JCOSIS (*Journal Computer Science and Information Syetem*)
E-ISSN : 3032-5676
Volume : 03 | Nomor : 02 | Tahun 2026 | Hal 77-85
DOI : https://doi.org/10.61567

Multi-Task Learning Berbasis Seleksi Fitur Stabil untuk Klasifikasi Komplikasi *Mikrovaskular Diabetes*

Hamzah Setiawan[1]*, Ade Eviyanti[2]
[1,2]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo, Indonesia
[1] hamzah@umsida.ac.id; [2] adeeviyanti@umsida.ac.id

Abstrak.
Tujuan : Mengevaluasi pengaruh seleksi fitur berbasis stabilitas dan pembobotan kelas terhadap klasifikasi retinopati, neuropati, dan nefropati diabetik menggunakan *multi-task learning* (MTL).
Metode/Design/Pendekatan: Dari 100.000 rekam dataset publik, 27.407 rekam berstatus diabetes positif dianalisis. Sebanyak 18 dari 41 prediktor numerik dipilih menurut frekuensi ranking ANOVA F pada 20 subsampel pelatihan. Model logistic regression dan tiga konfigurasi MTL diuji dengan pembagian train/validasi/test sebesar 19.184/4.111/4.112 rekam. Average precision (AP) digunakan untuk setiap label, sedangkan macro-AUPRC merupakan rata-rata AP ketiga label. Hasil MTL dilaporkan sebagai rerata ± simpangan baku dari tiga seed.
Hasil/Temuan: MTL 41 fitur memperoleh macro-AUPRC tertinggi (0,1485 ± 0,0028), dibandingkan MTL 18 fitur (0,1454 ± 0,0024) dan MTL 18 fitur dengan weighted BCE (0,1449 ± 0,0043). AP per label pada MTL 41 fitur adalah 0,1440 (retinopati), 0,1847 (neuropati), dan 0,1167 (nefropati). Pengurangan prediktor sebesar 56,1% belum memperbaiki kinerja rata-rata.
Kebaharuan/Originalitas/Nilai: Studi ablasi pada tiga label komplikasi secara bersama memperlihatkan bahwa konsistensi seleksi fitur pada subsampel tidak menjamin peningkatan macro-AUPRC. Evaluasi satu pembagian data belum membuktikan kesiapan klinis.

Kata Kunci: diabetes; komplikasi mikrovaskular; *multi-task learning*; seleksi fitur berbasis frekuensi *resampling*; klasifikasi multilabel.

Abstract.
Purpose: *To evaluate stability-based feature selection and class weighting for multi-task learning (MTL) classification of diabetic retinopathy, neuropathy, and nephropathy.*
Methods/Study design/approach: *From 100,000 public-dataset records, 27,407 diabetes-positive records were included. ANOVA F ranking across 20 training subsamples selected 18 of 41 numeric predictors. Logistic regression and three MTL configurations were assessed using train/validation/test partitions of 19,184/4,111/4,112 records. Average precision (AP) quantified performance for each label; macro-AUPRC was the unweighted mean of the three label-wise AP scores. MTL results are reported as mean ± standard deviation across three seeds.*
Result/Findings: *Full-feature MTL had the highest macro-AUPRC (0.1485 ± 0.0028), compared with selected-feature MTL (0.1454 ± 0.0024) and selected-feature MTL with weighted BCE (0.1449 ± 0.0043). Label-wise AP for full-feature MTL was 0.1440 for retinopathy, 0.1847 for neuropathy, and 0.1167 for nephropathy. Reducing predictors by 56.1% did not improve mean classification performance.*
Novelty/Originality/Value: *This three-label ablation study shows that consistently selected features do not necessarily improve macro-AUPRC. A single internal split cannot establish clinical readiness.*

Keywords*: diabetes; microvascular complications; multi-task learning; resampling-frequency feature selection; multilabel classification.*

Article history:

Received, 2026-10-19
Revised, 2026-10-22
Accepted, 2026-10-28

[1]*Corresponding author.
Hamzah Setiawan,
Email addresses: hamzah@umsida.ac.id

This is an open access article under the CC BY-SAlicense.

PENDAHULUAN

Komplikasi mikrovaskular diabetes mencakup *retinopati*, *neuropati*, dan *nefropati*. Ketiganya memiliki manifestasi dan faktor risiko yang saling berkaitan, sehingga klasifikasi masing-masing komplikasi memerlukan perhatian terhadap karakteristik label dan kondisi data. Kanbour et al. [1] dan Sambyal et al. [2] merangkum pendekatan pembelajaran mesin untuk komplikasi tersebut, sedangkan Xiao et al. [3] mengkaji faktor risiko menggunakan *gradient boosting*. Schallmoser et al. [4] dan He et al. [5] menelaah pengembangan dan evaluasi model menggunakan rancangan data berbeda, termasuk validasi eksternal. Perbedaan populasi, karakteristik masukan, definisi keluaran, dan rancangan validasi membuat nilai kinerja antarpublikasi tidak dapat dibandingkan secara langsung. Berbeda dari penelitian longitudinal, eksperimen ini berfokus pada klasifikasi status tiga komplikasi dari rekam dataset publik yang tersedia, bukan prediksi kejadian masa mendatang karena waktu onset tidak diketahui.

Dari sisi metode, Caruana [6] serta Zhang dan Yang [7] menjelaskan dasar pembelajaran representasi bersama pada multi-task learning (MTL). Kim et al. [8] menggunakan MTL untuk mempelajari faktor risiko bersama dan khusus pada komplikasi diabetes. Meinshausen dan Bühlmann [9] serta Nogueira et al. [10] menguraikan stabilitas pemilihan fitur, sedangkan Zhang et al. [11] mengembangkan pendekatan seleksi fitur bersama dan spesifik tugas berbasis regularisasi. Pendekatan tersebut menjawab persoalan yang berbeda: pemanfaatan informasi antartugas, konsistensi pemilihan fitur, dan pemilihan fitur untuk beberapa tugas. Namun, konsistensi sebuah subset fitur pada pengambilan sampel berulang tidak dengan sendirinya menunjukkan bahwa subset tersebut meningkatkan kinerja klasifikasi multitugas.

Pada data dengan kasus positif yang jarang, evaluasi kinerja perlu mempertimbangkan ketidakseimbangan label. Krawczyk [12] membahas tantangan pembelajaran pada kelas tidak seimbang, dan Saito serta Rehmsmeier [13] menunjukkan kelebihan *precision-recall* untuk mengevaluasi identifikasi kelas positif. Telaah Tan et al. [14] menekankan perlunya evaluasi kritis model komplikasi diabetes, sedangkan panduan TRIPOD+AI oleh Collins et al. [15] memberikan konteks prinsip transparansi pelaporan model kesehatan. Literatur tersebut menjadi alasan metodologis untuk melaporkan *average precision* (AP) pada setiap label dan macro-AUPRC, tanpa menyamakan hasil eksperimen komputasional dengan bukti manfaat klinis.

Kesenjangan penelitian dan kontribusi. Literatur terdahulu telah mengevaluasi klasifikasi komplikasi diabetes dengan pembelajaran mesin [1], [2], pemanfaatan representasi bersama serta faktor risiko khusus tugas melalui MTL [8], dan pemilihan fitur pada banyak tugas maupun konsistensi pemilihannya [9], [10], [11]. Akan tetapi, dari studi yang ditelaah belum tersedia jawaban langsung mengenai apakah seleksi 18 dari 41 prediktor berdasarkan frekuensi ranking ANOVA F pada 20 subsampel data latih mampu mempertahankan atau meningkatkan kinerja MTL saat *retinopati*, *neuropati*, dan *nefropati* diklasifikasikan bersama sebagai tiga label yang tidak seimbang pada dataset yang digunakan. Penelitian ini mengisi celah evaluasi tersebut melalui studi ablasi yang membandingkan *logistic regression* dan MTL dengan fitur lengkap, MTL dengan fitur hasil resampling, serta MTL dengan fitur hasil *resampling* dan *weighted* BCE. Kontribusi yang dapat dipertanggungjawabkan ialah bukti empiris mengenai kompromi antara reduksi jumlah prediktor, konsistensi pemilihan internal, dan AP per label serta macro-AUPRC pada satu pembagian dataset publik. Penelitian ini tidak mengklaim algoritma seleksi baru, perbaikan performa yang universal, atau manfaat klinis yang telah teruji.

Pemilihan fitur berulang dilakukan hanya pada data latih untuk menghindari penggunaan informasi dari data validasi dan pengujian. Frekuensi fitur terpilih digunakan sebagai ukuran kestabilan empiris; frekuensi tersebut tidak menyediakan jaminan kontrol kesalahan dari *stability selection formal* [9], [10]. Risiko kebocoran informasi juga diperiksa mengacu pada Kaufman et al. [16].

Metrik utama adalah macro-AUPRC, yaitu rata-rata tak berbobot AP dari *retinopati*, *neuropati*, dan *nefropati*. Setiap AP merangkum *precision* pada tingkat *recall* yang berbeda untuk satu label; setiap komplikasi memperoleh bobot evaluasi yang sama walaupun prevalensinya berbeda. Pada data penelitian ini, proporsi positif masing-masing label hanya 7,11%, 9,49%, dan 5,86%. Akurasi tidak dijadikan metrik primer karena model yang selalu memilih kelas negatif dapat memperoleh akurasi tinggi tanpa menemukan satu pun kasus positif. AUROC tetap bermanfaat sebagai metrik diskriminasi, tetapi pada prevalensi rendah dapat terlihat baik meskipun *precision* untuk kasus positif masih terbatas. Penilaian precision-recall lebih langsung menunjukkan kompromi antara temuan positif benar dan salah [13]. Macro-AUPRC dilengkapi pelaporan AP per label agar hasil agregat tidak menyembunyikan perbedaan kinerja antar komplikasi.

Dataset publik yang dianalisis memiliki 100.000 rekam dan 209 variabel menurut audit berkas penelitian. Audit internal menunjukkan sebagian rekam dengan status diabetes negatif tetap memiliki label komplikasi positif. Temuan tersebut merupakan karakteristik berkas yang digunakan, bukan estimasi epidemiologis ataupun bukti

kondisi pasien nyata. Asal dataset, status sintetis, dan definisi pembentukan label belum terverifikasi sehingga interpretasi klinis harus dibatasi. Tanpa waktu onset yang valid, penelitian ini hanya menilai klasifikasi status komplikasi, bukan prognosis atau deteksi dini. Kebutuhan audit kualitas data dan pelaporan model secara transparan selaras dengan prinsip pencegahan kebocoran informasi serta pelaporan model kesehatan [16], [15].

Tujuan operasional eksperimen ini adalah (1) mengukur frekuensi pemilihan setiap prediktor melalui 20 subsampel data latih dan menentukan subset 18 fitur; (2) menghitung perubahan absolut macro-AUPRC serta AP retinopati, neuropati, dan nefropati antara MTL 41 fitur dan MTL 18 fitur serta membandingkannya dengan logistic regression; dan (3) mengukur perubahan ketiga AP dan macro-AUPRC ketika weighted BCE diterapkan pada MTL 18 fitur. Seluruh pengujian merupakan klasifikasi multilabel status komplikasi pada rekam diabetes positif. Penelitian tidak menguji prediksi onset komplikasi, signifikansi statistik selisih model, ataupun kesiapan penerapan klinis.

METODE PENELITIAN

Eksperimen komputasional ini menggunakan berkas diabetes_large_dataset(1).csv yang diunggah untuk audit, berisi 100.000 rekam dan 209 kolom. Berdasarkan kolom diabetes_status, terdapat 27.407 rekam positif, 70.573 negatif, serta 2.020 nilai kosong. Populasi analisis dibatasi pada 27.407 rekam dengan diabetes_status = 1. Identitas repositori asal, URL/DOI, lisensi, tanggal pengambilan asli, dan status sintetis atau observasional tidak terdapat dalam isi CSV dan tidak dapat diverifikasi hanya melalui berkas ini. Informasi tersebut harus dilengkapi dari halaman unduhan atau dokumentasi penerbit data sebelum publikasi. Tidak ada kolom tanggal onset komplikasi, sehingga analisis hanya mengklasifikasikan status yang tercatat.

Tiga keluaran diperlakukan sebagai multilabel biner, dengan urutan R = diabetic_retinopathy, N = diabetic_neuropathy, F = diabetic_nephropathy (Yes = 1; No = 0). Pada populasi analisis, positif retinopati berjumlah 1.949 (7,11%), neuropati 2.601 (9,49%), dan nefropati 1.607 (5,86%); tidak terdapat nilai label kosong. Audit kombinasi label menghasilkan 000 = 22.178, 001 = 1.054, 010 = 1.922, 011 = 304, 100 = 1.376, 101 = 198, 110 = 324, dan 111 = 51 rekam. Sebanyak 877 rekam (3,20%) memiliki dua atau tiga komplikasi secara bersamaan, termasuk 51 dengan ketiganya. Stratifikasi menggunakan kode tiga bit tersebut dengan random_state = 42 pada pembagian train/validation/test 70%/15%/15% (19.184/4.111/4.112). Nilai seluruh strata telah dihitung dari CSV.

Definisi operasional ketiga label dalam eksperimen terbatas pada nilai *Yes*/*No* dalam kolom diabetic_retinopathy, diabetic_neuropathy, dan diabetic_nephropathy. CSV tidak menyediakan protokol pemeriksaan mata, neurologis, ataupun pemeriksaan ginjal yang menetapkan label tersebut, serta tidak menyertakan ambang diagnosis dan sumber anotasinya. Karena itu, label diperlakukan sebagai status yang sudah tercatat, bukan diagnosis yang dapat diverifikasi secara mandiri. Audit menemukan 739 label retinopati, 980 neuropati, dan 606 nefropati positif pada rekam berstatus diabetes negatif. Temuan ini memerlukan pemeriksaan dokumentasi definisi dan kualitas label, dan tidak boleh diartikan sebagai estimasi prevalensi klinis.

Kaufman et al. [16] menjelaskan bahwa data *leakage* dapat terjadi ketika prediktor memuat informasi target yang seharusnya tidak tersedia pada saat prediksi. Prinsip tersebut menjadi dasar untuk meninjau indikator diagnosis, skor risiko turunan, dan variabel yang mungkin membentuk label komplikasi sebelum pemodelan.

Audit variabel mengidentifikasi kandidat kebocoran informasi berupa patient_id dan record_id; ketiga label komplikasi; diabetes_complication_count; diabetic_foot_problem dan cardiovascular_complication; variabel diagnosis atau turunan yang berhubungan langsung dengan target, khususnya kidney_disease_status, kidney_disease, proteinuria_status, creatinine_mg_dl, egfr_ml_min_1_73m2, urine_albumin_mg_g, dan urine_albumin_creatinine_ratio; serta skor risiko turunan seperti kidney_risk_score dan diabetes_risk_score. Variabel tersebut perlu dikeluarkan atau diaudit terhadap aturan pembentukan label sebelum dilibatkan sebagai prediktor. Daftar 18 fitur akhir yang terbaca pada grafik eksperimen adalah age, weight_kg, bmi, waist_circumference_cm, body_fat_percentage, hip_circumference_cm, random_glucose_mg_dl, fasting_glucose_mg_dl, hba1c_percent, fasting_insulin_uIU_ml, post_meal_glucose_mg_dl, two_hour_glucose_mg_dl, c_peptide_ng_ml, systolic_bp, triglycerides_mg_dl, diastolic_bp, diet_quality_score, dan hdl_mg_dl. Seluruh 18 nama ditemukan dalam CSV, tetapi daftar tepat 41 kandidat awal dan bukti semua kolom yang dikeluarkan dari pipeline memerlukan skrip untuk verifikasi lengkap. Median imputasi, standardisasi dengan statistik train, lalu clipping [−8, 8] dipertahankan sebagai prosedur yang dilaporkan.

Prosedur aktual: dari train saja, dibuat B = 20 subsampel acak tanpa pengembalian, masing-masing berukuran 70% train. Pada setiap subsampel dihitung statistik ANOVA F untuk setiap fitur terhadap masing-masing dari tiga

label, skor diubah menjadi log(1 + F), kemudian dirata-ratakan pada tiga label. Delapan belas fitur dengan skor agregat teratas dipilih per subsampel. Frekuensi fitur j dihitung S_j = (1/20) Σ_b I(j ∈ F_b), lalu 18 fitur dengan frekuensi tertinggi menjadi subset final. Seluruh perhitungan pemilihan dan urutan fitur menggunakan train saja. *Validation* digunakan untuk pemantauan dan keputusan pengembangan model; test hanya untuk evaluasi akhir. Belum tersedia *seed* terpisah untuk pembangkitan 20 subsampel dan aturan penyelesaian skor seri (ties), sehingga dua rincian tersebut masih perlu dilengkapi dari skrip.

Metode aktual ini disebut seleksi fitur berbasis frekuensi resampling, bukan stability selection formal Meinshausen–Bühlmann [9]. Frekuensi tinggi menunjukkan konsistensi pemeringkatan pada subsampel train yang saling tumpang tindih, tanpa jaminan kontrol salah pilih teoretis atau stabilitas di populasi eksternal [10], [17].

$$S_j = \frac{1}{B} \times \sum_{b=1}^{B} I(f_j \in F_b) \quad (1)$$

Rancangan lanjutan yang BELUM DIUJI meliputi seleksi fitur khusus tugas dan fitur bersama secara terpisah, penentuan ambang stabilitas melalui validasi, serta penggabungan subset berdasarkan tujuan tiap label. Eksperimen aktual tidak menggunakan aturan tersebut; tidak ada hasil dari ambang, regularisasi seleksi baru, atau optimasi subset khusus tugas yang boleh diatribusikan ke penelitian ini.

Arsitektur aktual MTL memakai shared MLP dengan lapisan tersembunyi 64 dan 32 unit, aktivasi ReLU, dropout 0,15, dan tiga logit output biner. Optimizer AdamW menggunakan learning rate 0,001 dan weight decay 0,001; batch size 512; maksimal 24 epoch; early stopping patience = 5 berdasarkan macro-AUPRC validation. Seed pelatihan 11, 23, dan 37. Loss adalah rerata BCE ketiga tugas, sedangkan varian berbobot memakai BCEWithLogits dan pos_weight_k = Nnegatif,k(train)/Npositif,k(train), dihitung terpisah per label hanya dari train. Baseline logistic regression dilatih per label dengan max_iter = 400 pada 41 dan 18 fitur. Parameter selain yang disebut (misalnya momentum/betas dan epsilon optimizer, toleransi minimum peningkatan early stopping, serta pilihan solver LR) belum tercantum di naskah atau log dan perlu dicatat menurut konfigurasi kode yang benar-benar dijalankan.

$$L_{total} = \sum_{k=1}^{3} w_k \times BCE_k + \lambda \times R(\theta) \quad (2)$$

Pelatihan tiga seed dilakukan pada split train/validation/test yang sama. Untuk setiap seed s dan label k, AP_k,s dihitung pada test dari skor kontinu; macro-AUPRC_s = (AP_R,s + AP_N,s + AP_F,s)/3. Tabel melaporkan rata-rata aritmetika antar tiga seed untuk AP tiap label dan macro-AUPRC MTL; simpangan baku macro-AUPRC dihitung antarseed dan ditampilkan sebagai mean ± SD (jenis sample SD, ddof=1 atau population SD, ddof=0, perlu diverifikasi dari kode). Nilai logistic regression adalah skor dari satu proses evaluasi dan tidak mempunyai SD yang dilaporkan. Test tidak boleh dipakai untuk memilih hiperparameter, jumlah fitur, model, atau epoch terbaik; log seleksi checkpoint berdasarkan validation perlu diverifikasi sebelum menyatakan prosedur sepenuhnya diaudit. Perangkat lunak yang dapat menyusun alur ini meliputi Python, NumPy, scikit-learn, dan PyTorch, tetapi penggunaan aktual beserta nomor versinya BELUM DAPAT DIPASTIKAN tanpa environment.yml, requirements.txt, atau log; tidak dicantumkan versi rekaan.

Tabel 1. Xxx xxx

Konfigurasi	Tujuan pembandingan	Status
LR per target, 41 fitur	Baseline single-task tabular	Sudah diuji
LR per target, 18 fitur resampling	Dampak reduksi fitur pada baseline	Sudah diuji
MTL, 41 fitur	Baseline representasi bersama	Sudah diuji
MTL, 18 fitur resampling	Dampak reduksi fitur pada MTL	Sudah diuji
MTL, 18 fitur + weighted BCE	Dampak pembobotan kelas	Sudah diuji

Tabel 1 hanya memuat lima konfigurasi yang benar-benar dievaluasi dalam hasil penelitian. Pengujian XGBoost per target, MLP single-task per target, dan MTL dengan seleksi fitur sekali jalan merupakan rancangan lanjutan/belum dievaluasi dan tidak digunakan untuk menyimpulkan keunggulan model. Pembagian data ditetapkan terlebih dahulu; seluruh fit imputasi, penskalaan, ranking ANOVA F, dan pos_weight menggunakan train saja. Validation digunakan untuk early stopping dan keputusan pengembangan, sedangkan test untuk pelaporan akhir. Kepatuhan penuh terhadap protokol ini masih memerlukan pemeriksaan skrip/log asli.

Definisi metrik: untuk label k, precision_k(t) = TP_k(t)/(TP_k(t)+FP_k(t)) dan recall_k(t) = TP_k(t)/(TP_k(t)+FN_k(t)). Average precision dihitung sebagai AP_k = Σm (R_k,m − R_k,m−1) P_k,m, dengan titik m diurutkan menurut ambang skor dan P_k,m precision pada titik recall R_k,m. Macro-AUPRC =

(AP_retinopati + AP_neuropati + AP_nefropati)/3. Istilah macro-AUPRC dalam artikel ini mengacu secara operasional pada mean average precision tiga label, bukan otomatis identik dengan luas kurva PR hasil interpolasi trapesium. Implementasi yang dapat mereproduksi definisi tersebut di Python adalah sklearn.metrics.average_precision_score(y_true[:, k], y_score[:, k]) pada tiap label dan numpy.mean(AP_per_label. Skor model berupa probabilitas positif atau nilai keputusan kontinu, bukan label biner yang sudah di-threshold.

HASIL DAN PEMBAHASAN

Temuan audit deskriptif awal

Audit CSV mengonfirmasi 100.000 rekam dan 209 kolom, serta 27.407 rekam dengan diabetes_status = 1. Jumlah label positif pada kelompok ini adalah 1.949 retinopati (7,11%), 2.601 neuropati (9,49%), dan 1.607 nefropati (5,86%). Sebanyak 877 rekam memiliki dua atau tiga label positif; distribusi delapan kombinasi tercantum pada metode. Rekam diabetes_status = 0 juga mengandung label komplikasi positif: 739 retinopati, 980 neuropati, dan 606 nefropati. Fakta ini menekankan perlunya audit asal dataset dan aturan anotasi, bukan membuktikan ketidaksesuaian medis tanpa dokumentasi sumber.

Tabel 2. Distribusi label komplikasi pada rekam diabetes positif

Target komplikasi	Jumlah positif (n = 27.407)	Prevalensi positif (%)
Retinopati	1.949	7,11%
Neuropati	2.601	9,49%
Nefropati	1.607	5,86%

Audit seluruh dataset menemukan komplikasi positif pada rekam yang memiliki diabetes_status = 0: retinopati 739, neuropati 980, dan nefropati 606. Keadaan ini memerlukan verifikasi definisi label dan provenance. Eksperimen utama mengecualikan status diabetes 0 dan kosong, tetapi pembatasan tersebut tidak menyelesaikan seluruh ketidakpastian kualitas data.

Hasil perbandingan model klasifikasi

Temuan utama studi ablasi adalah bahwa MTL dengan seluruh 41 prediktor memperoleh macro-AUPRC tertinggi, yaitu 0,1485 ± 0,0028, dibandingkan MTL 18 fitur hasil seleksi frekuensi resampling (0,1454 ± 0,0024) dan MTL 18 fitur dengan weighted BCE (0,1449 ± 0,0043). Hasil ini menunjukkan bahwa pada split internal yang digunakan, representasi bersama MTL memperoleh informasi yang lebih mendukung pemeringkatan label positif ketika seluruh prediktor tetap disertakan. Pengurangan 56,1% fitur menghasilkan model dengan masukan lebih ringkas, tetapi tidak meningkatkan macro-AUPRC. Angka tersebut merupakan hasil deskriptif dan tidak mendukung klaim keunggulan yang signifikan secara statistik.

Tabel 3. Perbandingan average precision (AP) dan macro-AUPRC pada data pengujian

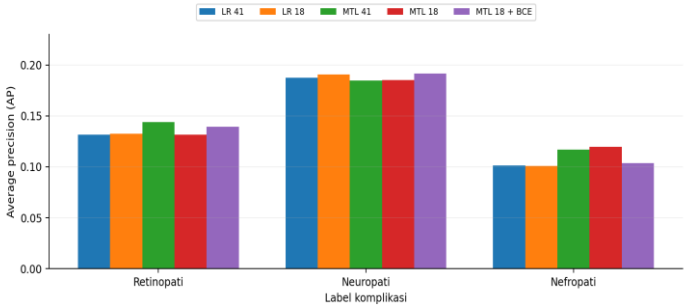
Model	Macro-AUPRC	AP Retinopati	AP Neuropati	AP Nefropati
LR (41 fitur)	0,1402	0,1316	0,1873	0,1015
LR (18 fitur)	0,1413	0,1323	0,1905	0,1011
MTL (41 fitur)	0,1485 ± 0,0028	0,1440	0,1847	0,1167
MTL (18 fitur)	0,1454 ± 0,0024	0,1316	0,1850	0,1196
MTL (18 fitur, weighted BCE)	0,1449 ± 0,0043	0,1393	0,1916	0,1038

MTL 41 fitur melampaui LR 41 fitur sebesar 0,0083 dan LR 18 fitur sebesar 0,0072 pada macro-AUPRC. Keunggulan numerik atas MTL 18 fitur sebesar 0,0031, sedangkan atas MTL 18 fitur berbobot sebesar 0,0036. Secara praktis, selisih yang kecil ini menggambarkan kompromi antara pengurangan jumlah masukan dan kemampuan pemeringkatan rekam berlabel positif. Belum ada interval kepercayaan selisih antarmodel atau uji statistik; simpangan baku dari tiga seed hanya mencerminkan variasi pelatihan pada split test yang sama. Oleh sebab itu, perbedaan numerik belum dapat dinyatakan signifikan. Tabel 4 menyajikan pembanding baseline AP yang berbasis prevalensi label untuk membantu interpretasi besaran AP, bukan sebagai ukuran manfaat klinis.

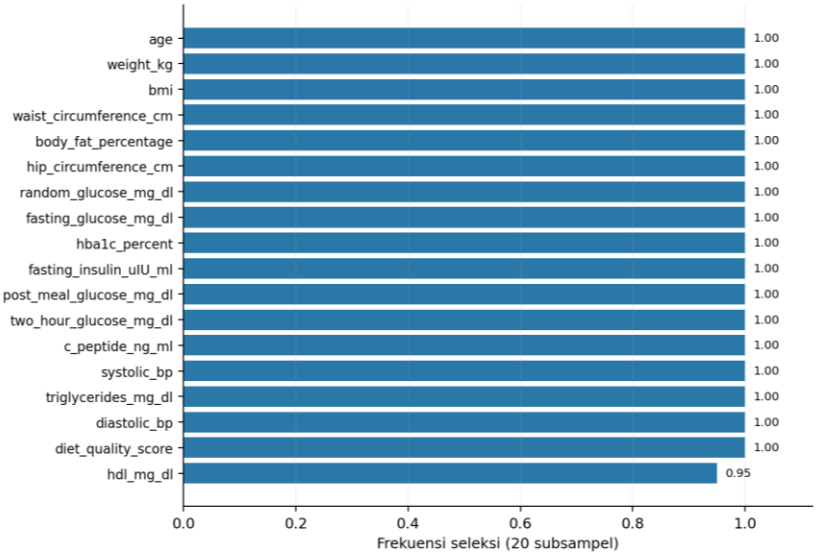
Tabel 4. AP MTL 41 fitur dibanding baseline prevalensi kelas positif (baseline acak).

Label	Prevalensi / baseline AP	AP MTL 41 fitur	Selisih absolut
Retinopati	0,0711	0,1440	+0,0729
Neuropati	0,0949	0,1847	+0,0898
Nefropati	0,0586	0,1167	+0,0581

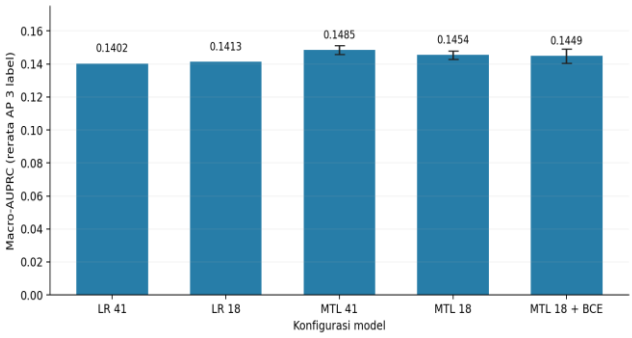
Dibandingkan penelitian terdahulu, kontribusi hasil ini bersifat metodologis dan terbatas pada dataset yang digunakan. Kim et al. [8] mengeksplorasi faktor risiko bersama dan spesifik melalui MTL, sementara Zhang et al. [11] mengembangkan pendekatan seleksi fitur multitugas berbasis regularisasi. Penelitian sekarang tidak menguji metode yang sama, tetapi menilai apakah pemilihan ulang fitur dengan ANOVA F memberikan manfaat setelah diterapkan sebelum MTL. Hasil ablasi justru memperlihatkan pengurangan 56,1% prediktor dengan penurunan macro-AUPRC sebesar 0,0031 terhadap MTL 41 fitur. Studi Schallmoser et al. [4] dan He et al. [5] menggunakan rancangan sumber data atau validasi berbeda sehingga nilai performanya tidak layak dibandingkan langsung. Dengan demikian, hasil penelitian ini lebih tepat diposisikan sebagai bukti empiris tentang pertukaran antara parsimoni fitur, kestabilan seleksi, dan kinerja klasifikasi internal, bukan sebagai bukti peningkatan hasil klinis. Analisis SHAP [18] dan studi Ahmed et al. [19] dapat menjadi konteks interpretasi fitur maupun komplikasi, tetapi penelitian ini tidak melaporkan analisis kausal.



Gambar 1. AP setiap komplikasi pada lima konfigurasi yang benar-benar diuji. Batang LR menyajikan satu hasil evaluasi tiap label; batang MTL menyajikan rerata AP dari tiga seed pelatihan (11, 23, 37) pada test set yang sama. Tidak ada batang galat per label karena simpangan bakunya tidak tersedia. Sumber: hasil pengolahan sendiri dari Tabel 3.



Gambar 2. Frekuensi seleksi 18 fitur akhir dari 20 subsampel train (70% train per subsampel). Sumbu horizontal menunjukkan proporsi pemilihan: 1,00 = 20/20 dan 0,95 = 19/20. Angka merupakan frekuensi teramati tanpa estimasi galat. Sumber: hasil pengolahan sendiri dari seleksi ANOVA F berbasis resampling.



Gambar 3. Perbandingan macro-AUPRC lima konfigurasi pada test set. Dua skor LR berasal dari satu evaluasi; tiga skor MTL merupakan rerata tiga seed, dengan batang galat ±1 simpangan baku antarseed. Batang galat bukan interval kepercayaan dan tidak menggambarkan signifikansi. Sumber: hasil pengolahan sendiri dari Tabel 3.

Tabel 5. Nama dan frekuensi seleksi 18 fitur akhir pada 20 subsampel data latih.

No.	Nama fitur	Terpilih / 20	Frekuensi
1	age	20	1,00
2	weight_kg	20	1,00
3	bmi	20	1,00
4	waist_circumference_cm	20	1,00
5	body_fat_percentage	20	1,00
6	hip_circumference_cm	20	1,00
7	random_glucose_mg_dl	20	1,00
8	fasting_glucose_mg_dl	20	1,00
9	hba1c_percent	20	1,00
10	fasting_insulin_uIU_ml	20	1,00
11	post_meal_glucose_mg_dl	20	1,00
12	two_hour_glucose_mg_dl	20	1,00
13	c_peptide_ng_ml	20	1,00
14	systolic_bp	20	1,00
15	triglycerides_mg_dl	20	1,00
16	diastolic_bp	20	1,00
17	diet_quality_score	20	1,00
18	hdl_mg_dl	19	0,95

Tabel 5 memperinci 18 fitur akhir dan frekuensinya, selaras dengan Gambar 2. Sebanyak 17 fitur tercatat pada semua 20 subsampel (1,00), sedangkan hdl_mg_dl muncul pada 19 subsampel (0,95). Setiap subsampel memuat 70% rekam dari partisi train yang sama, sehingga antarsubsampel saling tumpang tindih. Frekuensi ini hanya menunjukkan konsistensi internal pemeringkatan ANOVA F; frekuensi tinggi tidak membuktikan hubungan kausal, kinerja lebih baik, atau kestabilan pada populasi dan sumber data independen. Bukti generalisasi memerlukan repeated independent splits dan validasi eksternal [9], [10].

Respons terhadap pembobotan kelas berbeda menurut target. Dibandingkan MTL 18 fitur tanpa bobot, weighted BCE meningkatkan AP retinopati dari 0,1316 menjadi 0,1393 (Δ = +0,0077) dan AP neuropati dari 0,1850 menjadi 0,1916 (Δ = +0,0066), tetapi menurunkan AP nefropati dari 0,1196 menjadi 0,1038 (Δ = −0,0158). Penurunan nefropati lebih besar secara absolut daripada peningkatan pada kedua target lain, sehingga macro-AUPRC turun dari 0,1454 menjadi 0,1449 (Δ = −0,0005). Secara praktis, peningkatan kemampuan pemeringkatan rekam positif pada neuropati dapat dibarengi penurunan kemampuan serupa pada nefropati; evaluasi agregat saja akan menutupi perbedaan ini. Karena tidak tersedia interval kepercayaan atau uji perbedaan AP per label, perubahan tersebut belum tentu signifikan secara statistik. Hasil ini tidak membenarkan klaim bahwa weighted BCE atau strategi penyeimbangan tertentu selalu lebih baik untuk semua komplikasi [12], [14].

Studi ablasi yang sudah dijalankan membandingkan logistic regression dengan dan tanpa seleksi, MTL seluruh fitur, MTL terpilih, serta MTL terpilih dengan loss berbobot. Belum dilakukan pencarian sistematis terhadap jumlah fitur, bentuk regularisasi stabilitas, ukuran encoder, atau parameter loss. Pengujian tambahan tersebut direkomendasikan sebelum pengajuan ke jurnal.

Gambar 1 menunjukkan bahwa perubahan AP akibat reduksi fitur tidak seragam: dibanding MTL 41 fitur, AP MTL 18 fitur berubah sebesar −0,0124 untuk retinopati, +0,0003 untuk neuropati, dan +0,0029 untuk nefropati. Pada weighted BCE, AP retinopati dan neuropati meningkat terhadap MTL 18 fitur tanpa bobot, tetapi AP

nefropati menurun. Gambar 3 memperlihatkan bahwa selisih tersebut tidak menghasilkan peningkatan macro-AUPRC. Penilaian AP per label lebih tepat daripada satu skor agregat semata untuk membaca konsekuensi model pada kelas positif yang jarang [12], [13].

Ancaman validitas dan kelayakan klinis

Dalam Tabel 4, AP MTL 41 fitur melebihi baseline prevalensi sebesar 0,0729 untuk retinopati, 0,0898 untuk neuropati, dan 0,0581 untuk nefropati. Baseline prevalensi merupakan nilai harapan AP pemeringkatan acak, bukan kinerja alat diagnosis. Walaupun ketiga AP sekitar dua kali prevalensi masing-masing, precision positif tetap terbatas jika dilihat dari nilai AP absolut (0,1440; 0,1847; 0,1167). Kinerja operasional, seperti precision pada ambang tindakan tertentu, kalibrasi, konsekuensi false positive/false negative, dan manfaat klinis tidak dinilai. Ancaman validitas meliputi asal data dan aturan pembentukan label yang belum terverifikasi, potensi target leakage, satu split internal, dan tidak adanya validasi eksternal. Prosedur yang dilaporkan membatasi imputasi, standardisasi, seleksi fitur, dan pos_weight pada train serta early stopping pada validation; kepatuhan terhadap larangan memilih parameter dari test tetap perlu diverifikasi dengan skrip atau log.

Dataset tidak menyediakan informasi waktu onset komplikasi yang dapat dipastikan. Oleh karena itu, studi ini hanya mengevaluasi klasifikasi status komplikasi pada rekam yang tersedia, bukan prediksi kejadian di masa depan, prognosis, atau deteksi dini. Pemanfaatan model sebagai pendukung keputusan klinis masih merupakan kemungkinan penelitian lanjutan, bukan rekomendasi penerapan saat ini. Kajian berikutnya perlu memverifikasi asal dan definisi label, mengevaluasi data independen melalui validasi eksternal, menguji kalibrasi serta kesesuaian pada kelompok pasien yang berbeda, dan melibatkan tenaga kesehatan dalam penilaian manfaat serta risikonya. Persyaratan pelaporan dan validasi yang transparan sejalan dengan Collins et al. [15], sedangkan studi validasi eksternal He et al. [5] menjadi konteks penelitian lanjutan, bukan tahapan yang telah dilaksanakan dalam eksperimen ini.

KESIMPULAN

Evaluasi klasifikasi multilabel pada 27.407 rekam berstatus diabetes positif menghasilkan tiga temuan utama. Pertama, MTL dengan 41 fitur memperoleh macro-AUPRC tertinggi sebesar 0,1485 ± 0,0028 pada pembagian data internal yang diuji. Kedua, seleksi fitur berbasis frekuensi resampling mengurangi prediktor dari 41 menjadi 18 (56,1%), tetapi macro-AUPRC menurun menjadi 0,1454 ± 0,0024, atau lebih rendah 0,0031 secara absolut. Ketiga, penerapan weighted BCE pada MTL 18 fitur menghasilkan macro-AUPRC 0,1449 ± 0,0043 dan belum memberikan manfaat yang konsisten karena perubahan AP berbeda antara retinopati, neuropati, dan nefropati. Perbedaan skor ini bersifat deskriptif dan belum diuji signifikansinya. Karena tidak tersedia waktu onset komplikasi, hasil penelitian hanya mendukung klasifikasi status pada rekam yang dianalisis, bukan prediksi dini atau prognosis, serta belum menunjukkan kesiapan penggunaan klinis. Penelitian lanjutan diprioritaskan secara berurutan untuk memverifikasi asal dataset dan definisi label, mengevaluasi model dengan pembagian data berulang, melaporkan ketidakpastian estimasi serta selisih kinerja, kemudian melakukan validasi eksternal pada data independen.

REFERENSI

[1] S. Kanbour et al., “Machine learning models for prediction of diabetic microvascular complications,” Journal of Diabetes Science and Technology, vol. 18, no. 2, pp. 273–286, 2024, doi: 10.1177/19322968231223726.

[2] N. Sambyal, P. Saini, and R. Syal, “A review of statistical and machine learning techniques for microvascular complications in type 2 diabetes,” Current Diabetes Reviews, vol. 17, no. 2, pp. 143–155, 2021, doi: 10.2174/1573399816666200511003357.

[3] M. Xiao, Y. Fu, Y. Li, Q. Liu, X. Qiao, H. Zhang, X. Zhu, and J. Wang, “Machine learning screening of risk factors for diabetic microvascular complications and construction of a gradient boosting decision tree predictive model,” Frontiers in Endocrinology, vol. 17, art. 1784699, 2026, doi: 10.3389/fendo.2026.1784699.

[4] S. Schallmoser, T. Zueger, M. Kraus, M. Saar-Tsechansky, C. Stettler, and S. Feuerriegel, “Machine learning for predicting micro- and macrovascular complications in individuals with prediabetes or diabetes: Retrospective cohort study,” Journal of Medical Internet Research, vol. 25, e42181, 2023, doi: 10.2196/42181.

[5] F. He, C. N. Y. Ling, S. Nusinovici, C.-Y. Cheng, T. Y. Wong, J. Li, and C. Sabanayagam, "Development and external validation of machine learning models for diabetic microvascular complications: Cross-sectional study with metabolites," Journal of Medical Internet Research, vol. 26, e41065, 2024, doi: 10.2196/41065.

[6] R. Caruana, "Multitask learning," Machine Learning, vol. 28, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.

[7] Y. Zhang and Q. Yang, "An overview of multi-task learning," National Science Review, vol. 5, no. 1, pp. 30–43, 2018, doi: 10.1093/nsr/nwx105.

[8] E. Kim, D. S. Pieczkiewicz, M. R. Castro, P. J. Caraballo, and G. J. Simon, "Multi-task learning to identify outcome-specific risk factors that distinguish individual micro and macrovascular complications of type 2 diabetes," AMIA Summits on Translational Science Proceedings, pp. 122–131, 2018, PMID: 29888055.

[9] N. Meinshausen and P. Bühlmann, "Stability selection," Journal of the Royal Statistical Society: Series B, vol. 72, no. 4, pp. 417–473, 2010, doi: 10.1111/j.1467-9868.2010.00740.x.

[10] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," Journal of Machine Learning Research, vol. 18, no. 174, pp. 1–54, 2018.

[11] J. Zhang, J. Miao, K. Zhao, and Y. Tian, "Multi-task feature selection with sparse regularization to extract common and task-specific features," Neurocomputing, vol. 340, pp. 76–89, 2019, doi: 10.1016/j.neucom.2019.02.035.

[12] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

[13] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[14] K. R. Tan et al., "Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review," Journal of Diabetes Science and Technology, vol. 17, no. 2, pp. 474–489, 2023, doi: 10.1177/19322968211056917.

[15] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.

[16] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, art. 15, 2012, doi: 10.1145/2382577.2382579.

[17] R. Knauer, N. M. Brisson, G. N. Duda, D. Falla, D. W. Evans, and E. Rodner, "Best subset selection meets resampling: Tackling selection instability at very low numbers of events per variable," The International Journal of Biostatistics, online ahead of print, 2026, doi: 10.1515/ijb-2025-0024.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

[19] T. S. Ahmed et al., "Ocular microvascular complications in diabetic retinopathy: Insights from machine learning," BMJ Open Diabetes Research & Care, vol. 12, no. 1, e003758, 2024, doi: 10.1136/bmjdrc-2023-003758.