JSiI | Jurnal Sistem Informasi
Vol. 13 No. 1 | March 2026, P. 46-52

p-ISSN: 2406-7768
e-ISSN: 2581-2181

EVALUASI KLASIFIKASI HASIL CATUR BLITZ PADA DATASET TIDAK SEIMBANG SKALA BESAR MENGGUNAKAN COST-SENSITIVE LEARNING

Khairuddin[1]

[1] *Jurusan Sistem Informasi Fakultas Teknik dan Komputer Universitas Putera Batam, Indonesia*
Jln. R. Soeprapto Muka Kuning, Kibing, Batu Aji, Kota Batam, Kepulauan Riau
[1]pb221510062@upbatam.ac.id

Abstract

The rapid growth of online chess platforms has generated large-scale structured game data that can be utilized for data-driven analysis. In blitz mode games, match outcomes are categorized into win, lose, and draw; however, the distribution of these outcomes is inherently imbalanced, with draw representing a small minority of the dataset. This study aims to evaluate the effectiveness of cost-sensitive learning through balanced class weighting in improving classification performance on an imbalanced large-scale blitz chess dataset. A total of 100,000 rated blitz games were extracted from the Lichess open database and processed through preprocessing, feature extraction, and stratified data splitting. Three supervised learning algorithms - Support Vector Machine (SVM), Decision Tree, and Random Forest - were implemented. Model performance was evaluated using Macro F1-score as the primary metric, along with accuracy and 5-fold stratified cross-validation. The results indicate that without cost-sensitive learning, the recall for the minority class (draw) approaches zero despite achieving higher overall accuracy (0.54). In contrast, applying balanced class weighting significantly improves minority class detection, increasing recall for draw up to 0.73 with a Macro F1-score of approximately 0.40, although overall accuracy decreases to 0.45. This demonstrates the trade-off between global performance and minority class sensitivity. Feature importance analysis further reveals that move count is the most influential predictor of match outcomes. These findings confirm that imbalance-aware learning plays a critical role in large-scale chess outcome classification and highlight the importance of appropriate evaluation metrics in imbalanced datasets.

Keywords: Chess outcome classification, Imbalanced classification, Cost-sensitive learning, Support Vector Machine, Macro F1-score, Large-scale chess dataset

I. PENDAHULUAN

Perkembangan platform permainan daring telah menghasilkan data dalam jumlah besar yang berpotensi dimanfaatkan untuk analisis berbasis data [1]. Salah satu contoh yang menonjol adalah catur daring, yang menunjukkan partisipasi masif, dengan lebih dari 123 juta pertandingan yang dimainkan oleh hampir satu juta pemain hanya pada satu platform online. Setiap pertandingan catur online terdokumentasi dalam bentuk data digital yang mencakup informasi hasil pertandingan, rating pemain, jenis kontrol waktu, serta urutan langkah permainan [2]. Karakteristik data tersebut membuka peluang penerapan teknik machine learning untuk mengekstraksi pola dan informasi yang bernilai [3].

Dalam klasifikasi hasil pertandingan catur blitz, keluaran umumnya dibagi ke dalam tiga kelas, yaitu win, lose, dan draw. Namun, berdasarkan eksplorasi dataset pertandingan blitz, distribusi kelas sering kali tidak seimbang, di mana kelas draw memiliki proporsi yang jauh lebih kecil dibandingkan kelas win dan lose [4]. Ketidakseimbangan distribusi kelas (*class imbalance*) merupakan permasalahan umum dalam pembelajaran mesin, di mana jumlah sampel antar kelas tidak proporsional. Kondisi ini menyebabkan model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas, terutama pada kasus kejadian langka, sehingga menurunkan kemampuan deteksi terhadap kelas minoritas [5]. Pada konteks data tidak seimbang, penggunaan akurasi sebagai satu-satunya metrik evaluasi sering kali menyesatkan [6], karena model dapat mencapai nilai akurasi tinggi dengan hanya memprediksi kelas mayoritas, tanpa mampu mengidentifikasi kelas minoritas secara memadai [7].

Penelitian sebelumnya telah mengeksplorasi klasifikasi hasil pertandingan catur menggunakan berbagai pendekatan machine learning [4]. Namun, analisis eksplisit terhadap dampak ketidakseimbangan kelas serta evaluasi berbasis Macro F1-score pada dataset blitz berskala besar masih terbatas [8]. Selain itu, kajian mengenai efektivitas pendekatan cost-sensitive learning telah banyak digunakan untuk menangani permasalahan data tidak seimbang [9], namun penerapannya dalam konteks klasifikasi hasil catur blitz belum banyak

Received: Feb 19th, 2026 | Published: March 30th, 2026

p-ISSN: 2406-7768
e-ISSN: 2581-2181

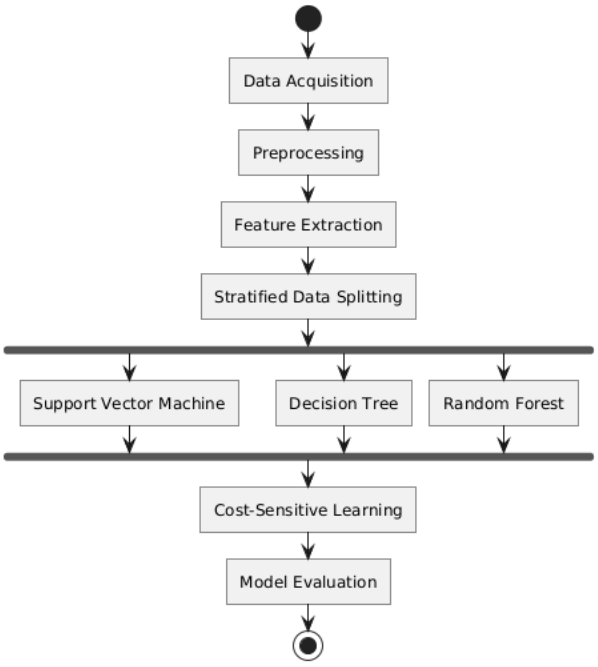
dibahas secara sistematis, sehingga diperlukan evaluasi yang lebih komprehensif. Oleh karena itu, diperlukan kajian yang tidak hanya berfokus pada peningkatan akurasi, tetapi juga mempertimbangkan keseimbangan performa antar kelas pada kondisi distribusi data yang tidak seimbang.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengevaluasi performa klasifikasi hasil catur blitz pada dataset tidak seimbang skala besar dengan menerapkan pendekatan cost-sensitive learning. Dataset yang digunakan terdiri dari 100.000 partai blitz rated yang diperoleh dari database terbuka Lichess. Penelitian ini mengimplementasikan algoritma Support Vector Machine, Decision Tree, dan Random Forest, serta membandingkan performa model dengan dan tanpa penyesuaian bobot kelas untuk menganalisis dampak penanganan ketidakseimbangan kelas terhadap deteksi kelas minoritas.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental berbasis komputasi untuk mengevaluasi performa model klasifikasi hasil pertandingan catur blitz pada dataset tidak seimbang dalam skala besar. Seluruh tahapan penelitian dirancang agar dapat direplikasi dan dianalisis secara objektif.

Secara umum, tahapan penelitian meliputi pengumpulan data, pra-pemrosesan, ekstraksi fitur, pembagian data, penerapan algoritma klasifikasi, penanganan ketidakseimbangan kelas, serta evaluasi performa model.



Gambar 1. Alur Penelitian

A. Dataset

Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Lichess Open Database dalam format Portable Game Notation (PGN) terkompresi (.pgn.zst). Data yang digunakan adalah partai rated dengan mode permainan blitz dan dapat diakses melalui laman resmi Lichess https://database.lichess.org/.

Sebanyak 100.000 partai diproses melalui tahap penyaringan berdasarkan atribut *Event* dan *TimeControl* untuk memastikan kesesuaian dengan karakteristik permainan blitz. Data kemudian diekstraksi menjadi format terstruktur yang memuat informasi rating pemain, kontrol waktu, jumlah langkah, serta hasil pertandingan.

Dataset dibagi menggunakan teknik stratified sampling untuk mempertahankan proporsi distribusi kelas pada data latih dan data uji [10].

Tabel 1. Karakteristik Umum Dataset

Karakteristik	Keterangan
Sumber Data	Lichess Open Database
Jenis Permainan	Catur Online
Mode Permainan	Rated Blitz
Periode Data	Desember 2025
Format Data	PGN (.pgn.zst)
Total Data	100.000 pertandingan
Data Latih	70.000 (70%)
Data Uji	30.000 (30%)
Jumlah Kelas	3 (win, lose, draw)
Distribusi Win	49.742 (49,74%)
Distribusi Lose	46.157 (46,16%)
Distribusi Draw	4.101 (4,10%)
Fitur yang Digunakan	Rating Difference, Move Count, Base Time
Tipe Fitur	Numerik
Karakteristik Dataset	Tidak seimbang (kelas draw sebagai minoritas)

Distribusi kelas pada dataset menunjukkan kondisi tidak seimbang, di mana kelas draw hanya merepresentasikan 4,10% dari total 100.000 partai, sedangkan kelas win dan lose masing-masing mencapai 49,74% dan 46,16%. Ketimpangan ini mencerminkan karakteristik alami pertandingan blitz yang cenderung menghasilkan kemenangan atau kekalahan dibandingkan hasil remis. Kondisi distribusi tersebut berimplikasi pada perlunya perhatian khusus dalam strategi pemodelan agar kelas minoritas tetap terakomodasi secara proporsional.

B. Pra-Pemrosesan Data

Tahap pra-pemrosesan bertujuan untuk mengubah data mentah dalam format PGN terkompresi (.pgn.zst) menjadi dataset terstruktur yang siap digunakan dalam proses analisis. Proses ini dilakukan melalui beberapa langkah sistematis sebagai berikut.

Received: Feb 19th, 2026 | Published: March 30th, 2026

p-ISSN: 2406-7768
e-ISSN: 2581-2181

1) *Parsing File PGN:* Dataset Lichess disimpan dalam format PGN terkompresi menggunakan algoritma Zstandard. Untuk menghindari penggunaan memori yang berlebihan, file dibaca secara streaming menggunakan pustaka zstandard dan python-chess. Setiap partai diparsing untuk mengekstraksi informasi header yang relevan, antaranya WhiteElo, BlackElo, Result, TimeControl dan Event. Parsing dilakukan secara iteratif hingga diperoleh 100.000 partai yang memenuhi kriteria penelitian.
2) *Penyaringan Mode Permainan:* Tidak seluruh partai dalam database digunakan. Penyaringan dilakukan dengan memilih hanya partai yang memiliki indikator "Blitz" pada atribut *Event*. Langkah ini bertujuan untuk memastikan bahwa dataset benar-benar merepresentasikan pertandingan rated mode blitz.
3) *Transformasi Label:* Atribut *Result* pada file PGN memiliki tiga kemungkinan nilai, yaitu "1-0", "0-1", dan "1/2-1/2". Nilai-nilai tersebut kemudian ditransformasikan menjadi tiga kelas kategorikal, yaitu *win*, *lose*, dan *draw*. Transformasi ini dilakukan untuk menyederhanakan proses klasifikasi multikelas serta memastikan konsistensi label pada seluruh dataset.

C. Extraksi Fitur

Setelah tahap pra-pemrosesan selesai, dilakukan proses ekstraksi fitur untuk membentuk variabel input yang akan digunakan dalam pemodelan klasifikasi.

1) *Rating Difference:* Fitur ini diperoleh dengan menghitung selisih antara rating pemain putih dan rating pemain hitam berupa Rating Difference = WhiteElo - BlackElo. Nilai ini merepresentasikan perbedaan tingkat kemampuan antar pemain dalam satu pertandingan.
2) *Move Count:* Jumlah langkah permainan dihitung berdasarkan total *play* yang tercatat pada partai dan dibagi dua untuk memperoleh jumlah langkah penuh. Fitur ini merepresentasikan durasi relatif pertandingan.
3) *Base Time:* Atribut *TimeControl* pada PGN memiliki format "base+increment". Pada penelitian ini hanya digunakan nilai waktu dasar (*base time*), yang diperoleh dengan memisahkan nilai sebelum simbol "+" dan mengonversinya ke dalam bentuk numerik (detik). Fitur-fitur tersebut dipilih karena: Tersedia secara konsisten pada seluruh data, bersifat numerik sehingga kompatibel dengan algoritma klasifikasi yang digunakan, Memiliki relevansi langsung terhadap dinamika pertandingan blitz.

D. Pembagian Data dan Standarisasi

Setelah proses ekstraksi fitur selesai, dataset dibagi menjadi data latih dan data uji untuk keperluan pelatihan dan evaluasi model.

1) *Pembagian Data:* Dataset dibagi menggunakan rasio 70:30 sebagai data latih dan data uji. Proses pembagian dilakukan dengan teknik stratified sampling memanfaatkan fungsi *train_test_split* yang tersedia dalam pustaka *scikit-learn* untuk mempertahankan proporsi distribusi kelas pada kedua subset. Pendekatan ini penting mengingat dataset memiliki karakteristik tidak seimbang, di mana kelas *draw* merupakan kelas minoritas. Parameter pembagian data ditetapkan dengan test size sebesar 0,3, nilai random state 42, serta opsi stratify menggunakan label untuk mempertahankan proporsi distribusi kelas pada data latih dan data uji. Penggunaan nilai random state bertujuan menjaga konsistensi hasil eksperimen dan memastikan replikasi penelitian dapat dilakukan.
2) *Standarisasi Fitur:* Proses standarisasi dilakukan terhadap fitur numerik menggunakan metode z-score normalization melalui StandardScaler. Standarisasi bertujuan untuk melakukan normalisasi sehingga setiap fitur memiliki rata-rata sebesar nol dengan simpangan baku satu. Standarisasi hanya diterapkan pada model Support Vector Machine (SVM) karena algoritma ini berbasis perhitungan jarak dan optimasi margin sehingga sensitif terhadap perbedaan skala antar fitur dalam pembentukan hyperplane optimal. Normalisasi atau standarisasi fitur telah dilaporkan dapat meningkatkan stabilitas dan performa klasifikasi SVM pada berbagai permasalahan supervised learning [11]. Proses standarisasi dilakukan dengan menghitung parameter skala berdasarkan data latih, kemudian menerapkan transformasi yang sama pada data uji. Mekanisme tersebut dilakukan guna memastikan tidak terjadi *data leakage*, yaitu keadaan ketika informasi dari dataset pengujian memengaruhi proses pelatihan dan berpotensi menghasilkan estimasi performa yang bias [12]. Sebaliknya, proses standarisasi tidak diterapkan pada Decision Tree dan Random Forest karena kedua algoritma tersebut berbasis pemisahan nilai atribut dan tidak bergantung pada skala numerik fitur.

E. Metode Klasifikasi

Penelitian ini menerapkan tiga algoritma klasifikasi berbasis supervised learning, yaitu Support Vector Machine (SVM), Decision Tree, dan Random Forest. Ketiga algoritma dipilih untuk merepresentasikan pendekatan berbasis margin, pohon keputusan tunggal, dan ensemble learning.

1) *Support Vector Machine (SVM):* Support Vector Machine (SVM) merupakan algoritma klasifikasi berbasis teori pembelajaran statistik yang beroperasi dengan membangun bidang pemisah optimal untuk memaksimalkan margin antar kelas dalam ruang fitur berdimensi tinggi [13]. Untuk menangani hubungan non-linear antar fitur, SVM mengandalkan mekanisme kernel yang mengubah representasi data ke ruang fitur dengan dimensi yang lebih kompleks. Pada penelitian ini, kernel yang dipilih adalah Radial Basis Function (RBF) karena kemampuannya dalam memodelkan batas keputusan non-linear [14]. Model SVM diimplementasikan menggunakan pustaka *scikit-learn* dengan konfigurasi parameter meliputi kernel Radial Basis Function (RBF) dengan parameter C bernilai 0,1 serta parameter gamma menggunakan skema *scale*. Parameter C berperan dalam mengontrol tingkat regularisasi model, sedangkan gamma menentukan pengaruh masing-masing titik data dalam pembentukan hyperplane optimal.
2) *Decision Tree:* Decision Tree merupakan algoritma klasifikasi berbasis struktur pohon yang membagi data

Received: Feb 19th, 2026 | Published: March 30th, 2026

p-ISSN: 2406-7768
e-ISSN: 2581-2181

berdasarkan nilai atribut tertentu untuk memaksimalkan homogenitas kelas pada setiap node [15]. Pada penelitian ini digunakan kriteria pemisahan *Gini impurity* dengan parameter hasil tuning berupa *criterion* gini, *max depth* sebesar 10, dan *min samples split* sebesar 2. Pembatasan kedalaman pohon diterapkan untuk mengurangi risiko *overfitting* pada data latih, mengingat pohon dengan kedalaman berlebihan cenderung mempelajari pola lokal yang bersifat *noise* sehingga menghasilkan performa yang terlalu optimistis pada data pelatihan [16].

3) *Random Forest:* Random Forest merupakan metode ensemble learning yang membangun sejumlah pohon keputusan melalui teknik bootstrap sampling dan menggabungkan hasil prediksi menggunakan mekanisme majority voting untuk meningkatkan stabilitas dan akurasi model [17]. Pada penelitian ini digunakan Random Forest dengan konfigurasi *number of estimators* sebanyak 200, parameter *class_weight* menggunakan skema *balanced*, serta nilai *random state* sebesar 42. Pendekatan *ensemble learning* ini diterapkan untuk meningkatkan stabilitas model dan mengurangi varians dibandingkan Decision Tree tunggal [18].
4) *Perbandingan Model:* Ketiga algoritma dilatih menggunakan dataset yang sama dan dievaluasi menggunakan metrik yang identik untuk memastikan perbandingan yang adil. Selain itu, pada setiap model dilakukan penerapan strategi cost-sensitive learning dengan parameter class_weight="balanced" serta dibandingkan dengan versi tanpa pembobotan kelas untuk mengukur dampak ketidakseimbangan distribusi data terhadap performa klasifikasi.

F. Strategi Cost-Sensitive Learning

Pada penelitian ini, data yang digunakan memiliki proporsi kelas yang tidak proporsional, dengan kelas *draw* hanya merepresentasikan sekitar 4% dari keseluruhan data. Ketidakseimbangan distribusi kelas dapat menyebabkan model lebih condong memprediksi kelas dominan sehingga kemampuan mendeteksi kelas minoritas menjadi berkurang [19]. Untuk mengatasi permasalahan tersebut, diterapkan pendekatan cost-sensitive learning melalui pemberian bobot kelas (class weighting) pada proses pelatihan model. Pendekatan ini meningkatkan pembobotan terhadap kesalahan prediksi pada proses klasifikasi pada kelas minoritas sehingga meningkatkan sensitivitas model terhadap kelas tersebut [20].

1) *Prinsip Cost-Sensitive Learning:* Dalam implementasi penelitian ini, pembobotan kelas dilakukan menggunakan parameter class_weight = "balanced". Parameter tersebut secara otomatis menghitung bobot kelas secara proporsional terbalik terhadap frekuensi kemunculan kelas dalam dataset. Secara umum, bobot kelas dirumuskan sebagai:

$$w_i = \frac{n}{k \cdot n_i} \qquad (1)$$

Di mana n adalah jumlah total sampel, k adalah jumlah kelas, dan n_i adalah jumlah sampel pada kelas ke-i. Pendekatan ini menyebabkan kelas minoritas memperoleh bobot lebih besar dibandingkan kelas mayoritas [20].

2) *Implementasi pada Model:* Strategi *cost-sensitive learning* diimplementasikan pada tiga metode klasifikasi, yakni Support Vector Machine, *Decision Tree*, dan *Random Forest*, melalui parameter class_weight="balanced" pada masing-masing model. Sebagai pembanding, setiap model juga dilatih tanpa pembobotan kelas untuk mengamati perbedaan performa yang dihasilkan. Pendekatan komparatif ini bertujuan untuk mengevaluasi sejauh mana pembobotan kelas mampu meningkatkan kemampuan model dalam mengenali kelas minoritas, serta mengukur *trade-off* antara peningkatan sensitivitas kelas minoritas dan penurunan performa pada kelas mayoritas.
3) *Tujuan Penerapan Strategi:* Penerapan *cost-sensitive learning* dalam penelitian ini dimaksudkan untuk menekan dominasi prediksi pada kelas mayoritas dan meningkatkan sensitivitas model dalam mendeteksi kelas *draw*, serta menghasilkan performa klasifikasi yang lebih seimbang antar kelas. Evaluasi dampak strategi ini dianalisis menggunakan metrik Macro F1-score serta nilai *recall* pada kelas minoritas, karena metrik tersebut lebih representatif dalam kondisi distribusi kelas yang tidak seimbang [21].

G. Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa klasifikasi dalam memprediksi hasil pertandingan catur blitz ke dalam tiga kelas, yaitu *win*, *lose*, dan *draw*. Mengingat karakteristik dataset yang tidak seimbang, evaluasi tidak hanya berfokus pada akurasi, tetapi juga mempertimbangkan metrik yang mampu merepresentasikan keseimbangan performa antar kelas.

1) *Accuracy:* Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data uji. Secara matematis, accuracy dirumuskan sebagai:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

Di mana TP adalah *True Positive*, TN adalah *True Negative*, FP adalah *False Positive*, dan FN adalah *False Negative*. Meskipun accuracy memberikan gambaran umum performa model, metrik ini kurang representatif pada dataset yang tidak seimbang [22].

2) *Precision, Recall, dan F1-Score:* Untuk mengevaluasi performa tiap kelas secara lebih rinci, digunakan metrik precision, recall, dan F1-score. Precision merepresentasikan perbandingan antara jumlah prediksi yang benar pada kelas positif terhadap seluruh prediksi positif yang dihasilkan model. Sementara itu, recall menunjukkan tingkat keberhasilan model dalam mengidentifikasi seluruh sampel aktual pada suatu kelas. Nilai F1-score dihitung sebagai rata-rata harmonik dari precision dan recall untuk merefleksikan keseimbangan keduanya. Karena penelitian ini melibatkan klasifikasi multikelas, digunakan pendekatan

Received: Feb 19th, 2026 | Published: March 30th, 2026

p-ISSN: 2406-7768
e-ISSN: 2581-2181

macro averaging untuk menghitung rata-rata F1-score secara merata pada setiap kelas tanpa dipengaruhi dominasi kelas mayoritas. Penggunaan Macro F1-score dipilih karena lebih representatif dalam kondisi dataset tidak seimbang [23].

3) *Confusion Matrix:* Confusion matrix digunakan untuk menganalisis distribusi kesalahan klasifikasi pada setiap kelas. Matriks ini menampilkan uraian terperinci mengenai hasil prediksi yang akurat serta yang tidak akurat pada masing-masing kombinasi kelas aktual dan kelas prediksi. Analisis confusion matrix membantu dalam memahami pola kesalahan model, khususnya dalam mengidentifikasi apakah kelas minoritas (draw) sering salah diklasifikasikan sebagai kelas mayoritas.
4) *Cross-Validation:* Untuk meningkatkan reliabilitas hasil evaluasi, diterapkan teknik Stratified K-Fold Cross-Validation dengan jumlah lipatan sebanyak 5 (5-fold) [10]. Pendekatan stratifikasi memastikan bahwa distribusi kelas pada setiap lipatan tetap proporsional terhadap dataset asli. Evaluasi cross-validation dilakukan menggunakan metrik Macro F1-score guna memperoleh estimasi performa model yang lebih stabil dan tidak bergantung pada satu pembagian data saja. Nilai rata-rata dan standar deviasi dari hasil cross-validation digunakan untuk menilai konsistensi performa model.

III. HASIL DAN PEMBAHASAN

A. Hasil

1) *Distribusi Dataset:* Dataset terdiri dari 100.000 partai catur blitz rated dengan tiga kelas hasil pertandingan.

Tabel 2. Distribusi Kelas Dataset

Kelas	Jumlah	Persentase
Win	49.749	49,75%
Lose	46.152	46,15%
Draw	4.099	4,10%

Distribusi tersebut menunjukkan ketidakseimbangan kelas yang signifikan, di mana kelas *draw* hanya merepresentasikan sekitar 4% dari keseluruhan data. Kondisi ini berpotensi menyebabkan model cenderung mengoptimalkan prediksi pada kelas mayoritas apabila tidak diterapkan strategi penanganan ketidakseimbangan.

2) *Performa Model Tanpa Cost-Sensitive Learning:* Evaluasi awal dilakukan tanpa penerapan pembobotan kelas (*class_weight*). Hasil pengujian pada data uji ditunjukkan pada Tabel IV.

Tabel 3. Performa Model Tanpa Class Weight

Model	Accuracy	Macro F1	Recall Draw
SVM	0,54	0,35	0,00
Decision Tree	0,52	0,37	0,03
Random Forest	0,49	0,38	0,11

Model tanpa pembobotan menunjukkan nilai accuracy yang relatif lebih tinggi. Namun demikian, performa pada kelas minoritas sangat rendah. Pada model SVM, recall kelas *draw* mencapai 0,00, yang menunjukkan bahwa model gagal mendeteksi kelas tersebut. Hal ini mengindikasikan bahwa accuracy tidak cukup representatif dalam kondisi distribusi kelas yang tidak seimbang.

3) *Performa Model Dengan Cost-Sensitive Learning:* Selanjutnya dilakukan eksperimen dengan menerapkan class_weight="balanced".

Tabel 4. Performa Model Dengan Class Weight

Model	Accuracy	Macro F1	Recall Draw
SVM	0,45	0,40	0,73
Decision Tree	0,44	0,39	0,73
Random Forest	0,47	0,37	0,21

Penerapan cost-sensitive learning meningkatkan recall kelas *draw* secara signifikan pada SVM dan Decision Tree, dari hampir nol menjadi 0,73. Namun demikian, peningkatan ini diikuti dengan penurunan accuracy keseluruhan sekitar 8%. Selain itu, precision kelas *draw* relatif rendah (sekitar 0,10–0,12 pada model balanced), yang menunjukkan adanya peningkatan jumlah *false positive*. Hal ini menegaskan adanya trade-off antara sensitivitas terhadap kelas minoritas dan ketepatan prediksi pada kelas mayoritas.

4) *Cross-Validation:* Untuk mengevaluasi stabilitas model, dilakukan 5-fold stratified cross-validation dengan metrik Macro F1-score.

Tabel 5. Hasil Cross-Validation (Macro F1, 5-Fold)

Model	Mean	Std Dev
SVM	0,397	0,0017
Decision Tree	0,390	0,0055
Random Forest	0,369	0,0045

Nilai standar deviasi yang relatif kecil menunjukkan bahwa performa model stabil terhadap variasi pembagian data. Nilai Macro F1 cross-validation yang mendekati hasil pada data uji menunjukkan konsistensi model dan meminimalkan risiko overfitting.

5) *Perbandingan Dampak Cost-Sensitive Learning pada SVM:* Untuk memperjelas dampak pembobotan kelas, dilakukan perbandingan langsung pada model SVM.

Tabel 6. Perbandingan SVM Dengan dan Tanpa Class Weight

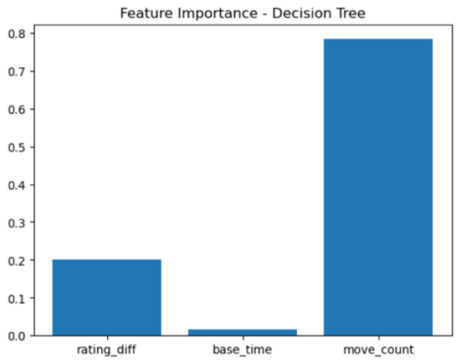
Model	Accuracy	Macro F1	Recall Draw
SVM (Tanpa Weight)	0,54	0,35	0,00
SVM (Balanced)	0,45	0,40	0,73

Received: Feb 19th, 2026 | Published: March 30th, 2026

p-ISSN: 2406-7768
e-ISSN: 2581-2181

Tanpa pembobotan kelas, model mengoptimalkan prediksi pada kelas mayoritas sehingga menghasilkan accuracy lebih tinggi, tetapi gagal mendeteksi kelas minoritas. Sebaliknya, dengan pembobotan kelas, model menjadi lebih sensitif terhadap kelas *draw*, meskipun akurasi keseluruhan menurun.

6) *Analisis Feature Importance:* Analisis *feature importance* menggunakan *Decision Tree* menunjukkan kontribusi relatif masing-masing fitur. Hasil perhitungan menunjukkan bahwa *move count* memiliki nilai kepentingan sekitar 0,78, diikuti oleh *rating difference* sebesar sekitar 0,20, dan *base time* sekitar 0,02.



Gambar 2. Feature Importance Model Decision Tree pada Dataset Blitz

Fitur *move count* memiliki kontribusi paling dominan dalam menentukan hasil pertandingan blitz. Hal ini mengindikasikan bahwa durasi permainan memiliki pengaruh yang lebih besar dibandingkan selisih rating atau waktu dasar permainan.

B. *Pembahasan*

Hasil eksperimen menunjukkan bahwa ketidakseimbangan distribusi kelas memiliki dampak signifikan terhadap performa model klasifikasi. Model tanpa pembobotan cenderung mengoptimalkan akurasi keseluruhan dengan memprioritaskan kelas mayoritas, sehingga mengabaikan kelas minoritas.

Penerapan cost-sensitive learning terbukti meningkatkan kemampuan model dalam mendeteksi kelas draw. Peningkatan recall yang signifikan pada SVM dan Decision Tree menunjukkan bahwa pembobotan kelas efektif dalam mengurangi bias terhadap kelas mayoritas. Namun demikian, peningkatan sensitivitas tersebut diikuti oleh penurunan precision dan akurasi keseluruhan, yang mencerminkan adanya trade-off antara fairness antar kelas dan performa global.

Nilai Macro F1-score yang meningkat pada model balanced menunjukkan bahwa metrik ini lebih representatif dibandingkan accuracy dalam kondisi distribusi kelas tidak seimbang. Temuan ini konsisten dengan teori klasifikasi pada kondisi class imbalance, di mana evaluasi berbasis rata-rata makro memberikan gambaran yang lebih adil terhadap seluruh kelas.

Menariknya, Random Forest tidak menunjukkan peningkatan performa yang signifikan dibandingkan Decision Tree tunggal. Hal ini mengindikasikan bahwa dengan fitur metadata yang relatif sederhana, kompleksitas metode ensemble tidak memberikan keuntungan yang substansial.

Penelitian ini memiliki keterbatasan pada penggunaan fitur berbasis metadata tanpa mempertimbangkan informasi posisi papan atau evaluasi engine catur. Dengan demikian, kompleksitas taktis dan strategis permainan belum sepenuhnya terwakili dalam model.

Penelitian ini memberikan kontribusi empiris dalam mengevaluasi efektivitas cost-sensitive learning pada klasifikasi hasil pertandingan catur blitz berskala besar. Selain itu, penelitian ini menunjukkan pentingnya penggunaan metrik Macro F1-score dalam mengevaluasi performa model pada dataset tidak seimbang, serta mengungkap trade-off yang terjadi antara peningkatan sensitivitas kelas minoritas dan penurunan akurasi keseluruhan.

IV. KESIMPULAN

Penelitian ini berhasil mengevaluasi dampak ketidakseimbangan kelas terhadap performa klasifikasi hasil pertandingan catur blitz pada dataset berskala besar serta menguji efektivitas pendekatan cost-sensitive learning dalam mengatasinya. Hasil eksperimen menunjukkan bahwa model tanpa penanganan ketidakseimbangan cenderung bias terhadap kelas mayoritas dan gagal mendeteksi kelas minoritas secara memadai, meskipun menghasilkan nilai akurasi yang relatif tinggi. Penerapan pembobotan kelas melalui skema class_weight = "balanced" terbukti mampu meningkatkan sensitivitas terhadap kelas draw secara signifikan dan menghasilkan performa yang lebih seimbang antar kelas berdasarkan metrik Macro F1-score. Meskipun terjadi penurunan akurasi keseluruhan, pendekatan ini memberikan representasi evaluasi yang lebih adil pada kondisi distribusi data yang tidak seimbang.

Berdasarkan hasil pengujian dan validasi silang, model Support Vector Machine dengan pembobotan kelas menunjukkan performa paling konsisten dalam menjaga keseimbangan antara deteksi kelas minoritas dan stabilitas model. Dengan demikian, tujuan penelitian untuk mengevaluasi dan membuktikan efektivitas cost-sensitive learning pada klasifikasi hasil catur blitz tidak seimbang dapat dinyatakan tercapai. Temuan ini menegaskan pentingnya strategi pembelajaran yang sadar ketidakseimbangan serta penggunaan metrik evaluasi yang tepat dalam analisis klasifikasi multikelas berskala besar.

V. SARAN

Berdasarkan hasil penelitian yang telah dilakukan, pengembangan penelitian selanjutnya dapat diarahkan pada beberapa aspek penting. Pertama, penelitian ini masih memanfaatkan fitur berbasis metadata pertandingan, seperti rating difference, move count, dan base time, yang belum sepenuhnya merepresentasikan kompleksitas taktis dan strategis permainan catur. Oleh karena itu, integrasi fitur yang lebih kaya, seperti evaluasi posisi papan menggunakan engine

Received: Feb 19th, 2026 | Published: March 30th, 2026

p-ISSN: 2406-7768
e-ISSN: 2581-2181

catur, representasi konfigurasi bidak, atau ekstraksi pola langkah permainan, berpotensi meningkatkan kemampuan model dalam menangkap dinamika pertandingan secara lebih komprehensif.

Kedua, selain pendekatan cost-sensitive learning, penelitian selanjutnya dapat mengeksplorasi metode penanganan ketidakseimbangan kelas lainnya, seperti teknik oversampling (misalnya SMOTE), undersampling, maupun pendekatan ensemble khusus untuk data tidak seimbang, sehingga diperoleh perbandingan efektivitas yang lebih menyeluruh antar strategi penanganan class imbalance.

Ketiga, penggunaan algoritma yang lebih kompleks, seperti metode gradient boosting atau pendekatan berbasis deep learning, dapat dipertimbangkan untuk mengevaluasi potensi peningkatan performa pada dataset berskala besar. Selain itu, perluasan kajian pada mode permainan lain seperti rapid atau bullet juga dapat memberikan wawasan tambahan mengenai variasi distribusi hasil pertandingan dan implikasinya terhadap performa klasifikasi.

Referensi

[1] A. Rapp and A. Boldi, "The quantification of the gaming experience: Self-tracking practices and game metrics among casual players, esports players, and streamers," *Comput. Human Behav.*, vol. 174, Jan. 2026, doi: 10.1016/j.chb.2025.108826.

[2] S. Chowdhary, I. Iacopini, and F. Battiston, "Quantifying human performance in chess," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-27735-9.

[3] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc. Sci. Res.*, vol. 110, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

[4] K. Samara, A. Antreassian, M. Klug, and M. S. Hasan, "Machine Learning Approaches for Classifying Chess Game Outcomes: A Comparative Analysis of Player Ratings and Game Dynamics," *Electronics (Basel).*, vol. 15, no. 1, p. 1, Dec. 2025, doi: 10.3390/electronics15010001.

[5] Q. Zhou and B. Sun, "Adaptive K-means clustering based under-sampling methods to solve the class imbalance problem," *Data Inf. Manag.*, vol. 8, no. 3, Sep. 2024, doi: 10.1016/j.dim.2023.100064.

[6] T. Ishii, T. Yang, and M. Goto, "BIG-PU: An evaluation metric for exploration based on preference elicitation in recommender systems," *Expert Syst. Appl.*, vol. 308, p. 131077, May 2026, doi: 10.1016/j.eswa.2025.131077.

[7] M. Ali, "Classification of imbalanced travel mode choice dataset with SMOTE and prediction using interpretable machine learning," *Sustainable Futures*, vol. 10, Dec. 2025, doi: 10.1016/j.sftr.2025.101119.

[8] R. Drezewski and G. Wator, "Chess as sequential data in a chess match outcome prediction using deep learning with various chessboard representations," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 1760–1769. doi: 10.1016/j.procs.2021.08.180.

[9] Y. Bhargava, S. K. Shetty, and V. Baths, "Subjective Cognitive Decline Prediction on Imbalanced Data Using Data-Resampling and Cost-Sensitive Training Methods," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1964–1979. doi: 10.1016/j.procs.2024.04.186.

[10] S. Ünalan, O. Günay, I. Akkurt, K. Gunoglu, and H. O. Tekin, "A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning," *J. Radiat. Res. Appl. Sci.*, vol. 17, no. 4, p. 101080, Dec. 2024, doi: 10.1016/j.jrras.2024.101080.

[11] Y. Qin, Q. Xu, T. Kujala, X. Wang, and F. Cong, "Evaluating spatial normalization for SVM-based EEG decoding: A within- and between-subjects perspective," *Biomed. Signal Process. Control*, vol. 116, May 2026, doi: 10.1016/j.bspc.2026.109535.

[12] M. A. Lones, "Avoiding common machine learning pitfalls," Oct. 11, 2024, *Cell Press*. doi: 10.1016/j.patter.2024.101046.

[13] R. De Leone, F. Maggioni, and A. Spinelli, "A robust twin parametric margin support vector machine for multiclass classification," *EURO Journal on Computational Optimization*, vol. 13, Jan. 2025, doi: 10.1016/j.ejco.2025.100115.

[14] A. Alfaleh, N. Ben Khedher, S. M. Eldin, M. Alturki, I. Elbadawi, and R. Kumar, "Predicting thermal conductivity and dynamic viscosity of nanofluid by employment of Support Vector Machines: A review," *Energy Reports*, vol. 10, pp. 1259–1267, Nov. 2023, doi: 10.1016/j.egyr.2023.08.001.

[15] M. Curran, E. Howley, and J. Duggan, "Classification of System Dynamics model outputs using decision trees," *Machine Learning with Applications*, vol. 21, p. 100713, Sep. 2025, doi: 10.1016/j.mlwa.2025.100713.

[16] L. Barreñada, P. Dhiman, D. Timmerman, A.-L. Boulesteix, and B. Van Calster, "Understanding overfitting in random forest for probability estimation: a visualization and simulation study," *Diagn. Progn. Res.*, vol. 8, no. 1, Sep. 2024, doi: 10.1186/s41512-024-00177-1.

[17] C. Liu *et al.*, "Metabolomics for origin traceability of lamb: An ensemble learning approach based on random forest recursive feature elimination," *Food Chem. X*, vol. 29, Jul. 2025, doi: 10.1016/j.fochx.2025.102856.

[18] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Dec. 15, 2024, *Mesopotamian Academic Press*. doi: 10.58496/BJML/2024/007.

[19] J. C-Rella, G. Claeskens, R. Cao, and J. M. Vilar, "Instance-dependent cost-sensitive parametric learning," *Neurocomputing*, vol. 615, Jan. 2025, doi: 10.1016/j.neucom.2024.128875.

[20] B. BAKIRARAR and A. H. ELHAN, "Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research," *Turkiye Klinikleri Journal of Biostatistics*, vol. 15, no. 1, pp. 19–29, 2023, doi: 10.5336/biostatic.2022-93961.

[21] A. X. Wang, V. T. Le, H. N. Trung, and B. P. Nguyen, "Addressing imbalance in health data: Synthetic minority oversampling using deep learning," *Comput. Biol. Med.*, vol. 188, Apr. 2025, doi: 10.1016/j.compbiomed.2025.109830.

[22] J. Zheng, S. Wang, H. Yan, and H. Sun, "Binary classification for imbalanced datasets using a novel metric method," *Egyptian Informatics Journal*, vol. 33, p. 100890, Mar. 2026, doi: 10.1016/j.eij.2026.100890.

[23] S. Liu *et al.*, "Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies," *Osteoarthritis Cartilage*, vol. 31, no. 9, pp. 1242–1248, Sep. 2023, doi: 10.1016/j.joca.2023.05.006.

Received: Feb 19th, 2026 | Published: March 30th, 2026