JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika

[E-ISSN 2656-0321]

[Vol 20 No 2]

Evaluasi Empiris Kinerja dan Stabilitas Model pada Klasifikasi Data Tabular

Nahumi Nugrahaningsih [a,1,*], Felicia Sylviana [a,2], Abertun Sagit Sahay [a,3]

[a] Jurusan Teknik Informatika, Universitas Palangka Raya, Kampus UPR Jl. Yos Sudarso Palangka Raya, 73112
[1] nahumi@it.upr.ac.id*; [2] feliciasylviana@it.upr.ac.id ; [3] abertun@it.upr.ac.id
* corresponding author

ARTICLE INFO

Keywords
Benchmark Evaluation, Gradient Boosting, In-Context Learning, Tabular Classification, Tabular Data, Tabular Foundation Models

ABSTRACT

Data tabular banyak digunakan dalam berbagai aplikasi *machine learning*, dengan gradient boosting sebagai salah satu pendekatan yang umum digunakan untuk *task* klasifikasi. Perkembangan Tabular Foundation Models (FM) menawarkan alternatif melalui prapelatihan dan in-context learning. Namun, perbandingan FM dan gradient boosting dalam kondisi praktis tanpa hyperparameter tuning masih terbatas. Penelitian ini mengevaluasi tiga Tabular FM (TabPFN v2, TabICL, dan TabDPT) dan tiga metode gradient boosting (XGBoost, LightGBM, dan CatBoost) pada delapan dataset klasifikasi publik dengan karakteristik yang beragam, mencakup perbedaan ukuran data, jumlah fitur, jumlah kelas, dan distribusi kelas. Evaluasi dilakukan menggunakan konfigurasi bawaan melalui *stratified five-fold cross-validation* dan sepuluh *random seed*, dengan mempertimbangkan kinerja, ukuran dataset, ketidakseimbangan kelas, dan stabilitas. Hasil menunjukkan perbedaan kinerja yang signifikan ($p < 0{,}001$), dengan FM memberikan hasil terbaik pada tujuh dari delapan dataset. Keunggulan FM paling terlihat pada dataset kecil, tetapi keunggulan ini mengecil pada data yang tidak seimbang. CatBoost tetap kompetitif dan tidak berbeda signifikan dari ketiga FM, sedangkan stabilitas keenam model relatif serupa. Hasil ini menunjukkan bahwa FM merupakan pilihan yang kuat untuk dataset kecil tanpa tuning, namun demikian keunggulannya terhadap gradient boosting bergantung pada karakteristik data.

1. Pendahuluan

Data tabular merupakan salah satu bentuk data yang paling umum digunakan dalam aplikasi machine learning, mulai dari bidang kesehatan dan keuangan hingga ilmu sosial. Klasifikasi data tabular banyak dipergunakan untuk mendukung berbagai jenis pengambilan keputusan penting; seperti prediksi risiko penyakit, penilaian kredit, dan deteksi penipuan [1], [2]. Dibandingkan dengan citra dan teks, data tabular memiliki sejumlah karakteristik khusus; antara lain tipe fitur yang beragam, nilai yang hilang, ketidakseimbangan kelas, serta interaksi antarfitur yang kompleks [3], [4], [5], [6].

Selama lebih dari satu dekade, metode gradient boosting (GBM) telah menjadi salah satu pendekatan utama untuk klasifikasi data tabular. XGBoost [7], LightGBM [8], dan CatBoost [9] banyak digunakan karena memiliki kinerja yang kuat dan konsisten, mampu menangani nilai yang hilang, serta dapat digunakan bersama metode interpretabilitas seperti SHAP. Karakteristik tersebut membuat GBM banyak digunakan dalam berbagai benchmark maupun aplikasi pada bidang yang membutuhkan interpretabilitas model [5], [10], [11], [12].

Dalam beberapa tahun terakhir, *Tabular Foundation Model* (FM) mulai berkembang sebagai pendekatan baru untuk klasifikasi data tabular. Model ini menggunakan arsitektur *transformer* yang telah melalui proses prapelatihan dan dapat melakukan prediksi melalui *in-context learning* [13], [14]. TabPFN [2], misalnya, menunjukkan kinerja yang kompetitif pada dataset kecil. Pengembangan

berikutnya memperluas kemampuan pendekatan ini. TabPFN v2 [15] mendukung dataset hingga 10.000 sampel, sedangkan TabICL [16] dirancang untuk menangani dataset hingga 500.000 sampel melalui mekanisme atensi dua tahap. Sementara itu, TabDPT [17] menggunakan *discriminative pre-training* untuk meningkatkan kemampuan generalisasi.

Sejumlah penelitian menunjukkan bahwa FM cenderung unggul pada dataset kecil, sedangkan GBM tetap kompetitif pada dataset yang lebih besar, mengandung lebih banyak *noise*, atau memiliki ketidakseimbangan kelas [11], [18]. Pola serupa juga ditemukan dalam beberapa benchmark berskala besar [19], [20]. Selain itu, kombinasi kedua kelompok model melalui teknik *ensemble* dilaporkan dapat meningkatkan kinerja [21], [22]. Meski demikian, hasil perbandingan yang tersedia belum sepenuhnya menggambarkan penggunaan model dalam kondisi praktis. Sebagian penelitian hanya mencakup satu domain atau sejumlah kecil dataset, sementara penelitian lainnya menggunakan *hyperparameter tuning* yang ekstensif. Kondisi ini membuat kinerja model sulit dipisahkan dari pengaruh proses *tuning* yang dilakukan.

Penelitian ini membandingkan tiga FM, yaitu TabPFN v2, TabICL, dan TabDPT, dengan tiga GBM, yaitu XGBoost, LightGBM, dan CatBoost, pada delapan dataset benchmark. Seluruh model dijalankan menggunakan konfigurasi *hyperparameter* bawaan untuk merepresentasikan skenario penggunaan praktis tanpa tuning yang ekstensif. Penelitian ini menjawab lima pertanyaan berikut:

- RQ1: Apakah Tabular Foundation Model secara konsisten mengungguli GBM pada berbagai dataset benchmark?
- RQ2: Apakah kinerja relatif kedua kelompok model berubah berdasarkan ukuran dataset?
- RQ3: Apakah FM lebih robust terhadap ketidakseimbangan kelas dibandingkan GBM pada konfigurasi bawaan?
- RQ4: Model mana yang memiliki kinerja paling stabil pada berbagai random seed?
- RQ5: Dalam kondisi seperti apa FM lebih tepat dipilih dibandingkan GBM?

Penelitian ini memberikan empat kontribusi utama. Pertama, menyajikan perbandingan empiris antara FM dan GBM pada beberapa dataset dengan konfigurasi bawaan. Kedua, mengevaluasi model berdasarkan empat aspek, yaitu akurasi, macro F1-score, AUC-ROC, dan stabilitas. Ketiga, menggunakan uji statistik nonparametrik untuk menguji perbedaan kinerja yang ditemukan. Keempat, memberikan panduan praktis dalam memilih antara FM dan GBM berdasarkan karakteristik dataset.

2. Metodologi Penelitian

2.1. Desain Penelitian

Penelitian ini menggunakan desain perbandingan empiris beberapa model klasifikasi. Seluruh model dievaluasi menggunakan konfigurasi *hyperparameter* bawaan tanpa proses tuning tambahan. Pengaturan ini memungkinkan kinerja model dibandingkan tanpa dipengaruhi oleh perbedaan strategi atau intensitas tuning. Selain itu, pengaturan tersebut mencerminkan penggunaan praktis ketika model diterapkan langsung dengan konfigurasi bawaan. Dengan rancangan yang sama untuk seluruh model, hasil evaluasi dapat dibandingkan secara konsisten.

2.2. Pemilihan Model

Enam model digunakan dalam penelitian ini, terdiri atas tiga *Tabular Foundation Model* (FM) dan tiga metode *Gradient Boosting* (GBM), seperti ditunjukkan pada Tabel 1. Pemilihan tiga model dari masing-masing kelompok memungkinkan perbandingan dilakukan tidak hanya antar-model, tetapi juga antara FM dan GBM sebagai dua pendekatan untuk klasifikasi data tabular. Ketiga FM mewakili pendekatan yang berbeda, yaitu *meta-learning* dengan prapelatihan pada TabPFN v2, *two-stage in-context learning* pada TabICL, dan *discriminative pre-training* pada TabDPT. XGBoost, LightGBM, dan CatBoost dipilih untuk mewakili GBM karena ketiganya telah banyak digunakan untuk klasifikasi data tabular [3], [11], [12].

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 2]

Tabel 1. Model yang Digunakan dalam Penelitian

Kelompok	Model	Tahun	Mekanisme	Referensi
Tabular FM	TabPFN v2	2024	Meta-learning Transformer, in-context learning	[15]
Tabular FM	TabICL	2024	Two-stage column-then-row attention, in-context learning	[16]
Tabular FM	TabDPT	2024	Discriminative pre-training Transformer	[17]
Gradient Boosting	XGBoost	2016	Regularized gradient boosted decision trees	[7]
Gradient Boosting	LightGBM	2017	Histogram-based gradient boosting	[8]
Gradient Boosting	CatBoost	2018	Ordered boosting, dukungan fitur kategorikal	[9]

2.3. Datasets

Delapan dataset publik dari UCI Machine Learning Repository digunakan dalam penelitian ini. Dataset mencakup lima domain, yaitu medis, pangan, umum, sosial, dan keuangan, dengan tugas klasifikasi biner maupun multikelas. Ukuran dataset berkisar antara 303 hingga 48.842 sampel sebelum dilakukan *subsampling*. Karakteristik masing-masing dataset disajikan pada Tabel 2.

Dua dataset berukuran besar, Adult Income dan Bank Marketing, diambil sebagian menjadi 5.000 sampel menggunakan *stratified random sampling* (random_state = 42). Pembatasan ini mempertimbangkan karakteristik TabPFN v2 yang dirancang untuk dataset hingga 10.000 sampel. Pada ukuran yang lebih besar, waktu komputasi meningkat karena seluruh data latih digunakan sebagai konteks saat inferensi. Selain itu, penggunaan dataset dalam ukuran aslinya dapat lebih menguntungkan GBM yang memiliki skalabilitas lebih baik. *Stratified sampling* digunakan agar distribusi kelas pada sampel tetap mengikuti distribusi data asal.

Tabel 2. Dataset Benchmark yang Digunakan dalam Penelitian. *Diambil 5.000 sampel

ID	Dataset	Domain	n	Fitur	Task	Distribusi Kelas	Referensi
HD	Heart Disease	Medis	303	13	Biner	Pos: 54,1% / Neg: 45,9%	[23]
PD	Pima Diabetes	Medis	767	8	Biner	Pos: 34,8% / Neg: 65,2%	[24]
CKD	CKD	Medis	400	24	Biner	CKD: 62,5% / Not: 37,5%	[25]
BC	Breast Cancer	Medis	569	30	Biner	Mal: 37,3% / Ben: 62,7%	[26]
WQ	Wine Quality	Pangan/Kimia	6.497	11	Multikelas (3)	Low: 3,8% / Med: 76,6% / High: 19,7%	[27]
CE	Car Evaluation	Umum	1.728	6	Multikelas (4)	Unacc: 70,0% / Acc: 22,2% / Good: 4,0% / Vgood: 3,8%	[28]
AI*	Adult Income	Sosial	48.842	14	Biner	≤50K: 76,1% / >50K: 23,9%	[29]
BM*	Bank Marketing	Keuangan	45.211	16	Biner	No: 88,3% / Yes: 11,7%	[30]

2.4. Preprocessing

Seluruh dataset diproses dengan prosedur preprocessing yang sama. Nilai yang hilang pada fitur numerik diisi dengan median, sedangkan pada fitur kategorikal digunakan nilai yang paling sering muncul (modus). Fitur kategorikal kemudian dikodekan dengan label encoding, sementara fitur numerik distandardisasi menggunakan StandardScaler. Untuk mencegah kebocoran data, standardisasi

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 2]

hanya mengacu pada data latih di setiap fold. Fitur yang memiliki lebih dari 60% nilai hilang dihapus sebelum proses imputasi, sedangkan variabel target dikodekan menjadi bilangan bulat (0, 1, ..., k−1).

Beberapa penyesuaian dilakukan sesuai karakteristik dataset. Pada Heart Disease, lima tingkat keparahan penyakit disederhanakan menjadi dua kelas, yaitu tidak ada penyakit (0) dan terdapat penyakit (1), mengikuti praktik yang umum digunakan [23]. Pada CKD, spasi dan karakter tab pada kolom target dibersihkan untuk menghindari munculnya label kelas ganda. Sementara itu, skor pada Wine Quality dikelompokkan menjadi tiga tingkat kualitas: rendah (3–4), sedang (5–6), dan tinggi (7–9).

2.5. Experimental Setup

Setiap model dievaluasi menggunakan *Stratified K-Fold cross-validation* (k = 5) dengan 10 *seed*. Tidak dilakukan *hyperparameter tuning*, sehingga seluruh model menggunakan konfigurasi bawaan. Standardisasi fitur dilakukan berdasarkan data latih pada setiap *fold*, kemudian diterapkan pada data uji. Secara keseluruhan, setiap model menjalani 400 proses evaluasi. Rincian pengaturan eksperimen disajikan pada Tabel 3.

Tabel 3. Pengaturan Eksperimen

Parameter	Pengaturan
Validasi silang	Stratified K-Fold (k = 5)
Jumlah seed	10 (seed 0–9)
Total evaluasi per model	10 seed × 5 fold × 8 dataset = 400
Imputasi nilai hilang	Median (numerik); modus (kategorikal)
Pengkodean fitur kategorikal	Label encoding, konsisten untuk seluruh model
Standardisasi fitur	StandardScaler (fit pada data latih, transform pada data uji)
Strategi *hyperparameter*	Konfigurasi bawaan untuk seluruh model (tanpa tuning)
Subsampling Adult & Bank	Stratified random sampling hingga n = 5.000 (random_state = 42)

2.6. Metrik Evaluasi

Kinerja model dievaluasi menggunakan empat metrik, yaitu akurasi, *macro F1-score*, AUC-ROC, dan simpangan baku akurasi antar-*seed*. Akurasi mengukur proporsi prediksi yang benar, sedangkan *macro F1-score* memberikan bobot yang sama pada setiap kelas sehingga lebih sesuai untuk melihat kinerja pada *data* yang tidak seimbang. AUC-ROC digunakan untuk mengukur kemampuan model dalam membedakan kelas, dengan pendekatan *one-vs-rest* pada klasifikasi multikelas. Stabilitas model dilihat dari simpangan baku akurasi yang diperoleh dari berbagai *seed*. Waktu pelatihan dan inferensi tidak disertakan karena penggunaan lingkungan komputasi berbasis *cloud* dapat menyebabkan variasi waktu eksekusi yang sulit dikendalikan.

2.7. Dimensi Analisis dan Uji Statistik

Analisis dilakukan dalam lima dimensi yang sesuai dengan RQ1-RQ5, sebagaimana dirangkum pada Tabel 4. *Perbedaan* kinerja antar-model diuji menggunakan metode nonparametrik. Uji Friedman digunakan untuk melihat apakah terdapat perbedaan kinerja secara keseluruhan, kemudian dilanjutkan dengan uji *post-hoc* Nemenyi untuk mengetahui pasangan model yang berbeda. Sebagai analisis tambahan, uji Wilcoxon *signed-rank* dilakukan pada setiap pasangan model dengan koreksi Bonferroni.

Tabel 4. Dimensi Analisis dan Pertanyaan Penelitian

RQ	Dimensi	Pertanyaan Penelitian	Analisis

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 2]

RQ1	Kinerja keseluruhan	Apakah FM tabular secara konsisten mengungguli GBM pada berbagai dataset?	Uji Friedman dan post-hoc Nemenyi berdasarkan rata-rata akurasi
RQ2	Pengaruh ukuran dataset	Apakah keunggulan FM berubah seiring ukuran dataset?	Perbandingan rata-rata akurasi pada dataset kecil, sedang, dan besar
RQ3	Ketidakseimbangan kelas	Apakah FM lebih robust terhadap ketidakseimbangan kelas dibandingkan GBM tanpa teknik tambahan?	Perbandingan macro F1-score pada dataset seimbang dan tidak seimbang
RQ4	Stabilitas kinerja	Model mana yang menunjukkan kinerja paling konsisten pada berbagai seed?	Simpangan baku akurasi dan distribusi melalui box plot
RQ5	Panduan praktis	Dalam kondisi apa FM lebih tepat digunakan dibandingkan GBM?	Sintesis temuan RQ1–RQ4 sebagai dasar pemilihan model

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil evaluasi tiga *Tabular Foundation Model* (TabPFN v2, TabICL, dan TabDPT) dan tiga metode *Gradient Boosting* (XGBoost, LightGBM, dan CatBoost) pada delapan dataset *benchmark*. Hasil disusun berdasarkan pertanyaan penelitian, disertai interpretasi singkat untuk memperjelas temuan. Pembahasan yang mengaitkan hasil dengan penelitian sebelumnya disajikan pada bagian akhir.

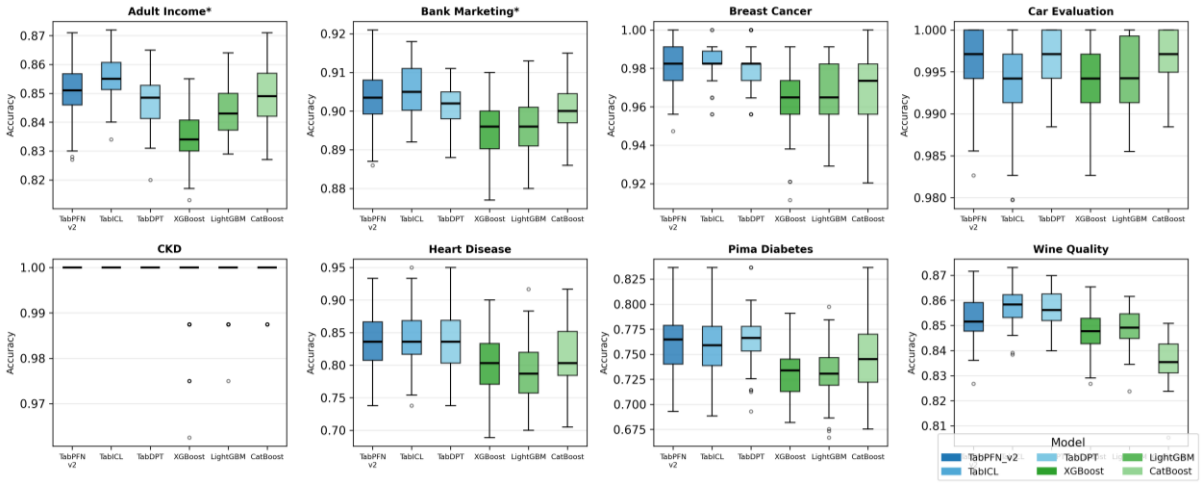
3.1. Kinerja Klasifikasi Keseluruhan (RQ1)

Tabel 5 menyajikan rata-rata akurasi, *macro F1-score*, dan AUC-ROC dari 10 *seed* × 5-*fold cross-validation* untuk setiap model dan dataset. Agar lebih ringkas, tabel hanya menampilkan nilai rata-rata, sedangkan simpangan baku disajikan secara terpisah pada Tabel 10. Nilai terbaik untuk setiap dataset dan metrik dicetak tebal.

Secara umum, FM memperoleh hasil terbaik atau setara terbaik pada tujuh dari delapan dataset. Keunggulan FM paling terlihat pada Heart Disease, Pima Diabetes, dan Wine Quality. Pada CKD, ketiga FM mencapai nilai sempurna (1,000) untuk akurasi, *macro F1-score*, dan AUC-ROC, sedangkan CatBoost sebagai GBM terbaik mencapai akurasi 0,999. Hasil berbeda ditemukan pada Car Evaluation, ketika CatBoost menghasilkan akurasi (0,997) dan *macro F1-score* (0,993) tertinggi. Dataset ini seluruhnya terdiri atas fitur kategorikal dengan struktur hierarkis yang jelas, karakteristik yang sesuai dengan model berbasis pohon.

Di antara ketiga FM, tidak ada satu model yang selalu unggul. TabPFN v2 dan TabICL menempati posisi teratas pada sebagian besar dataset, sedangkan TabDPT tetap kompetitif meskipun tidak menunjukkan dominasi yang jelas. Pada kelompok GBM, CatBoost secara umum memberikan hasil yang lebih baik dibandingkan XGBoost dan LightGBM.

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 2]



Gambar 1. Distribusi akurasi setiap model pada masing-masing dataset (10 seed × 5 fold = 50 pengamatan per box). Kotak biru menunjukkan Foundation Model, sedangkan kotak hijau menunjukkan metode Gradient Boosting.

3.2. Peringkat Statistik (RQ1)

Berdasarkan akurasi, ketiga FM menempati tiga peringkat teratas. TabPFN v2 dan TabICL sama-sama memperoleh peringkat rata-rata 2,125, diikuti TabDPT sebesar 2,500. CatBoost berada di posisi keempat dengan 3,750, sedangkan LightGBM dan XGBoost masing-masing memperoleh 5,000 dan 5,500. Uji Friedman menunjukkan bahwa secara keseluruhan terdapat perbedaan kinerja yang signifikan antar-model ($\chi^2 = 34{,}21$; df = 5; $p < 0{,}001$).

Meskipun berada di peringkat keempat, CatBoost tidak berbeda secara signifikan dari ketiga FM berdasarkan uji Nemenyi (p = 0,507-0,765). Sebaliknya, sebagian besar perbandingan FM dengan XGBoost dan LightGBM menunjukkan perbedaan yang signifikan. Tidak ditemukan perbedaan yang signifikan di antara ketiga FM. Dalam kelompok GBM, CatBoost juga menunjukkan kinerja yang lebih baik secara signifikan dibandingkan LightGBM dan XGBoost.

3.3. Kinerja Berdasarkan Ukuran Dataset (RQ2)

Untuk melihat pengaruh ukuran data, dataset dikelompokkan menjadi tiga kategori, yaitu kecil (n < 600), sedang (600–2.000), dan besar (2.000–5.000). FM menghasilkan rata-rata akurasi yang lebih tinggi pada ketiga kelompok. Selisih terbesar ditemukan pada dataset kecil, dengan rata-rata akurasi 0,940 untuk FM dan 0,921 untuk GBM. Selisih tersebut menyempit pada dataset sedang (0,878 dibandingkan 0,867) dan dataset besar (0,870 dibandingkan 0,861).

Pola ini menunjukkan bahwa keunggulan FM lebih terlihat ketika jumlah data terbatas. Namun, hasil tersebut perlu dibaca dengan hati-hati karena kelompok ukuran terdiri atas dataset dan tugas klasifikasi yang berbeda. Dengan demikian, penurunan akurasi pada kelompok yang lebih besar tidak dapat langsung diartikan sebagai akibat bertambahnya ukuran dataset. Hasil ini lebih tepat dipandang sebagai pola yang muncul pada dataset yang digunakan dalam penelitian ini.

3.4. Kinerja pada Data Tidak Seimbang (RQ3)

Dataset juga dikelompokkan berdasarkan keseimbangan kelas (Tabel 5). Pada dataset seimbang (Heart Disease, Breast Cancer, CKD), FM menunjukkan *macro F1-score* yang lebih tinggi dibandingkan GBM. TabPFN v2 dan TabICL mencapai sekitar 0,940, sementara CatBoost sebagai

GBM terbaik memperoleh 0,925. Pada dataset tidak seimbang (Pima Diabetes, Car Evaluation, Wine Quality, Adult Income, Bank Marketing), perbedaannya jauh lebih kecil. *Macro F1-score* FM berada pada rentang 0,760–0,766, sedangkan GBM berada pada 0,754–0,755.

Tabel 5. Mean Macro F1-Score by Class Balance Group. Bold = best per group.

Group	TabPFN v2	TabICL	TabDPT	XGBoost	LightGBM	CatBoost
Balanced	**0.940**	**0.940**	0.938	0.918	0.916	0.925
Imbalanced	0.762	**0.766**	0.760	0.755	0.754	0.755

Hasil ini menunjukkan bahwa FM belum memberikan keunggulan yang jelas dalam menangani ketidakseimbangan kelas. Pada konfigurasi bawaan, kinerja kedua kelompok model relatif berdekatan. Karena itu, penanganan ketidakseimbangan kelas tetap perlu dipertimbangkan secara khusus, terlepas dari kelompok model yang digunakan [31], [32].

3.5. Stabilitas Kinerja Antar-Seed (RQ4)

Stabilitas model dilihat dari simpangan baku akurasi pada berbagai *seed* dan *fold*. Secara umum, perbedaannya relatif kecil. Nilai rata-rata simpangan baku berkisar antara 0,0145 hingga 0,0165, yang menunjukkan bahwa keenam model memiliki tingkat stabilitas yang cukup serupa dalam pengaturan eksperimen ini.

Variasi terbesar justru muncul pada dataset tertentu, terutama Heart Disease dan Pima Diabetes, dengan simpangan baku sekitar 0,03-0,05 pada hampir seluruh model. Hal ini menunjukkan bahwa karakteristik dan tingkat kesulitan dataset tampaknya lebih berpengaruh terhadap variasi hasil dibandingkan kelompok model yang digunakan.

3.6. Panduan Praktis Pemilihan Model (RQ5)

Berdasarkan keseluruhan hasil, TabPFN v2 dan TabICL menjadi pilihan yang kuat untuk dataset kecil ketika *hyperparameter tuning* tidak dilakukan. Keduanya secara konsisten berada pada peringkat atas dan menunjukkan keunggulan terbesar dibandingkan GBM pada kelompok dataset kecil.

CatBoost dapat menjadi pilihan yang kompetitif ketika data memiliki banyak fitur kategorikal atau struktur yang jelas. Model ini merupakan satu-satunya GBM yang tidak menunjukkan perbedaan signifikan dengan ketiga FM berdasarkan uji Nemenyi. Sementara itu, pada data yang tidak seimbang, perbedaan antara FM dan GBM relatif kecil. Dalam kondisi tersebut, strategi penanganan ketidakseimbangan kelas menjadi pertimbangan yang lebih penting daripada sekadar memilih salah satu kelompok model.

3.7. Pembahasan

Secara keseluruhan, FM menempati peringkat lebih tinggi dibandingkan GBM. Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa tabular foundation model dapat memberikan kinerja yang kompetitif pada berbagai *task* [15], [16], [33]. Namun, CatBoost menunjukkan pola yang menarik. Walaupun berada di peringkat keempat, perbedaannya dengan ketiga FM tidak signifikan berdasarkan uji Nemenyi. Artinya, keunggulan FM terhadap GBM tidak berlaku sama kuat untuk semua model, terutama ketika CatBoost digunakan sebagai pembanding.

Pengecualian terlihat pada Car Evaluation, ketika CatBoost justru mengungguli seluruh FM. Hasil ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa model berbasis pohon tetap kuat pada dataset berdimensi rendah dengan struktur aturan yang jelas [3], [11].

Keunggulan FM paling jelas pada kelompok dataset kecil dan cenderung menyempit pada kelompok yang lebih besar. Pola ini sejalan dengan karakteristik in-context learning yang memungkinkan FM bekerja dengan baik ketika data latih terbatas [2], [34]. Namun, karena penelitian ini hanya mencakup dataset hingga 5.000 sampel, temuan tersebut belum dapat digeneralisasikan ke dataset yang jauh lebih besar.

Pada data tidak seimbang, *macro F1-score* kedua kelompok model hanya berbeda sedikit. Temuan ini menunjukkan bahwa keunggulan FM yang terlihat pada evaluasi secara umum tidak banyak terlihat ketika kelas dalam dataset tidak seimbang. Strategi seperti pembobotan kelas atau *resampling* tetap relevan untuk kedua kelompok model. Namun studi yang menggunakan *benchmark* yang lebih besar menemukan bahwa FM tetap dapat menunjukkan performa yang kompetitif [16], [35].

Stabilitas kinerja juga tidak menunjukkan perbedaan yang besar antar-model. Variasi yang lebih tinggi justru muncul pada dataset tertentu, terutama Heart Disease dan Pima Diabetes. Temuan ini mengindikasikan bahwa karakteristik dataset dapat lebih menentukan stabilitas hasil dibandingkan pilihan antara FM dan GBM.

3.8. Keterbatasan

Penelitian ini memiliki beberapa keterbatasan. Pertama, seluruh model diuji menggunakan konfigurasi bawaan. XGBoost dan LightGBM dapat memperoleh peningkatan kinerja melalui *hyperparameter tuning*, sehingga hasil penelitian ini perlu dipahami sebagai perbandingan dalam kondisi tanpa *tuning*. Kedua, ukuran dataset dibatasi hingga 5.000 sampel setelah *subsampling*, sehingga kinerja model pada dataset yang lebih besar belum dapat disimpulkan. Ketiga, meskipun dataset yang digunakan berasal dari beberapa domain, penelitian ini belum mencakup data berdimensi sangat tinggi (>100 fitur) maupun bentuk data tabular khusus seperti *electronic health records* atau data deret waktu yang telah diagregasi.

4. Kesimpulan

Penelitian ini membandingkan tiga *Tabular Foundation Model* (TabPFN v2, TabICL, dan TabDPT) dengan tiga metode *Gradient Boosting* (XGBoost, LightGBM, dan CatBoost) pada delapan dataset *benchmark* menggunakan konfigurasi bawaan. Hasil pengujian menunjukkan adanya perbedaan kinerja yang signifikan antar-model ($p < 0{,}001$). Secara umum, ketiga FM menempati tiga peringkat teratas dan memberikan hasil terbaik pada tujuh dari delapan dataset. Satu-satunya pengecualian adalah Car Evaluation, ketika CatBoost memperoleh hasil terbaik. CatBoost juga menjadi GBM yang paling kompetitif karena kinerjanya tidak berbeda secara signifikan dari ketiga FM berdasarkan uji Nemenyi. Sementara itu, pada data yang tidak seimbang, perbedaan kinerja antara FM dan GBM relatif kecil. Stabilitas keenam model juga tidak jauh berbeda, dan variasi hasil lebih terlihat pada dataset yang memang lebih sulit.

Berdasarkan temuan tersebut, TabPFN v2 dan TabICL dapat menjadi pilihan awal untuk dataset kecil dan relatif seimbang ketika *hyperparameter tuning* tidak dilakukan. CatBoost tetap menjadi alternatif yang kuat, terutama untuk data dengan banyak fitur kategorikal. Penelitian selanjutnya dapat memperluas evaluasi pada dataset yang lebih besar, membandingkan model setelah *tuning*, serta menguji strategi khusus untuk menangani ketidakseimbangan kelas.

Daftar Pustaka

[1] V. Borisov, T. Leemann, K. Seßler, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7499–7519, 2024, doi: 10.1109/TNNLS.2022.3229161.

[2] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second," presented at the Proceedings of the 11th International Conference on Learning Representations, OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=cp5PvcI6w8_

[3] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning Is Not All You Need," Information Fusion, vol. 81, pp. 84–90, 2022, doi: 10.1016/j.inffus.2021.11.011.

[4] S. A. Fayaz, M. Zaman, S. Kaul, and M. A. Butt, "Is Deep Learning on Tabular Data Enough? An Assessment," International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022, doi: 10.14569/IJACSA.2022.0130650.

[5] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting Deep Learning Models for Tabular Data," presented at the Advances in Neural Information Processing Systems, Curran Associates, 2021, pp. 18932–18943. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html

[6] A. Kadra, M. Lindauer, F. Hutter, and J. Grabocka, "Well-Tuned Simple Nets Excel on Tabular Datasets," presented at the Advances in Neural Information Processing Systems, Curran Associates, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/c902b497eb972281fb5b4e206db38ee6-Abstract.html

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[8] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," presented at the Advances in Neural Information Processing Systems, Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," presented at the Advances in Neural Information Processing Systems, Curran Associates, 2018.

[10] A. Y. Yildiz and A. Kalayci, "Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data," presented at the 2025 IEEE Conference on AI and Data Analytics, IEEE, 2025.

[11] D. McElfresh, S. Khandagale, J. Valverde, and C. White, "When Do Neural Nets Outperform Boosted Trees on Tabular Data?," presented at the Advances in Neural Information Processing Systems, Curran Associates, 2023.

[12] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?," presented at the Advances in Neural Information Processing Systems, Curran Associates, 2022, pp. 507–520. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html

[13] G. Badaro, M. Saeed, and P. Papotti, "Transformers for Tabular Data Representation: A Survey of Models and Applications," Transactions of the Association for Computational Linguistics, vol. 11, pp. 227–249, 2023, doi: 10.1162/tacl_a_00544.

[14] S. Ö. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," presented at the Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2021, pp. 6679–6687. doi: 10.1609/aaai.v35i8.16826.

[15] N. Hollmann, S. Müller, L. Purucker, and F. Hutter, "Accurate Predictions on Small Data with a Tabular Foundation Model," Nature, vol. 637, pp. 319–326, 2025, doi: 10.1038/s41586-024-08328-6.

[16] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "TabICL: A Tabular Foundation Model for In-Context Learning on Large Data," in Proceedings of the 42nd International Conference on Machine Learning, PMLR, 2025, pp. 50817–50847. [Online]. Available: https://proceedings.mlr.press/v267/qu25d.html

[17] J. Ma, "TabDPT: Scaling Tabular Foundation Models on Real Data."

[18] J. Luo, Y. Quan, and S. Xu, "Robust-GBDT: Leveraging Robust Loss for Noisy and Imbalanced Classification with GBDT," Knowledge and Information Systems, vol. 67, pp. 2481–2510, 2025, doi: 10.1007/s10115-024-02290-3.

[19] H.-J. Ye, S.-Y. Liu, H.-R. Cai, Q.-L. Zhou, and D.-C. Zhan, "A Closer Look at Deep Learning Methods on Tabular Datasets."

[20] A. Shmuel, O. Glickman, and T. Lazebnik, "A Comprehensive Benchmark of Machine and Deep Learning Across Diverse Tabular Datasets," arXiv preprint arXiv:2408.14817, 2024, doi: 10.48550/arXiv.2408.14817.

[21] B. Amirshahi and S. Lahmiri, "Bankruptcy Prediction Using Optimal Ensemble Models Under Balanced and Imbalanced Data," Expert Systems, vol. 41, no. 5, p. e13444, 2024, doi: 10.1111/exsy.13444.

[22] Sarwo and Y. D. Prabowo, "Enhancing Classification Performance Through the Synergistic Use of XGBoost, TabPFN, and LGBM Models," presented at the Proceedings of the 2023 15th International Congress on Advanced Applied Informatics Winter, IEEE, 2023. doi: 10.1109/IIAI-AAI-Winter60127.2023.

[23] R. Detrano, "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease," American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989, doi: 10.1016/0002-9149(89)90524-9.

[24] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," presented at the Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1988, pp. 261–265.

[25] P. Soundarapandian, L. Rubini, and P. Eswaran, "Chronic Kidney Disease Dataset," UCI Machine Learning Repository. University of California, Irvine, 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[26] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction for Breast Tumor Diagnosis," presented at the Proceedings of SPIE, SPIE, 1993, pp. 861–870. doi: 10.1117/12.148698.

[27] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling Wine Preferences by Data Mining from Physicochemical Properties," Decision Support Systems, vol. 47, no. 4, pp. 547–553, 2009, doi: 10.1016/j.dss.2009.05.016.

[28] M. Bohanec, "Car Evaluation Dataset," UCI Machine Learning Repository. University of California, Irvine, 1997. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/car+evaluation

[29] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," presented at the Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 202–207.

[30] S. Moro, P. Cortez, and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing," Decision Support Systems, vol. 62, pp. 22–31, 2014, doi: 10.1016/j.dss.2014.03.001.

[31] D.-H. Won, K.-S. Shin, and S. Youm, "Synthetic Data Augmentation for Imbalanced Tabular Data: A Comparative Study of Generation Methods," Electronics, vol. 15, no. 4, p. 883, 2026, doi: 10.3390/electronics15040883.

[32] R. Kanász, P. Drotar, P. Gnip, and M. Zoricák, "Clash of Titans on Imbalanced Data: TabNet vs XGBoost," presented at the Proceedings of the 2024 IEEE Conference on Artificial Intelligence, IEEE, 2024. doi: 10.1109/CAI59869.2024.

[33] R. A. Babu, V. Vishwa Priya, M. K. Mishra, and B. Swarna, "Transformer-Based Tabular Foundation Models: Outperforming Traditional Methods with TabPFN," International Journal of Engineering, Science and Information Technology, vol. 5, no. 1, pp. 112–119, 2025.

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 2]

[34] A. P. De Rosa, A. D'Ambrosio, C. Marzi, and F. Esposito, "XGBoost vs. TabPFN in Neuroimaging Machine Learning-Based Analysis," presented at the Convegno Nazionale di Bioingegneria, 2023.

[35] Y. Zeng, T. Dinh, W. Kang, and A. C. Muller, "TABFLEX: Scaling Tabular Learning to Millions with Linear Attention," presented at the Proceedings of Machine Learning Research, PMLR, 2025.