau kerugian baik fisik, mental, maupun materil. Menurut UU No. 1 Tahun 1970 tentang Keselamatan Kerja, kecelakaan kerja merupakan kejadian yang terjadi dalam hubungan kerja, yang mengakibatkan luka, penyakit, cacat, bahkan kematian. Tingkat keparahan kecelakaan kerja diklasifikasikan ke dalam beberapa kategori, seperti ringan, sedang, dan berat, yang menunjukkan dampak kecelakaan terhadap pekerja.

Dengan kemajuan teknologi, khususnya dalam bidang machine learning, kini tersedia metode yang lebih objektif dan akurat untuk klasifikasi kecelakaan. Salah satu algoritma yang efektif untuk klasifikasi adalah Random Forest, yang mampu menangani data besar dan variabel kompleks.

Algoritma Random Forest merupakan algoritma klasifikasi dan regresi berbasis ensemble learning yang membentuk kumpulan (forest) dari beberapa pohon keputusan (decision tree). Setiap pohon menghasilkan prediksi dan hasil akhir ditentukan berdasarkan voting mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi). Keunggulan utama Random Forest adalah ketahanannya terhadap overfitting, kemampuannya menangani data kategorikal dan numerik, serta efektivitasnya dalam mengukur pentingnya fitur (feature importance). Algoritma ini juga dikenal mampu bekerja dengan baik pada dataset yang besar dan kompleks.

Dalam penelitian ini, penulis menggunakan dataset dari Occupational Safety and Health Administration (OSHA) – Injury Tracking Application (ITA) yang berisi ribuan kasus kecelakaan kerja di berbagai sektor industri selama periode Januari 2015 hingga September 2024. Dataset ini mencakup informasi seperti waktu kejadian, jenis cedera, bagian tubuh yang terluka, penyebab kecelakaan, dan tingkat keparahan. Dengan menggunakan algoritma Random Forest, penelitian ini bertujuan untuk membangun model klasifikasi otomatis yang dapat memprediksi tingkat keparahan kecelakaan kerja berdasarkan fiturfitur tertentu. Hasil penelitian diharapkan dapat memberikan kontribusi pada sistem monitoring K3 secara lebih efisien dan berbasis data.

METODE PENELITIAN

Penelitian ini merupakan penelitian terapan (applied research) dengan pendekatan kuantitatif dan menggunakan metode eksperimen komputasional. Tujuan dari penelitian ini adalah untuk membangun model klasifikasi otomatis menggunakan algoritma Random Forest guna memprediksi tingkat keparahan kecelakaan kerja berdasarkan data historis yang telah tersedia. Penelitian ini memanfaatkan pendekatan machine learning berbasis supervised learning, di mana data yang digunakan telah memiliki label tingkat keparahan kecelakaan. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi praktis dalam mendukung upaya

keselamatan kerja melalui pemanfaatan teknologi analitik berbasis data. Dalam skala pengukuran, peneliti mengkalisifikasikan keparahan kecekalaan kerja sebagai berikut:

0 = Ringan
1 = Sedang
2 = Berat

Penelitian ini diawali dengan identifikasi masalah terkait kesulitan klasifikasi tingkat keparahan kecelakaan kerja akibat banyaknya variabel yang terlibat. Data yang digunakan berasal dari Workplace Accident Severity Dataset (OSHA–ITA) periode Januari 2015–September 2024 dengan variabel independen meliputi Amputation, Hospitalized, NatureTitle, Part of Body Title, EventTitle, dan SourceTitle, serta target Severity yang dikategorikan menjadi tiga kelas: Ringan (0), Sedang (1), dan Berat (2). Tahap pra-pemrosesan meliputi pembersihan missing values, encoding variabel kategorikal menggunakan OrdinalEncoder, dan penanganan ketidakseimbangan kelas dengan SMOTE yang hanya diterapkan pada data pelatihan. Data kemudian dibagi menjadi data latih dan uji dengan perbandingan 70:30 menggunakan stratifikasi distribusi kelas. Model dibangun menggunakan Random Forest Classifier dengan optimasi hyperparameter melalui GridSearchCV, sedangkan evaluasi dilakukan menggunakan confusion matrix serta analisis feature importance. Penelitian ini menggunakan perangkat keras berupa laptop berprosesor Intel Core i7 dengan RAM 8 GB dan SSD 256 GB, serta perangkat lunak Python 3.10 dengan pustaka pandas, scikit-learn, matplotlib, dan seaborn.

HASIL PENELITIAN DAN PEMBAHASAN

Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari Occupational Safety and Health Administration (OSHA) – Injury Tracking Application (ITA) dengan periode Januari 2015 hingga September 2024. Total terdapat 95.194 data insiden dengan 27 variabel yang mencakup data numerik, kategorikal, dan teks. Dataset ini berisi informasi mengenai identitas insiden, lokasi, jenis cedera, bagian tubuh terdampak, jenis kejadian, hingga status medis pekerja.

Tabel 1. Deskripsi variabel pada dataset penelitian

Nama Kolom	Tipe Data	Keterangan
ID	*int64*	Nomor identitas unik setiap insiden kecelakaan.
UPA	*int64*	Kode identifikasi unit/area kerja.
EventDate	*Object*	Tanggal terjadinya insiden kecelakaan.
Employer	*Object*	Nama perusahaan atau instansi tempat kejadian.
Address1	*Object*	Alamat utama lokasi kejadian.
Address2	*Object*	Alamat tambahan lokasi kejadian.
City	*Object*	Kota tempat kejadian.
State	*Object*	Negara bagian atau provinsi tempat kejadian.
Zip	*float64*	Kode pos lokasi kejadian.
Latitude	*float64*	Koordinat lintang lokasi kejadian.
Longitude	*float64*	Koordinat bujur lokasi kejadian.
Primary NAICS	*Object*	Kode klasifikasi industri menurut *NAICS.*

Hospitalized	*float64*	Status rawat inap (1 = ya, 0 = tidak).
Amputation	*float64*	Status amputasi (1 = ya, 0 = tidak).
Inspection	*float64*	Nomor identifikasi inspeksi terkait insiden.
Final Narrative	*Object*	Deskripsi singkat kronologi kejadian.
Nature	*int64*	Kode numerik jenis cedera.
NatureTitle	*Object*	Nama/deskripsi jenis cedera.
Part of Body	*int64*	Kode numerik bagian tubuh yang terkena.
Part of Body Title	*Object*	Nama/deskripsi bagian tubuh yang terkena.
Event	*int64*	Kode numerik jenis kejadian.
EventTitle	*Object*	Nama/deskripsi jenis kejadian.
Source	*int64*	Kode numerik sumber penyebab kecelakaan.
SourceTitle	*Object*	Nama/deskripsi sumber penyebab kecelakaan.
Secondary Source	*float64*	Kode numerik sumber sekunder penyebab kecelakaan.
Secondary Source Title	*Object*	Nama/deskripsi sumber sekunder penyebab kecelakaan.
FederalState	*int64*	Kode status federal atau negara bagian.

Permasalahan Kualitas Data

Hasil inspeksi awal menemukan beberapa permasalahan kualitas data, yaitu: (1) adanya nilai kosong pada variabel seperti *Latitude*, *Longitude*, dan *Secondary Source Title*; (2) ketidakkonsistenan format penulisan pada variabel kategorikal; dan (3) distribusi kelas yang tidak seimbang, di mana kelas Berat mendominasi dibandingkan kelas Sedang dan Ringan. Permasalahan ini berpotensi menimbulkan bias pada model jika tidak ditangani.

Tabel 2. Permasalahan kualitas data

Permasalahan	**Dampak**	**Solusi**
Nilai kosong pada variabel *Latitude, Longitude*, dan *Secondary Source Title*	Dapat mengganggu proses analisis dan mengurangi kualitas prediksi	*Data cleaning* dengan menghapus kolom tidak relevan atau melakukan imputasi nilai yang hilang
Perbedaan format penulisan pada variabel kategorikal	Menimbulkan penggandaan kategori secara tidak sengaja	Normalisasi format penulisan (penyeragaman huruf besar/kecil)
Ketidakseimbangan jumlah data antar kelas	Model cenderung bias terhadap kelas mayoritas	Penyeimbangan data menggunakan metode *SMOTE* (*Synthetic Minority Over-sampling Technique*)

Proses Preprocessing Data

Tahapan preprocessing dilakukan dengan:

1. Cleansing Data, menghapus variabel administratif, kolom dengan banyak *missing values*, serta variabel redundan (kode numerik yang tumpang tindih dengan deskripsi).
2. Penanganan Missing Values, menggunakan imputasi sederhana, seperti mengisi nilai 0 untuk data biner (*Hospitalized* dan *Amputation*) dan label "unknown" untuk variabel kategorikal.
3. Pembuatan Kolom Target (Severity), mengklasifikasikan tingkat keparahan menjadi tiga kategori: Berat, Sedang, dan Ringan berdasarkan kombinasi enam variabel utama (*Hospitalized, Amputation, Nature Title, Part of Body Title, Event Title, Source Title*).
4. Encoding Variabel Kategorikal, mengubah data kategorikal menjadi representasi numerik agar dapat diproses oleh algoritma machine learning.

Pembagian Data Training dan Testing

Dataset hasil preprocessing dibagi menjadi data training (70%) dan data testing (30%) menggunakan *train_test_split* dengan parameter *stratify*. Dari total 28.559 sampel, diperoleh 19.991 data training dan 8.568 data testing. Proses stratifikasi menjaga proporsi distribusi kelas agar tetap konsisten antara data training dan testing. Namun, hasil distribusi menunjukkan adanya class imbalance yang signifikan, dengan kelas Berat jauh lebih dominan dibanding kelas Sedang dan Ringan.

Penyeimbangan Data dengan SMOTE

Untuk mengatasi ketidakseimbangan kelas, diterapkan Synthetic Minority Over-sampling Technique (SMOTE) pada data training. SMOTE menghasilkan data sintetis untuk kelas minoritas sehingga distribusi kelas menjadi lebih seimbang. Penerapan hanya dilakukan pada data training guna menghindari *data leakage*, sedangkan data testing tetap mempertahankan distribusi asli untuk evaluasi objektif.

Evaluasi Model

Evaluasi model dilakukan menggunakan algoritma Random Forest dengan dua skenario: sebelum dan sesudah penerapan SMOTE.

1. Evaluasi Sebelum SMOTE

Model cenderung hanya mengenali kelas mayoritas (Berat) dengan akurasi tinggi (69,87%), namun gagal mengklasifikasikan kelas minoritas (Ringan dan Sedang) dengan baik (recall = 0). Hal ini menegaskan dampak negatif dari ketidakseimbangan kelas.

2. Evaluasi Sesudah SMOTE

Setelah SMOTE, performa model meningkat signifikan pada kelas minoritas. Akurasi keseluruhan mencapai 99,78%, dengan recall kelas Ringan dan Sedang masing-masing meningkat menjadi 91,5% dan 92,1%. Nilai F1-score juga menunjukkan peningkatan yang signifikan, menandakan keseimbangan antara presisi dan recall pada semua kelas.

Analisis Feature Importance

Hasil analisis *feature importance* menunjukkan bahwa variabel Amputation (0,32) dan Hospitalized (0,28) memiliki kontribusi terbesar dalam klasifikasi tingkat keparahan. Faktor medis ini menjadi penentu utama kategori kecelakaan kerja berat. Variabel lain seperti Nature

Title (0,16), Part of Body Title (0,12), Event Title (0,07), dan Source Title (0,05) tetap berperan sebagai pendukung untuk memperkaya interpretasi model.

Pembahasan

Hasil penelitian membuktikan bahwa penerapan SMOTE mampu meningkatkan kinerja model dalam mengenali kelas minoritas yang sebelumnya terabaikan. Faktor medis terbukti lebih dominan dibandingkan faktor kontekstual dalam menentukan tingkat keparahan kecelakaan kerja. Temuan ini sejalan dengan teori kesehatan kerja, di mana amputasi dan rawat inap merupakan indikator utama keparahan. Dengan demikian, model tidak hanya efektif sebagai alat prediksi, tetapi juga dapat digunakan sebagai sistem pendukung keputusan dalam upaya pencegahan kecelakaan kerja di industri.

KESIMPULAN

Simpulan

Berdasarkan hasil penelitian dan pembahasan pada Bab IV, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Model klasifikasi tingkat keparahan kecelakaan kerja berhasil dibangun menggunakan algoritma Random Forest dengan dataset OSHA–ITA (2015–2024). Penerapan SMOTE terbukti meningkatkan kinerja model dalam mengenali kelas minoritas (Ringan dan Sedang), dengan recall masing-masing 91,5% dan 92,1%, meskipun akurasi global sedikit menurun. Hal ini menegaskan pentingnya teknik penyeimbangan data untuk menghasilkan model yang lebih representatif.
2. Analisis feature importance menunjukkan bahwa variabel paling berpengaruh terhadap klasifikasi adalah Amputation (indikator utama kecelakaan berat), diikuti oleh Hospitalized, serta variabel cedera dan konteks lain seperti Nature Title, Part of Body Title, Event Title, dan Source Title.

Saran

Berdasarkan hasil penelitian ini, beberapa saran yang dapat diberikan untuk pengembangan lebih lanjut adalah:

1. Penelitian selanjutnya dapat mencoba algoritma lain seperti XGBoost, LightGBM, atau Deep Learning untuk dibandingkan kinerjanya dengan Random Forest.
2. Gunakan teknik penyeimbangan data lain selain SMOTE, misalnya ADASYN atau Cost-Sensitive Learning.
3. Manfaatkan variabel teks melalui Natural Language Processing (NLP) agar model lebih kaya informasi.
4. Lakukan validasi dengan data kecelakaan nyata untuk menguji generalisasi model.
5. Integrasikan model ke dalam Decision Support System (DSS) bidang K3 sebagai dukungan analisis risiko dan pencegahan kecelakaan.

DAFTAR PUSTAKA

S. Chauhan and S. Sachdeva, "A machine learning approach for classification of occupational injury severity," Procedia Computer Science, vol. 167, pp. 2191–2200, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.03.270

G. López and M. Paz, "*Analysis and prediction of workplace accidents using machine learning: A Random Forest approach," Safety Science,* vol. 110, pp. 57–65, 2018. [Online].Available:https://doi.org/10.1016/j.ssci.2018.07.002

Y. H. Huang et al., *"Developing a classification model for construction accident severity using data mining techniques," Journal of Safety Research,* vol. 62, pp. 203–210, 2017. [Online]. Available: https://doi.org/10.1016/j.jsr.2017.06.007

E. Petridou et al., "Machine learning techniques for classification of work-related accidents: An empirical study in the construction sector," Automation in Construction, vol. 98, pp. 321–330, 2019. [Online]. Available: https://doi.org/10.1016/j.autcon.2018.11.018

Y. Kim and S. Park, "*Predicting industrial accident severity with Random Forest and imbalanced data techniques*," Expert *Systems with Applications*, vol. 158, p. 113584, 2020. *[Online].Available:*https://doi.org/10.1016/j.eswa.2020.113584

Y. Guo, T. W. Yiu, and V. A. González, "*Predicting accident severity in the construction industry using artificial neural networks," Automation in Construction, vol. 84, pp. 1–9, 2017.[Online].Available:*https://doi.org/10.1016/j.autcon.2017.08.009

A. J. P. Tixier et al., *"Application of machine learning to construction injury prediction," Automation in Construction, vol. 69, pp. 102–114, 2016. [Online]. Available:* https://doi.org/10.1016/j.autcon.2016.05.016

A. Dutta dan A. P. S. Rathore, "*Estimating Ergonomic Compatibility of Cars: A Fuzzy Approach," Procedia Computer Science, vol. 167, hlm. 506–515, 2020,* https://doi: 10.1016/j.procs.2020.03.270.

M. Marzouk dan M. Enaba, "*Text analytics to analyze and monitor construction project contract and correspondence," Automation in Construction, vol. 98, hlm. 265–274, Feb 2019,* https://doi: 10.1016/j.autcon.2018.11.018.

Y.-F. Huang dan P.-H. Chen, *"Fake news detection using an ensemble learning model based on Self-Adaptive Harmony Search algorithms," Expert Systems with Applications, vol. 159, hlm. 113584, Nov 2020.* https://doi: 10.1016/j.eswa.2020.113584.

W. Zhou, "*Traffic Accident Severity Prediction Based on Data Cleaning and Machine Learning," IEEE Access, vol. 12, 2024. [Online].Available:*https://doi.org/10.1109/ACCESS.2024.3333345

A. Khairuddin et al., "*Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization," International Journal of Occupational Safety and Ergonomics, vol. 28, no. 3, pp. 321–335, 2022. [Online]. Available:* https://doi.org/10.1080/10803548.2022.2057755

A. Adefabi, O. Ayodeji, *and* M. A. Abdulraheem, *"Predicting Accident Severity Using Random Forest Model," Safety Science, vol. 163, 2023. [Online]. Available:* https://doi.org/10.1016/j.ssci.2023.106038

K. Koç, Ö. Ekmekcioğlu, *and* A. P. Gurgun, "*Prediction of Construction Accident Outcomes Based on an Imbalanced Dataset," Journal of Construction Engineering and Management, vol. 149, no. 2, 2023. [Online]. Available:* https://doi.org/10.1061/jcem.0002344

Zermane, A., M. Z. M. Tohir, H. Zermane, M. R. Baharudin, *and* H. M. Yusoff, "*Predicting Fatal Fall from Heights Accidents Using Random Forest Classification," Journal of Occupational Health, vol. 64, no. 3, 2022. [Online]. Available:* https://doi.org/10.1002/1348-9585.12367