PREDIKSI DIAGNOSA PENYAKIT BERDASAR ANAMNESA MENGGUNAKAN CLINICALBERT LARGE LANGUAGE MODEL

DISEASE DIAGNOSE BASED ON ANAMNESIS USING CLINICALBERT LARGE LANGUAGE MODEL

Achmad Rizal[1], Intan Dzikria[2*], Maulana Achmad Rifa'i[3]
E-mail: [1)] m11315816@mail.ntust.edu.tw, [2,*)] intandzikria@untag-sby.ac.id
[3)] maulanaalan0822@gmail.com,

[1] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
[2] Sistem dan Teknologi Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas, Universitas 17 Agustus 1945 Surabaya
[3] Teknik Informatika, Fakultas Teknologi Elektro dan Informatika Cerdas, Universitas 17 Agustus 1945 Surabaya

Abstrak

International Classification of Diseases (ICD-10) digunakan sebagai standar kodifikasi internasional untuk klasifikasi penyakit dan dicatat oleh dokter saat melakukan pemeriksaan pasien dan anamnesis. Hasil anamnesis merupakan keluhan pasien, riwayat penyakit pribadi dan keluarga untuk menentukan diagnosis. Penelitian ini memanfaatkan Large Language Model (LLM) untuk membantu dokter dalam memprediksi kode ICD-10 yang sesuai dengan anamnesis pasien. Penelitian ini menggunakan model BERT, terkhusus pada model BERT yang sebelumnya sudah dilatih dengan dataset MIMIC-III, yakni merujuk pada klasifikasi kesehatan, lalu selanjutnya dilatih kembali menggunakan real dataset sejumlah 9241 data anamnesis dan ICD-10 dari salah satu fasilitas kesehatan. Hasil menunjukkan bahwa model mampu melakukan prediksi diagnosis menggunakan kode ICD-10 pada top 10 dengan tingkat akurasi sebesar 52% dan nilai F1 Score sebesar 49%. Penelitian ini menegaskan bahwa kinerja model sangat berpengaruh terhadap jumlah data yang digunakan untuk training dan dapat memengaruhi hasil akhir dari nilai akurasi prediksi.

Kata kunci**:** *rekam medis, diagnosis kode ICD – 10, clinicalbert, large language model*

Abstract

The International Classification of Diseases (ICD-10) is used as an international standard for disease coding and is recorded by physicians during patient examinations and anamnesis. The results of anamnesis include patient complaints, as well as personal and family medical history, which are used to determine a diagnosis. This study leverages a Large Language Model (LLM) to assist physicians in predicting appropriate ICD-10 codes based on patient anamnesis. The research utilizes the BERT model, specifically a version pre-trained on the MIMIC-III dataset, which is widely used for healthcare-related classification tasks. The model is then further fine-tuned using a real-world dataset consisting of 9,241 anamnesis records and their corresponding ICD-10 codes from a healthcare facility. The results show that the model is capable of predicting ICD-10 diagnosis codes for the top 10 classes with an accuracy of 52% and an F1-score of 49%. This study highlights that model performance is strongly influenced by the amount of training data, which significantly affects the final prediction accuracy.

Keywords*: medical records, ICD-10 classification, clinicalbert model, large language model*

1. PENDAHULUAN

Fasilitas pelayanan kesehatan adalah institusi pelayanan kesehatan yang dipengaruhi oleh perkembangan medis, teknologi, dan kehidupan sosial ekonomi masyarakat. Fasilitas pelayanan kesehatan harus mampu meningkatkan pelayanan kesehatan dengan bantuan tenaga medis dan non-medis melalui pelayanan rawat inap, rawat jalan, dan gawat darurat, serta sering digunakan sebagai pusat pendidikan dan penelitian medis [1]. Rekam medis merupakan berkas yang berisikan informasi identitas pasien, anamnesis dokter, diagnosis dokter, dan tindakan medis yang telah dilakukan. Informasi tersebut menjadi tindakan lebih lanjut dalam pelayanan medis baik [2].

International Classification of Diseases (ICD) adalah sistem standar internasional yang digunakan untuk mengkategorikan dan mengodekan penyakit yang telah dikeluarkan oleh Organisasi Kesehatan Dunia (WHO). Kode ICD-10 digunakan sebagai standar diagnosis yang dicatat di dalam rekam medis sebagai dokumentasi klinis pelayanan kesehatan. Dokter menentukan diagnosis pasien berdasarkan ICD-10 melalui proses anamnesis kesehatan pasien yang didasarkan pada ilmu pengetahuan medis, waktu yang cukup lama, dan subjektivitas tinggi seorang dokter.

Prediksi data dalam bidang kesehatan telah banyak dilakukan oleh penelitian terdahulu, baik untuk memprediksi kebutuhan obat maupun diagnosis penyakit tertentu [3]. Dalam konteks diagnosis, berbagai penelitian telah memanfaatkan teknik *Machine Learning* untuk memprediksi penyakit berdasarkan data pasien. Salah satu studi yang cukup banyak dilakukan adalah prediksi penyakit jantung menggunakan algoritma seperti *Decision Tree, Naïve Bayes,* dan *Support Vector Machine* [4]. Beberapa penelitian terdahulu mengembangkan model ClinicalBERT untuk memahami teks klinis dan memprediksi kode ICD – 10 [5]. Namun, berbagai penelitian tersebut masih memiliki keterbatasan dalam hal kemampuan menangani data teks medis yang kompleks serta memahami konteks klinis secara mendalam. Oleh karena itu, penelitian ini menggunakan pendekatan *single-label classification* untuk memprediksi kode ICD-10 berdasarkan teks anamnesis pasien dengan memanfaatkan model berbasis *Large Language Model* (LLM), yang diharapkan mampu memahami konteks klinis secara lebih baik dibandingkan dengan metode konvensional.

Melalui kecerdasan buatan, *natural language processing* dapat menciptakan sistem prediksi kode ICD – 10 secara lebih efisien dan akurat [6]. Prediksi dapat dilakukan melalui ekstraksi pengetahuan atas keluhan pasien agar sesuai dengan bahasa medis. Penelitian ini bertujuan untuk mengembangkan prediksi ICD-10 dengan *large language models* (LLMs). ClinicalBERT adalah model bahasa yang dilatih secara khusus menggunakan teks klinis dan catatan medis agar mampu memahami bahasa medis secara kontekstual dan semantik [7]. Penelitian ini menerapkan proses *fine-tuning* untuk mencapai performa model prediksi yang lebih baik [7] untuk memahami informasi yang ada di dalam teks keluhan medis yang diberi label kode ICD-10 secara lebih akurat.

Hasil penelitian ini diharapkan mampu membantu dokter dalam memprediksi kode diagnosis penyakit ICD-10 melalui keluhan pasien secara lebih cepat. Penelitian ini juga berkontribusi pada peningkatan kualitas model LLM ClinicalBERT yang telah dilakukan *fine-tuning*.

2. LANDASAN TEORI

2.1 Rekam Medis Elektronik

Rekam medis elektronik merupakan sistem elektronik yang digunakan untuk pendataan dan dokumentasi pasien secara cepat, serta keamanan data dan privasi terjaga sehingga mutu pelayanan kesehatan dapat meningkat dengan adanya teknologi tersebut [2]. Rekam medis elektronik merupakan sebuah informasi administratif, klinis, dan demografi dari pasien.

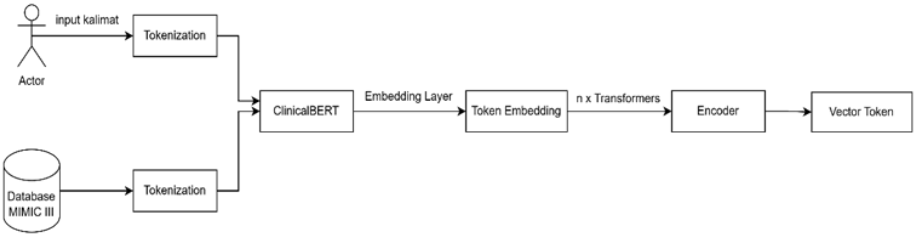
Sehingga dapat digunakan untuk analisis prevalensi penyakit dan pola pemanfaatan pelayanan kesehatan. Rekam medis elektronik mempunyai data yang sangat besar dan kurang dimanfaatkan untuk inovasi ilmu kesehatan, sehingga diperlukan analisis data untuk membantu sistem pengambilan keputusan dalam perawatan kesehatan [8].

Kode ICD-10 merupakan kode diagnosis penyakit dengan standar WHO yang dimasukkan ke dalam rekam medis elektronik oleh dokter sebagai bagian dari pencatatan rekam medis pasien. Penentuan ICD-10 dilakukan dengan proses pemeriksaan pasien atas keluhan yang dihadapi. Keluhan pasien atau anamnesis tersebut dapat menjadi data penting untuk melakukan prediksi diagnosis penyakit, karena mengandung informasi klinis berupa gejala, riwayat penyakit, serta kondisi pasien yang relevan dengan proses penentuan diagnosis. Beberapa penelitian mengungkapkan bahwa catatan klinis yang berisi informasi diagnosis dan perawatan pasien dapat digunakan untuk memprediksi kode ICD secara signifikan [9].

2.2 Large Language Model

Large language model (LLM) merupakan sebuah model yang dapat memproses bahasa manusia dengan memodelkan semantik dalam jumlah data teks yang besar. LLM bukan hanya pemrosesan bahasa saja namun dapat bermanfaat bagi pelayanan kesehatan serta memberikan kontribusi dalam konteks keamanan dan privasi data [10]. Model LLM ini telah di kembangkan dengan *deep learning* dan *natural language processing* yang bertransformasi dalam berbagai aspek praktik medis [11]. Model LLM modern memanfaatkan jaringan saraf, khususnya arsitektur Transformer. Arsitektur transformer dapat menggabungkan miliaran parameter yang dapat menangkap pola rumit dalam bahasa dan pengetahuan spesifik seperti domain medis.

ClinicalBERT adalah model bahasa yang dilatih secara khusus menggunakan teks klinis dan catatan medis *MIMIC – 3* oleh penyedia layanan kesehatan agar mampu memahami bahasa medis [9]. ClinicalBERT merupakan model *pretrained* yang dapat menghasilkan prediksi diagnosis. Gambar 1 menunjukkan diagram alur tahapan model ClinicalBERT.



Gambar 1 Diagram flow ClinicalBert

Seperti yang ditunjukkan pada Gambar 1, ClinicalBERT melakukan proses tokenisasi dari sebuah kalimat, kemudian melakukan pre-training dengan *database* MIMIC III yang sudah *tertokenisasi* sehingga dapat menangkap kesamaan kata klinis. Setelah itu menghasilkan token embedding yang semantik dengan data klinis kemudian dilakukan *transformer encoder* untuk menghasilkan *vector token* [9].

ClinicalBERT digunakan dalam penelitian bidang medis untuk mengekstraksi informasi penting dari rekam medis elektronik, melakukan klasifikasi diagnosis, serta memprediksi kode ICD-10 secara otomatis berdasarkan teks anamnesis pasien [9].

2.3 Fine-Tuning Model ClinicalBERT

Fine-tuning merupakan proses penyesuaian model bahasa yang telah dilatih sebelumnya dengan data yang lebih spesifik pada tugas tertentu agar model dapat beradaptasi secara optimal terhadap karakteristik domain dan label target [7]. Proses *fine-tuning* dilakukan dengan melatih

ulang seluruh parameter model ClinicalBERT menggunakan data keluhan yang telah diberi label kode ICD-10. Pada tahap ini bobot model diperbarui melalui mekanisme *backpropagation* dan optimasi berbasis fungsi kerugian, sehingga representasi yang dihasilkan oleh model menjadi lebih relevan terhadap tugas klasifikasi diagnosis [7]. Sehingga dengan menerapkan *fine-tuning* pada model ClinicalBERT, tugas klasifikasi diagnosis tidak hanya bergantung pada pengetahuan bahasa medis yang diperoleh saat pretraining, tetapi juga mampu menyesuaikan dengan distribusi data dan struktur label kode ICD-10 yang digunakan dalam bahasa medis tertentu.

2.4 Classification Fully Connected Layer

Fully Connected Layer berfungsi untuk memetakan representasi yang dihasilkan oleh model ClinicalBERT yang telah difine-tuning ke dalam ruang kelas target sesuai dengan tugas klasifikasi *Fully Connected Layer*. *Layer* ini berperan sebagai *classification head* yang memungkinkan model untuk mempelajari hubungan antara fitur semantik teks dan label kelas melalui proses optimasi dari fungsi kerugian [12]. Penggunaan *fully connected layer* dalam tugas klasifikasi diagnosis medis seperti kode ICD-10 memungkinkan model untuk menyesuaikan representasi bahasa klinis dengan struktur label diagnosis yang kompleks. Dengan demikian, fully connected layer tidak hanya berfungsi sebagai pemetaan akhir saja, tetapi juga sebagai komponen penting dalam proses fine-tuning antara representasi teks klinis dan kelas diagnosis.

2.5 Evaluasi Model

Akurasi sistem prediksi menggunakan LLM membutuhkan proses evaluasi [13]. Pada penelitian ini, evaluasi dilakukan menggunakan 3 metrik, yaitu *precision, recall,* dan *F1-score*. Ketiga metrik ini digunakan karena mampu memberikan gambaran performa model klasifikasi secara lebih komprehensif dibandingkan hanya menggunakan akurasi, terutama pada data yang tidak seimbang yang umum terjadi pada data medis [14].

Precision adalah presisi menunjukkan rasio sampel yang telah di klasifikasikan dengan benar dan menunjukkan beberapa prediksi yang benar oleh model [14]. Persamaan (1) menunjukkan formula untuk menghitung *precision*. Standar nilai *precision* adalah semakin mendekati 1, semakin baik, karena model memiliki tingkat kesalahan prediksi positif yang rendah. Nilai presisi yang tinggi sangat penting pada kasus medis untuk menghindari kesalahan diagnosis yang tidak tepat.

$$Precision = TP/(TP+FP) \quad (1)$$

Keterangan :
Precision = proporsi prediksi yang benar dari seluruh prediksi yang benar
TP = True Positif (prediksi yang benar)
FP = False Positif (prediksi yang salah)

Recall adalah untuk mengukur seberapa banyak jumlah prediksi yang positif oleh model [14]. Persamaan (2) menunjukkan formula untuk menghitung nilai *recall.* Nilai *recall* dikatakan baik apabila nilai mendekati 1, yang berarti model mampu mendeteksi sebagian besar kasus positif dengan baik dan memiliki tingkat false negative yang rendah. Nilai recall yang tinggi menjadi sangat penting karena berkaitan langsung dengan kemampuan model dalam mendeteksi kasus penyakit.

$$Recall = TP/(TP+FN) \quad (2)$$

Keterangan :

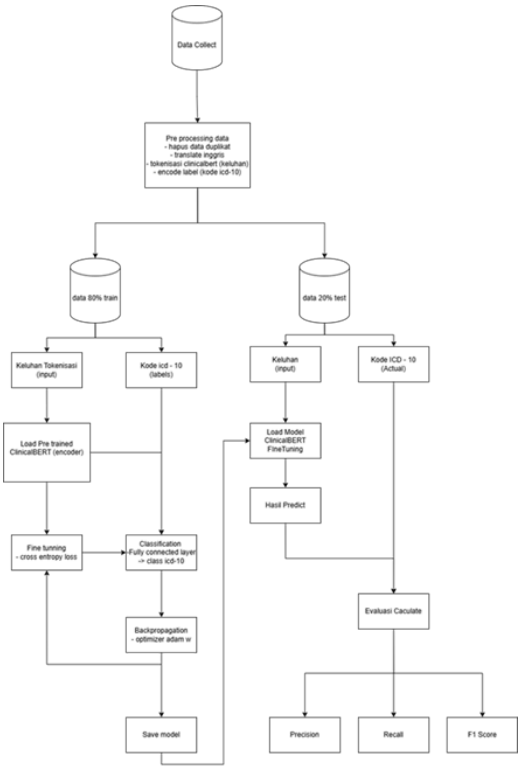
Recall = proporsi prediksi positif yang berhasil ditemukan
TP = True Positif (prediksi yang benar)
FP = False Positif (prediksi yang salah)

F1–Score adalah menggabungkan hasil dari precision dan recall menjadi 1 angka untuk memberikan keseimbangan terhadap data keduanya [14]. Persamaan (3) menunjukkan formula untuk menghitung nilai *F1-Score*. Standar nilai *F1-score* adalah apabila mendekati nilai 1, maka semakin baik untuk menunjukkan bahwa model memiliki keseimbangan antara ketepatan prediksi dan deteksi seluruh data positif.

$$F1\ Score = 2 \times (Precision \times Recall)/(Precision+Recall) \quad (3)$$

3. METODOLOGI

Penelitian ini mempunyai alur sistematis dalam pengembangan model klasifikasi diagnosis kode ICD-10 berdasarkan anamnesis pasien dengan memanfaatkan large language model ClinicalBERT. Gambar 2 menunjukkan kerangka penelitian secara menyeluruh.



Gambar 2 Kerangka Penelitian

Penelitian ini menggunakan data anamnesis dan kode ICD-10 dari sebuah rekam medis elektronik yang digunakan oleh sebuah fasilitas kesehatan, sejumlah 9.241 data. Tabel 1 menunjukkan data sampel yang digunakan untuk training model pada penelitian ini. Kemudian tahap pra-pemrosesan data dilakukan untuk memastikan data berada dalam kondisi yang sesuai sebelum digunakan oleh model. Tahap pra-pemrosesan meliputi pembersihan data, normalisasi teks, penggabungan kolom keluhan, *tokenisasi*, *padding*, serta *transformasi* label ke dalam bentuk numerik agar sesuai dengan kebutuhan model.

Data yang telah melalui tahap pra-pemrosesan kemudian dibagi menjadi dua bagian, yaitu data *training* (80%) dan data uji (20%). Data training digunakan untuk melatih model dan menyesuaikan parameter, sedangkan data testing digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi pada data yang belum pernah dilihat sebelumnya, sehingga hasil evaluasi lebih objektif dan tidak bias.

Tabel 1 Data Sample Training

Keluhan	Kode ICD – 10	Nama ICD-10
Cough, chills, dizziness	J06.9	Acute upper respiratory infection, unspecified
Leg pain, wounds, body heat	L20.8	Other atopic dermatitis
Upper back toothache	K04.0	Pulpitis
Left chest pain, heartburn, nausea	R07.3	Other chest pain
Ears feel like they are ringing	H60	Otitis externa

Pada tahap pelatihan, teks keluhan pasien dimasukkan ke dalam model *pretrained* ClinicalBERT yang berperan sebagai encoder untuk mengekstraksi representasi bahasa klinis dari teks. Representasi tersebut kemudian diproses melalui tahap *fine-tuning* untuk menyesuaikan parameter model agar mampu melakukan klasifikasi ke dalam kode ICD-10 sesuai dengan data yang digunakan. Selama proses pelatihan, model melakukan *optimasi* dengan meminimalkan nilai *loss*, sehingga parameter model diperbarui secara bertahap untuk mengenali pola hubungan antara teks keluhan dan diagnosis. Setelah model mencapai performa yang optimal, model disimpan untuk digunakan pada tahap evaluasi.

Pada tahap evaluasi, model yang telah di *fine-tuning* digunakan untuk memprediksi kode ICD-10 pada 20% data uji. Hasil prediksi kemudian dibandingkan dengan label sebenarnya untuk mengukur kinerja model menggunakan metrik evaluasi klasifikasi seperti *precision, recall,* dan *F1-score*. Hasil evaluasi ini digunakan untuk menilai efektivitas model dalam melakukan prediksi diagnosis secara akurat dan konsisten. Di sisi lain, pengujian ini menggunakan 100 data uji tambahan untuk menguji tingkat ketepatan prediksi diagnosis yang dibandingkan dengan penentuan diagnosis dokter pada data nyata.

4. HASIL DAN PEMBAHASAN

4.1 Perancangan Model

Penelitian ini merancang model untuk membangun sistem klasifikasi teks dalam memprediksi kode diagnosis ICD-10 berdasarkan keluhan atau anamnesis pasien. Model yang digunakan berbasis *Large Language Model* (LLM) dengan pendekatan *supervised learning*, menggunakan data berlabel berupa pasangan teks keluhan dan kode diagnosis. Arsitektur yang digunakan adalah model transformer, yaitu ClinicalBERT, yang telah melalui proses *pre-training* pada data klinis. Model kemudian di *fine-tuning* menggunakan dataset penelitian pada skenario top 50 dan top 10 kode ICD-10 untuk menyesuaikan dengan karakteristik data. Sebelum pelatihan, teks diproses melalui tahap tokenisasi, *padding*, dan *truncation* untuk menghasilkan representasi numerik dengan panjang sekuens yang seragam. Proses pelatihan dilakukan menggunakan parameter seperti *epoch*, *learning rate*, *batch size*, dan *optimizer*, dengan fungsi *loss cross-entropy* untuk mengukur kesalahan prediksi.

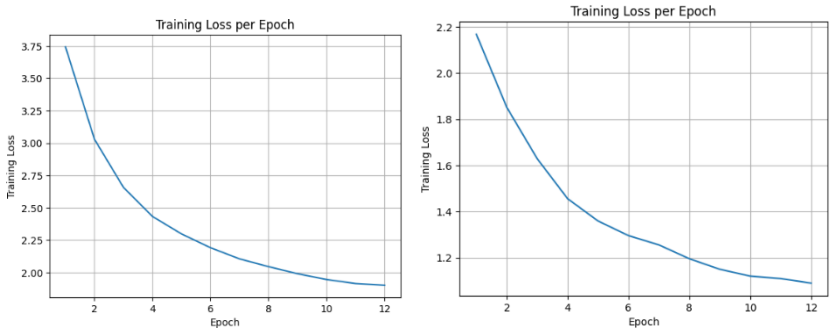
Selama pelatihan, nilai *loss* dipantau untuk memastikan model mengalami peningkatan performa. Model yang telah dilatih kemudian diuji menggunakan data testing untuk mengevaluasi kemampuan generalisasi. Output model berupa probabilitas setiap kelas ICD-10, di mana kelas dengan nilai tertinggi dipilih sebagai hasil prediksi. Evaluasi dilakukan

menggunakan metrik akurasi, *precision, recall*, dan *F1-Score*. Dengan pendekatan ini, sistem diharapkan mampu membantu proses klasifikasi diagnosis secara otomatis dengan tingkat akurasi yang baik.

4.2 Training Model

Hasil *training* model ClinicalBERT menunjukkan pola penurunan training loss yang konsisten pada kedua skenario, yaitu top 50 dan top 10 kode ICD-10. Training loss ini disebabkan oleh selisih antara hasil prediksi model dan label sebenarnya, yang dihitung menggunakan fungsi kerugian selama proses pelatihan. Penurunan nilai *training loss* menunjukkan bahwa model secara bertahap mampu menyesuaikan parameter dan meningkatkan kemampuannya dalam mengenali pola pada data teks klinis.

Pada skenario top 50, seperti yang ditunjukkan pada Gambar 3a, nilai *training loss* pada *epoch* awal berada pada angka 3.7444 dan secara bertahap menurun hingga mencapai 1.9021 pada *epoch* ke-12. Sementara itu, Gambar 3b menunjukkan skenario top 10, di mana nilai *training loss* dimulai dari 2.1588 dan menurun lebih cepat hingga mencapai 1.0337pada *epoch* ke-12.



Gambar 3a. Training Loss Top 50 Label, 3b. Training Loss Top 10 Label

Pada skenario top 50, model harus membedakan 50 kelas diagnosis, sehingga ruang klasifikasi menjadi lebih kompleks dan menyebabkan nilai *loss* awal lebih tinggi serta penurunan yang relatif lebih lambat. Sedangkan skenario top 10 hanya memiliki 10 kelas, sehingga model lebih mudah mempelajari pola dan melakukan pemisahan antarkelas, serta menghasilkan nilai *loss* yang lebih rendah dan konvergensi yang lebih cepat.

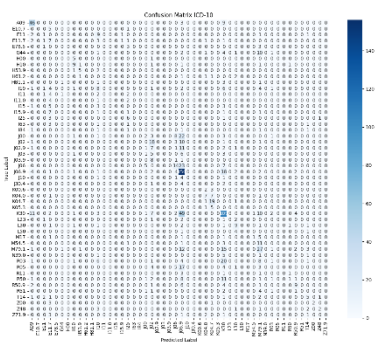
Dengan demikian, dapat disimpulkan bahwa kedua skenario pelatihan menunjukkan proses pembelajaran yang baik dan mencapai kondisi *konvergen*. Skenario top 10 memiliki keunggulan dalam hal kemudahan pembelajaran dan kecepatan konvergensi, sedangkan skenario top 50 merepresentasikan kasus yang lebih kompleks dan mendekati kondisi nyata dalam klasifikasi kode ICD-10.

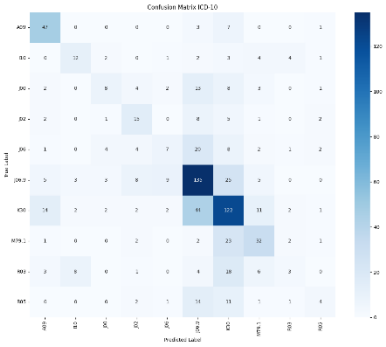
4.3 Evaluasi Model

Hasil evaluasi model dilakukan menggunakan metrik *precision, recall, F1-score*, dan *accuracy* untuk mengukur performa klasifikasi pada data testing. Model ClinicalBERT telah melalui proses pelatihan pada dua skenario, yaitu top 50 dan top 10 kode ICD-10. Gambar 4 menampilkan *confusion matrix* sebagai hasil evaluasi model. *Confusion matrix* merupakan metode evaluasi yang digunakan untuk melihat perbandingan antara label prediksi dan label sebenarnya pada setiap kelas dalam tugas klasifikasi.

Gambar 4a menunjukkan confusion matrix untuk Top50 ICD-10, dan Gambar 4b menunjukkan confusion matrix untuk Top10 ICD-10. Pada skenario top 50, model

menghasilkan nilai akurasi sebesar 0.39, dengan *macro average F1-score* sebesar 0.16 dan *weighted F1-score* sebesar 0.32. Hasil ini menunjukkan bahwa performa model masih relatif rendah dalam melakukan klasifikasi multi-kelas dengan jumlah kelas yang besar. Sebaliknya, pada skenario top 10, model menunjukkan peningkatan performa yang cukup signifikan dengan nilai akurasi sebesar 0.52, *macro average F1-score* sebesar 0.40, dan *weighted F1-score* sebesar 0.49. Nilai ini lebih tinggi dibandingkan dengan skenario top 50, yang menunjukkan bahwa model untuk top 10 lebih mampu melakukan klasifikasi ketika jumlah kelas lebih sedikit. Perbedaan hasil evaluasi antara kedua skenario ini menunjukkan bahwa jumlah kelas memiliki pengaruh yang signifikan terhadap performa model. Selain itu, ketidakseimbangan jumlah data antarkelas juga menjadi faktor penting yang memengaruhi performa model.



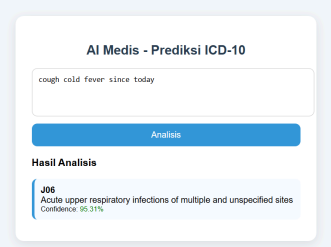


Gambar 4a Confusion Matrix Top 50 Label, 4b Confusion Matrix Top 10 Label

Penelitian ini telah menerapkan teknik *weighted loss*, yang merupakan metode penyesuaian fungsi kerugian dengan memberikan bobot yang berbeda pada setiap kelas berdasarkan distribusi jumlah data. Kelas dengan jumlah data yang lebih sedikit diberikan bobot yang lebih besar, sedangkan kelas dengan jumlah data yang lebih banyak diberikan bobot yang lebih kecil. Hasil evaluasi menunjukkan bahwa model masih cenderung lebih baik dalam memprediksi kelas dengan jumlah data yang lebih banyak dibandingkan kelas minoritas [15]. *Fine-tuning* model berbasis *transformer* seperti ClinicalBERT dipengaruhi oleh berbagai faktor, antara lain ukuran dataset, distribusi data, serta pemilihan *hyperparameter* seperti *learning rate* dan jumlah *epoch*. Penelitian sebelumnya menunjukkan bahwa proses fine-tuning bersifat tidak stabil dan sangat sensitif terhadap inisialisasi bobot serta urutan data pelatihan. Selain itu, pada kondisi dataset kecil, fine-tuning cenderung mengalami ketidakstabilan dan overfitting [16]. Oleh karena itu, diperlukan strategi optimal dalam pemilihan parameter agar model dapat menghasilkan performa yang lebih baik dan stabil.

4.4 Hasil Implementasi Model

Model klasifikasi yang telah ditraining menggunakan ClinicalBERT selanjutnya diimplementasikan ke dalam sistem untuk mendukung proses prediksi diagnosis kode ICD-10. Implementasi dilakukan dengan mengintegrasikan model ke dalam aplikasi berbasis web. Sistem dirancang untuk menerima input berupa teks keluhan pasien oleh pengguna, kemudian data tersebut dikirim ke model untuk dilakukan proses inferensi. Sistem akan menampilkan hasil prediksi berupa kode diagnosis dengan nilai probabilitas tertinggi sebagai rekomendasi utama, seperti yang ditunjukkan pada Gambar 5.



Gambar 5 User Interface Implementasi Model

Tabel 2 menunjukkan hasil prediksi untuk beberapa data uji anamnesis dan diagnosis. Penelitian ini menggunakan 100 data uji dengan kelas Top 10 ICD-10 untuk menilai kesesuaian hasil prediksi ICD-10 dengan penentuan ICD-10 secara nyata oleh dokter. Hasil menunjukkan kesesuaian prediksi sebesar 53% yang menunjukkan bahwa model LLM yang dibangun pada penelitian ini masih memiliki keterbatasan dalam memprediksi diagnosis secara akurat berdasarkan teks anamnesis pasien.

Tabel 2 Hasil Prediksi Data Uji

No	Keluhan	ICD-10 Dokter	ICD-10 Hasil Prediksi Model	Status
1	fever, cough, cold, no fever	J06.9	J06.9	Benar
2	There are no coughs and colds	J06.9	J06	Salah
3	SICK cough and cold	J06.9	J06.9	Benar
4	fever, cough, cold, no fever	J06.9	J06.9	Benar
5	cough cold summer	J06.9	J06	Salah

Rendahnya tingkat kesesuaian ini dapat disebabkan oleh beberapa faktor, seperti kompleksitas bahasa medis dalam teks anamnesis, keterbatasan jumlah data pelatihan, serta adanya kemiripan gejala antara beberapa diagnosis yang berbeda. Dengan demikian, hasil pengujian pada 100 data uji ini menunjukkan bahwa model masih memerlukan pengembangan lebih lanjut, baik dari sisi peningkatan kualitas dan kuantitas data, optimasi model ClinicalBERT, maupun penerapan metode tambahan agar mampu meningkatkan akurasi prediksi dan mendekati hasil diagnosis yang ditentukan oleh tenaga medis.

5. KESIMPULAN DAN SARAN

Penelitian ini bertujuan untuk membangun dan mengevaluasi model klasifikasi teks berbasis Large Language Model dalam memprediksi kode diagnosis ICD-10 berdasarkan keluhan pasien. Penelitian ini menggunakan model ClinicalBERT dengan pendekatan *supervised learning* serta dilakukan pengujian pada dua skenario, yaitu top 50 dan top 10 kode ICD-10. Hasil penelitian ini menunjukkan bahwa model mampu melakukan proses klasifikasi dengan baik, namun *fine-tuning* model berbasis *transformer* seperti BERT dipengaruhi oleh berbagai faktor, antara lain ukuran dataset, distribusi data, serta pemilihan *hyperparameter* seperti *learning rate* dan jumlah *epoch*. Selain itu, pada kondisi dataset kecil, *fine-tuning* cenderung mengalami ketidakstabilan dan *overfitting*. Maka, penelitian di masa depan perlu melakukan strategi optimal dalam pemilihan parameter agar model dapat menghasilkan performa yang lebih baik dan stabil.

Penelitian ini berkontribusi secara akademis pada pengembangan metode klasifikasi teks medis berbasis *Natural Language Processing* dengan memanfaatkan model ClinicalBERT dalam konteks prediksi kode ICD-10. Secara praktis, penelitian ini berkontribusi pada pengembangan sistem pendukung keputusan di bidang kesehatan yang dapat membantu tenaga medis dalam melakukan prediksi kode ICD-10 berdasarkan teks keluhan pasien secara lebih cepat dan efisien. Dengan adanya sistem ini, diharapkan proses penentuan kode diagnosis dapat menjadi lebih konsisten, terstandar, dan mendukung peningkatan kualitas layanan kesehatan.

6. DAFTAR RUJUKAN

[1] Kementerian Kesehatan Republik Indonesia, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 3 Tahun 2020 tentang Klasifikasi dan Perizinan Rumah Sakit," 2020

[2] Kementerian Kesehatan Republik Indonesia, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022 tentang Rekam Medis," 2022.

[3] I. Dzikria and M. M. Nizar, "Prediksi Pengadaan Obat di Ruang Obat Fasilitas Kesehatan Tingkat Pertama dengan Algoritma Holt-Winters Exponential Smoothing," Media Jurnal Informatika, vol. 15, no. 1, p. 28, 2023, doi: 10.35194/mji.v15i1.3175.

[4] A. Omari and H. AbuSharkh, "Heart Disease Prediction Using Machine Learning," in Communications in Computer and Information Science, vol. 2381, pp. 176–185, 2025, doi: 10.1007/978-3-031-82931-4_13.

[5] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[6] V. Klotzman, "Medical Diagnosis Coding Automation: Similarity Search vs. Generative AI," medRxiv preprint, 2024, doi: 10.1101/2024.04.26.24306470.

[7] I. Aden, C. H. T. Child, and C. C. Reyes-Aldasoro, "International Classification of Diseases Prediction from MIMIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art," Big Data and Cognitive Computing, vol. 8, no. 5, 2024, doi: 10.3390/bdcc8050047.

[8] Y. Syahidin and A. I. Suryani, "The Role of Artificial Intelligence in Optimizing Electronic Health Records for Early Detection of Disease," Dinasti International Journal of Education Management and Social Science, vol. 6, no. 3, pp. 2422–2433, 2025, doi: 10.38035/dijemss.v6i3.4024.

[9] K. Huang, J. Altosaar, and R. Ranganath, "Modeling Clinical Notes and Predicting Hospital Readmission," arXiv preprint arXiv:1904.05342, 2020..

[10] Y. Yao et al., "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly," High-Confidence Computing, vol. 4, no. 2, p. 100211, 2024, doi: 10.1016/j.hcc.2024.100211.

[11] S. Maity and M. J. Saikia, "Large Language Models in Healthcare and Medical Applications: A Review," Bioengineering, vol. 12, no. 6, p. 631, 2025, doi: 10.3390/bioengineering12060631.

[12] C. Eang and S. Lee, "Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of BERT and RNN," Applied Sciences, vol. 14, no. 18, 2024, doi: 10.3390/app14188388.

[13] S. A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports, vol. 12, no. 1, pp. 1–9, 2022, doi: 10.1038/s41598-022-09954-8.*

[14] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.*

[15] G. Ramesh et al., “A Review on NLP zero-shot and few-shot learning: methods and applications,” Discover Applied Sciences, vol. 7, no. 9, 2025, doi: 10.1007/s42452-025-07225-5.
[16] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting few-sample BERT fine-tuning," in Proc. 9th Int. Conf. Learn. Representations (ICLR), 2021. [Online]. Available: https://arxiv.org/abs/2006.05987