ZONAsi Jurnal Sistem Informasi
ZONAsi VOL. 8 NO. 2 Mei 2026
Page: 747 - 758 Akreditasi Sinta 4
ISSN: 2656-7407 (Online)
2656-7393 (Print)

MODEL *RETRIEVAL* BERBASIS *NATURAL LANGUAGE PROCESSING* UNTUK REKOMENDASI OTOMATIS ICD-10 DAN ICD-9 PADA SISTEM KLAIM RUMAH SAKIT

D Mastiin T Saputra[1], **Galih**[2]
[1,2] Prodi Teknik Informatika, Fakultas Teknik, Universitas Islam Nusantara
e-mail: [1]danawaterdish88@gmail.com*, [2]galihsetiana@gmail.com

Abstrak

Sejak penerapan Jaminan Kesehatan Nasional, pengkodean diagnosis dan prosedur menjadi komponen penting dalam penentuan tarif INA-CBG pada proses klaim rumah sakit. Namun praktik pengkodean manual sering menimbulkan jeda waktu, ketidakkonsistenan antarpetugas, serta potensi keterlambatan pengajuan klaim ke BPJS Kesehatan. Penelitian ini bertujuan mengembangkan model ringan berbasis Natural Language Processing (NLP) untuk menghasilkan rekomendasi otomatis kode ICD-10 dan ICD-9 dari teks diagnosis dokter dengan kebutuhan komputasi minimal. Metode yang digunakan meliputi normalisasi teks medis, pembentukan sentence embeddings menggunakan model Sentence Transformer, serta perhitungan kemiripan berbasis cosine similarity dalam skema retrieval multi-label. Dataset berasal dari hasil kodifikasi Rumah Sakit Khusus Ibu dan Anak Harapan Bunda Bandung periode Oktober 2024 hingga November 2025, dengan evaluasi dilakukan menggunakan pendekatan temporal hold-out pada data Desember 2025 dan Januari 2026. Hasil evaluasi menunjukkan performa stabil dengan nilai Hit-3 sebesar 80,7% pada Desember 2025 dan 77,5% pada Januari 2026. Nilai Recall-3 sebesar 70% pada Desember 2025 dan 68% pada Januari 2026, sementara Mean Reciprocal Rank (MRR) berada di atas 0,4, menunjukkan bahwa kode relevan umumnya muncul pada peringkat awal rekomendasi, analisis per unit layanan menunjukkan bahwa performa model lebih optimal pada kasus rawat inap dibandingkan rawat jalan, mengindikasikan bahwa karakteristik teks diagnosis dan kompleksitas kasus pada rawat inap lebih sesuai dengan pendekatan semantic retrieval yang digunakan. Temuan ini mengindikasikan bahwa pendekatan retrieval berbasis NLP mampu memberikan rekomendasi kode ICD secara cepat dan konsisten, serta berpotensi mendukung percepatan proses pengkodean dan meminimalkan risiko keterlambatan klaim layanan Kesehatan

Kata kunci: *Natural Language Processing, Otomatisasi, ICD 10, ICD 9, INA-CBG, Sentence Embeddings, Cosine Similarty, Rekam Medis, Validasi Temporal, Multi Label Data*

Abstract

Since the implementation of the National Health Insurance (JKN), the coding of diagnoses and procedures has become a important component in determining INA-CBG tariffs for hospital claims. However, manual coding practices often lead to time lags, inconsistencies between coders, and potential delays in claim submissions to BPJS Kesehatan. This study aims to develop a lightweight Natural Language Processing (NLP) model to generate automated ICD-10 and ICD-9 code recommendations from physician diagnostic text with minimal computational requirements, the methodology involves medical text normalization, the generation of sentence embeddings using a Sentence Transformer model, and cosine similarity calculations within a multi-label retrieval framework. The dataset was sourced from the codification records of Harapan Bunda Mother and Child Hospital (RSKIA) Bandung for the period of October 2024 to November 2025 The results demonstrated stable performance, with a Hit-3 value of 80.7% in December 2025 and 77.5% in January 2026. Recall-3 values were 70% and 68% for December 2025 and January 2026, respectively, while the Mean Reciprocal Rank (MRR) remained above 0.4, indicating that relevant codes consistently appeared at the top of the recommendation list. Analysis by service unit revealed that the model's performance was more optimal for inpatient cases compared to outpatient cases, indicating that the diagnostic text characteristics and case complexity in inpatient settings are more compatible with the semantic

retrieval approach used. These findings suggest that an NLP-based retrieval approach can provide rapid and consistent ICD code recommendations, potentially accelerating the coding process and minimizing the risk of claim delays in healthcare services.
Keywords: *Natural Language Processing, Automated, ICD Coding, INA-CBG Reimbursment, Sentence Embeddings, Cosine Similarity, Medical Record, Temporal Validation, Multi-label Retrieval*

1. PENDAHULUAN

Penerapan Jaminan Kesehatan Nasional (JKN) membuat proses pengkodean diagnosis dan prosedur medis sebagai elemen yang sangat penting dalam penentuan tarif INA-CBG pada fasilitas rumah sakit. Ketepatan koding diagnosis sangat menentukan dalam pembiayaan yang akan dibayarkan oleh BPJS Kesehatan terhadap rumah sakit serta berpengaruh terhadap kejadian penundaan penundaan pembayaran[1] namun praktik di berbagai rumah sakit masih menunjukkan adanya jeda waktu antara penentuan diagnosis oleh dokter dan proses pengkodean oleh perekam medis, Kondisi ini sering memicu keterlambatan klaim, ketidakkonsistenan antarpetugas, serta potensi ketidaksesuaian pembiayaan yang masih banyak terjadi permasalahan di lapangan antara BPJS Kesehatan dan FKRTL khususnya terkait pengodean.[2] Di Rumah Sakit Khusus Ibu dan Anak Harapan Bunda Bandung proses pengkodean biasanya dilakukan setelah layanan selesai dan menunjukkan variasi antarpetugas, sehingga mengindikasikan kebutuhan mendesak akan sistem rekomendasi kode yang cepat, konsisten, dan terstandardisasi.

Berbagai penelitian sebelumnya telah menerapkan pendekatan machine learning dan deep learning untuk klasifikasi teks medis. Beberapa ANN dilatih sesuai dengan empat kategori durasi rawat inap pada dataset kehidupan nyata. Dari puluhan ribu kode ICD-10, kami hanya mempelajari 613 yang dirujuk setidaknya 30 kali dan akhirnya memilih 346 yang memiliki efektivitas prediksi tertinggi[3] studi lain penggunaan klasifikasi teks multi-label berbasis jaringan Convolutional Neural Networks (CNN), dengan data yang diambil dari catatan SOAP dan daftar obat, untuk memprediksi kode ICD-10 merupakan pendekatan baru di bidang ini. Tujuan dari penelitian ini adalah untuk membangun model pembelajaran mendalam yang dapat membantu dokter dalam memilih kode ICD-10 yang paling relevan,[4] di Indonesia penelitian mengguakan text mining dapat digunakan untuk membantu petugas rekam medis dalam pengkodean penyakit berdasarkan kode ICD-10 penggunaan model deep learning mempercepat proses pengkodean penyakit, menghemat waktu dan sumber daya kesehatan. Yang memungkinkan pemberian perawatan yang lebih cepat dan efisien kepada pasien.[5] sementara model BERT-BiGRU untuk klasifikasi ICD-10 menunjukkan tantangan signifikan berupa ketidakseimbangan label serta kebutuhan penyesuaian khusus pada teks medis bahasa Indonesia.[6] Pada penelitian lainnya model pemrosesan teks dengan NLP

menggunakan HMM dan algoritma Viterbi untuk menghasilkan gejala-gejala jenis gangguan skizofrenia.[7] Temuan dari berbagai studi tersebut menunjukkan bahwa meskipun model berbasis neural network dapat mencapai performa tinggi, pendekatan tersebut membutuhkan sumber daya komputasi yang tinggi serta proses pelatihan yang tidak selalu praktis sehingga akan mempengaruhi kepada performa salah satu literatur menunjukan pengkodean ICD-10 membutuhkan waktu 3 menit 10 detik[8] untuk rumah sakit dengan sumber daya terbatas tentu akan menjadi masalah besar. Seluruh pendekatan tersebut memiliki keterbatasan utama, yakni belum mempertimbangkan karakteristik unik teks diagnosis medis berbahasa Indonesia yang sarat dengan singkatan klinis, variasi istilah, dan gaya penulisan dokumen medis.

Berdasarkan penelusuran literatur yang dilakukan penulis, penelitian di Indonesia yang memanfaatkan pendekatan SentenceTransformer untuk rekomendasi kode ICD masih sangat terbatas dan juga karakteristik pengkodean ICD yang bersifat multi-label dengan distribusi kode yang sangat tidak seimbang menjadikan pendekatan klasifikasi konvensional lebih kompleks dan berisiko mengalami bias terhadap label dominan.[9] Selain itu, penelitian yang ada masih berfokus pada model yang membutuhkan pelatihan intensif, sementara belum dieksplorasi pendekatan ringan berbasis non-training inference yang memungkinkan pembaruan cepat melalui regenerasi embedding tanpa retraining model besar.

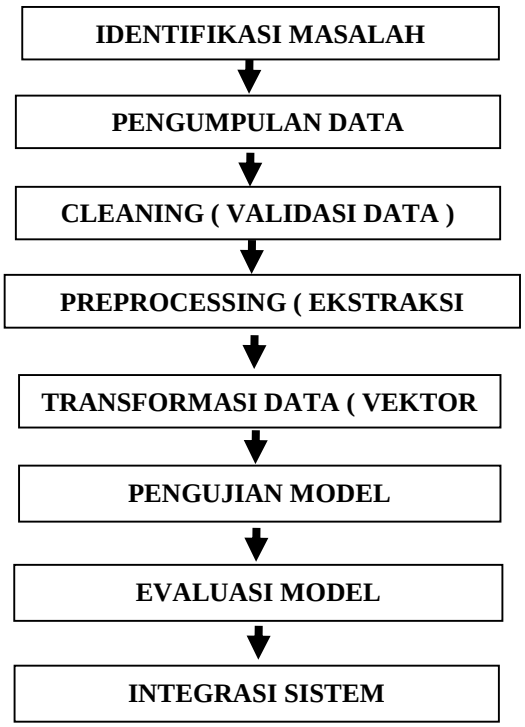
Berdasarkan kesenjangan tersebut, penelitian ini menawarkan kebaruan berupa pengembangan sistem rekomendasi kode ICD berbasis sentence embeddings dan cosine similarity yang dikenal cepat dalam

menangkap makna.[10] Model dikembangkan dengan memanfaatkan normalisasi singkatan medis lokal, pembentukan embedding menggunakan SentenceTransformer, serta integrasi tiga komponen teks diagnosis (diagnosa utama, diagnosa sekunder, dan plan). Pendekatan ini dirancang sebagai solusi ringan, efisien, dan dapat berjalan secara real-time melalui API, sehingga mudah diintegrasikan dengan sistem informasi rumah sakit dan relevan untuk digunakan dalam proses operasional sehari-hari.[11]

Dengan demikian, tujuan penelitian ini adalah merancang, mengembangkan, dan mengevaluasi model rekomendasi otomatis kode ICD-10 dan ICD-9 menggunakan pendekatan NLP ringan berbasis kemiripan semantik dengan cosine similarity untuk meningkatkan kecepatan, konsistensi, dan akurasi proses pengkodean, serta mendukung percepatan klaim pelayanan kesehatan pada fasilitas kesehatan di Indonesia.

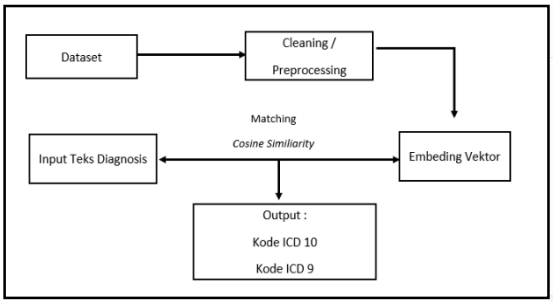
2. METODE PENELITIAN

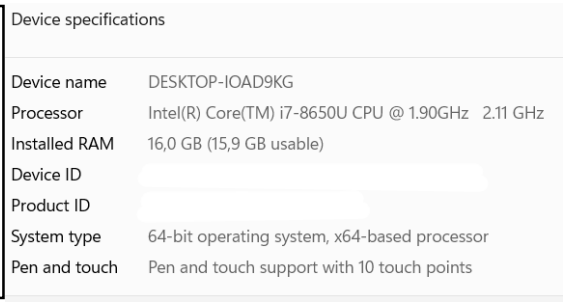
Penelitian ini menggunakan desain kuantitatif eksperimental dengan pendekatan Natural Language Processing (NLP) berbasis semantic similarity dalam skema retrieval multi-label. Desain ini dipilih karena penelitian bertujuan mengevaluasi kinerja sistem rekomendasi kode ICD-10 dan ICD-9 tanpa proses pelatihan ulang model (non-training inference). Sistem menghasilkan rekomendasi otomatis dengan menghitung tingkat kemiripan semantik antara teks diagnosis input dan basis data kodifikasi historis rumah sakit.



Gambar 1. Prosedur Penelitian

Pendekatan sentence embeddings digunakan untuk merepresentasikan teks diagnosis dalam bentuk vektor numerik berdimensi tetap, representasi ini memungkinkan perhitungan kemiripan semantik menggunakan cosine similarity. Metode ini dipilih karena mampu menangkap konteks kalimat secara utuh dibandingkan pendekatan berbasis frekuensi kata seperti Bag-of-Words atau TF-IDF, serta lebih ringan dibandingkan model deep learning klasifikasi yang memerlukan pelatihan intensif dan sumber daya komputasi besar





Gambar 2. Rancangan Model dan Spesifikasi CPU Uji Coba

Populasi penelitian mencakup seluruh catatan diagnosis dan prosedur medis yang telah dikodekan oleh petugas perekam medis di Rumah Sakit Khusus Ibu dan Anak Harapan Bunda Bandung. Data yang digunakan meliputi diagnosis utama, diagnosis sekunder, serta plan yang diberikan oleh dokter.

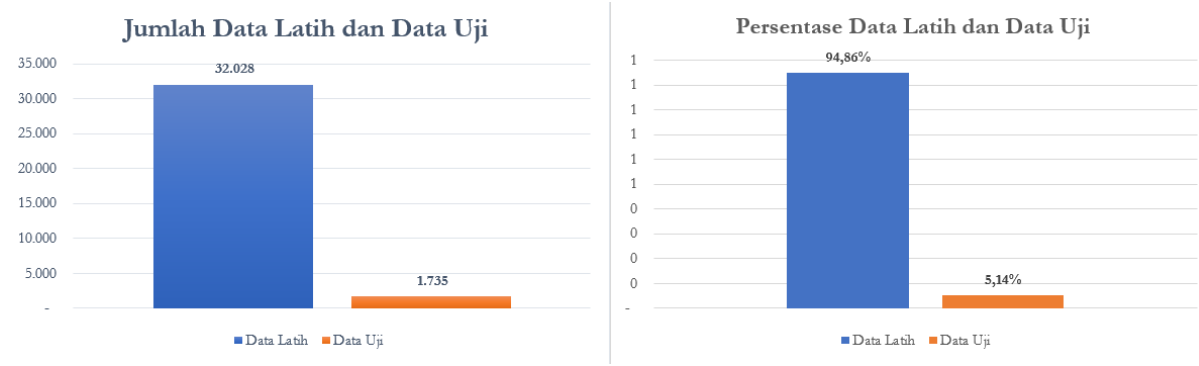
Tabel 1. Sample Dataset

Diagnosa	**Plan**	**ICD-10**	**ICD-9**
Abortus inkomplit	Kuretase	O03.4	69.02
Autism ec overscreen time	Lat.atensi dg gerakan 4 fleksibilitas	F80.8	93.01; 93.12
G2P1A0H1 gravid 38-39 minggu + panggul sempit + bekas sc	sctp	O33.8; O34.2; O82.0; Z37.0	74.1
ISPB	Infus cairan, injeksi antibiotik, nebulisasi	J22	
VL a/r frontalis dextra	Periksa dokter IGD Hecting 3 jahitan	S01.8	86.59
Weight feltering + Susp tb paru	kontrol selesai	Z09.8; R62.9	

Sampel penelitian terdiri dari 42.123 data dan setelah dilakukan proses cleaning seperti data berulang dan data null menjadi 32.208 data kodifikasi yang dikumpulkan dalam kurun satu tahun (Oktober 2024 – November 2025), dari data hasil rekomendasi dilakukan evaluasi model dengan pendekatan temporal hold-out dengan pembagian:

abel 2. Data Latih dan Data Uji

Data Latih Oktober 2024 – November 2025	Data Uji Desember 2025	Data Uji Januari 2026
32.208	857	878
32.208	1.735	

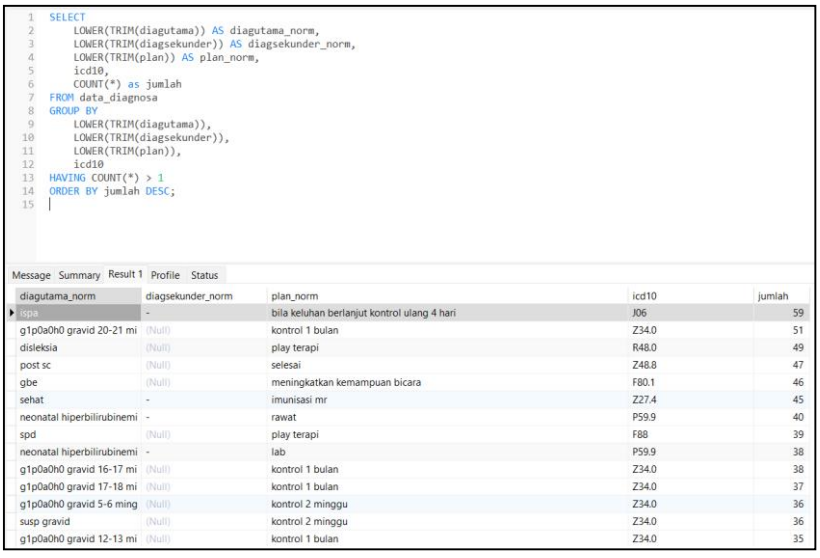


Gambar 3. Jumlah dan Persentase Data Latih dan Data Uji

Dalam konteks evaluasi medis pendekatan temporal hold-out dipilih untuk mensimulasikan implementasi nyata, di mana sistem digunakan untuk merekomendasikan kode pada periode setelah data historis tersedia, pendekatan lintas waktu yang digunakan menawarkan inovasi metodologis dibandingkan dengan validasi acak konvensional, karena lebih mencerminkan kondisi dunia nyata di mana pola misinformasi terus berkembang.[12]

2.1 Cleaning

Tahap cleaning dilakukan untuk memastikan kualitas data sebelum dilakukan proses embedding. Proses ini mencakup identifikasi dan penghapusan data duplikat menggunakan query SQL pada database rumah sakit. Duplikasi sangat berpengaruh dalam model berbasis vektor karena dapat menyebabkan bias representasi semantik, di mana data yang berulang akan mendominasi distribusi vektor dan menggeser kedekatan semantik antar-entri, sehingga menurunkan akurasi dan konsistensi rekomendasi ICD-10 dan ICD-9. Selain duplikasi, entri yang tidak memiliki informasi klinis yang lengkap atau tidak relevan juga dihapus untuk mencegah noise pada proses pembentukan sentence embeddings. Seluruh proses cleaning didokumentasikan dengan bukti visual berupa hasil query dan tampilan inspeksi data pada Navicat.



```sql
SELECT
    LOWER(TRIM(diagutama)) AS diagutama_norm,
    LOWER(TRIM(diagsekunder)) AS diagsekunder_norm,
    LOWER(TRIM(plan)) AS plan_norm,
    icd10,
    COUNT(*) as jumlah
FROM data_diagnosa
GROUP BY
    LOWER(TRIM(diagutama)),
    LOWER(TRIM(diagsekunder)),
    LOWER(TRIM(plan)),
    icd10
HAVING COUNT(*) > 1
ORDER BY jumlah DESC;
```

Message | Summary | Result 1 | Profile | Status

diagutama_norm	diagsekunder_norm	plan_norm	icd10	jumlah
ispa	-	bila keluhan berlanjut kontrol ulang 4 hari	J06	59
g1p0a0h0 gravid 20-21 mi	(Null)	kontrol 1 bulan	Z34.0	51
disleksia	(Null)	play terapi	R48.0	49
post sc	(Null)	selesai	Z48.8	47
gbe	(Null)	meningkatkan kemampuan bicara	F80.1	46
sehat	-	imunisasi mr	Z27.4	45
neonatal hiperbilirubinemi	-	rawat	P59.9	40
spd	(Null)	play terapi	F88	39
neonatal hiperbilirubinemi	-	lab	P59.9	38
g1p0a0h0 gravid 16-17 mi	(Null)	kontrol 1 bulan	Z34.0	38
g1p0a0h0 gravid 17-18 mi	(Null)	kontrol 1 bulan	Z34.0	37
g1p0a0h0 gravid 5-6 ming	(Null)	kontrol 2 minggu	Z34.0	36
susp gravid	(Null)	kontrol 2 minggu	Z34.0	36
g1p0a0h0 gravid 12-13 mi	(Null)	kontrol 1 bulan	Z34.0	35

Gambar 4. Proses Cleaning Pencarian Data Berulang

2.2 Preprocessing Data

Preprocessing terdiri dari dua tahap utama, yaitu normalisasi teks dan ekstraksi singkatan medis.

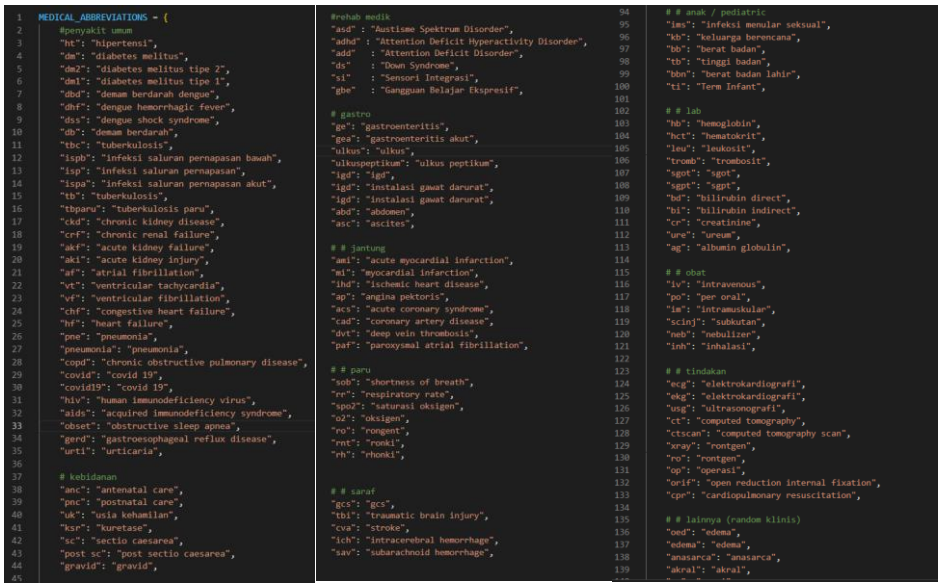
Tabel 3. Tahapan Preprocessing Data

No	Tahapan	Jenis
1	Normalisasi	Menghilangkan Spasi berlebih, pemisahan angka dan huruf, menggunakan filter regex
2	Ekstraksi	Melakukan ekstraksi singkatan-singkatan medis menggunakan fittur append python

Normalisasi diperlukan untuk menyamakan struktur penulisan diagnosis, memperbaiki pola teks, serta memastikan setiap entri berada dalam format yang konsisten sebelum dilakukan embedding. Teknik regex digunakan untuk menghapus karakter tak relevan, menormalkan spasi, dan melakukan pemisahan angka-huruf yang sering muncul pada penulisan diagnosis klinis.

2.3 Ekstraksi Data

Tahap ekstraksi merupakan komponen yang sangat signifikan dalam meningkatkan akurasi model. Teks diagnosis pada unit pelayanan ibu dan anak sering kali mengandung singkatan

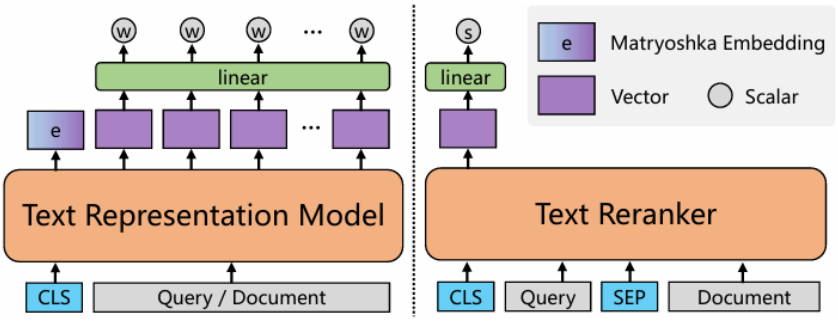


Gambar 5. Ekstraksi Data

Mengabaikan singkatan medis akan menyebabkan embedding kehilangan konteks, sehingga menurunkan nilai similarity. Dengan mengekstraksi singkatan, sistem dapat memperluas makna dan memperkaya representasi semantik.

2.4 Transformasi Data

Proses transformasi dilakukan dengan mengonversi seluruh 32.208 data diagnosis yang telah melalui tahap preprocessing ke dalam representasi vektor menggunakan model Sentence Transformer GTE-Multilingual-Base, model ini dipilih karena terbukti melalui penelitian dengan hasil pada benchmark pengambilan monolingual dan lintas bahasa menunjukkan bahwa Text Representation Model dan Rerangking mendekati peringkat yang lebih besar pada dataset reguler, dan mencapai kinerja yang lebih baik pada dataset konteks panjang. Ini berarti model lebih efisien untuk aplikasi industri.[13]



Gambar 6. Text Representation Model dan Text Reranker GTE-Multilingual Base

Setiap entri data terdiri dari Diagnosis Utama, Diagnosis Sekunder dan Rencana Tindakan (plan), Ketiga komponen tersebut digabungkan menjadi satu representasi teks sebelum dilakukan pemrosesan, dengan output target kode ICD-10 dan ICD-9-CM multi label, setiap entri data dilakukan juga tahapan preprocessing yang meliputi Normalisasi huruf, karakter dan singkatan medis, setelah preprocessing, teks dikonversi menjadi sentence embeddings menggunakan model SentenceTransformer. Embedding untuk seluruh data historis disimpan sebagai basis vektor referensi.

Tabel 4. Waktu Komputasi Embedding

Komponen	**Nilai**
Jumlah Data	32.208
Perangkat	Intel Core i7 Gen 8 (CPU)
GPU	Tidak Digunakan
Waktu Pemrosesan Total	±30 Menit
Rata-rata waktu per embedd	±50 milidetik
Output Size	224.292 KB

Proses embedding dilakukan pada perangkat dengan spesifikasi CPU Intel Core i7 generasi ke-8 tanpa akselerasi GPU, waktu komputasi untuk menghasilkan seluruh embedding adalah ±30 menit, yang menunjukkan bahwa pendekatan non-training inference dengan SentenceTransformer cukup efisien dan dapat dijalankan pada perangkat komputasi standar rumah sakit tanpa membutuhkan infrastruktur server berspesifikasi tinggi, temuan ini memperkuat argumen bahwa model berbasis semantic similarity tidak memerlukan sumber daya komputasi besar seperti model deep learning konvensional. Dengan demikian, pendekatan ini dinilai lebih adaptif dan mudah diimplementasikan pada fasilitas kesehatan dengan keterbatasan infrastruktur.

2.5 Evaluasi Model

Karena sistem dirancang sebagai alat rekomendasi pendukung (decision support), evaluasi berbasis peringkat lebih relevan dibandingkan akurasi klasifikasi tunggal, evaluasi model dilakukan menggunakan metrik peringkat yang umum digunakan pada sistem rekomendasi multi-label, yaitu Hit-K, Recall-K, Precission-K dan Mean Reciprocal Rank, evaluasi dilakukan pada K = 3 untuk mensimulasikan skenario rekomendasi tiga kandidat utama kepada petugas koding, hasil analisis digunakan untuk mengevaluasi efektivitas preprocessing dan memberikan rekomendasi pengembangan sistem.

Hit-K mengukur apakah minimal satu kode yang benar muncul dalam Top-K rekomendasi

$$Hit@K = \frac{1}{N}\sum_{i=1}^{N} 1\left(T_i \cap P_i^{(K)} \neq \varnothing\right)$$

Keterangan:
Ti = himpunan kode ICD yang benar (ground truth) pada data ke-i
Pi = himpunan kode ICD hasil rekomendasi model pada data ke-i
K = jumlah rekomendasi teratas yang dievaluasi
N = jumlah total data uji

Gambar 7. Rumus Hit-K dan Interpretasi

Rentang Nilai	Interpretasi Performa	Keterangan
0.90 – 1.00	Sangat Tinggi	Hampir seluruh kasus memiliki minimal satu kode benar dalam 3 rekomendasi teratas
0.75 – 0.89	Tinggi	Sebagian besar diagnosis berhasil direkomendasikan dengan tepat di Top-3
0.60 – 0.74	Cukup	Lebih dari setengah kasus memiliki minimal satu kode yang sesuai
0.50 – 0.59	Rendah	Model masih sering gagal menampilkan kode relevan dalam Top-3
< 0.50	Sangat Rendah	Mayoritas rekomendasi tidak mengandung kode yang benar

Recall-K mengukur proporsi kode benar yang berhasil muncul dalam Top-K rekomendasi

$$Recall@K = \frac{1}{N}\sum_{i=1}^{N} \frac{|T_i \cap P_i^{(K)}|}{|T_i|}$$

Keterangan:
Ti = himpunan kode ICD yang benar (ground truth) pada data ke-i
Pi = himpunan kode ICD hasil rekomendasi model pada data ke-i
K = jumlah rekomendasi teratas yang dievaluasi
N = jumlah total data uji

Gambar 8. Rumus Recall-K dan Interpretasi

Rentang Nilai	Interpretasi Performa	Keterangan
0.80 – 1.00	Sangat Baik	Sebagian besar kode yang benar berhasil ditangkap dalam Top-3
0.65 – 0.79	Baik	Mayoritas kode relevan berhasil ditemukan
0.50 – 0.64	Cukup	Sekitar setengah kode relevan terdeteksi
< 0.50	Kurang	Banyak kode relevan tidak muncul dalam rekomendasi

Precision-K mengukur proporsi rekomendasi dalam Top-K yang benar.

$$Precision@K = \frac{1}{N}\sum_{i=1}^{N} \frac{|T_i \cap P_i^{(K)}|}{K}$$

Keterangan:
Ti = himpunan kode ICD yang benar (ground truth) pada data ke-i
Pi = himpunan kode ICD hasil rekomendasi model pada data ke-i
K = jumlah rekomendasi teratas yang dievaluasi
N = jumlah total data uji

Gambar 9. Rumus dan Interpretasi Precision-K

Rentang Nilai	Interpretasi Performa	Keterangan
0.60 – 1.00	Sangat Baik	Sebagian besar rekomendasi Top-3 benar
0.45 – 0.59	Baik	Lebih dari separuh rekomendasi relevan
0.30 – 0.44	Cukup	Sebagian rekomendasi masih kurang tepat
< 0.30	Rendah	Banyak rekomendasi tidak relevan

MRR mengevaluasi posisi kemunculan kode relevan pertama dalam daftar rekomendasi

$$MRR = \frac{1}{N}\sum_{i=1}^{N} RR_i$$

Keterangan:
Ti = himpunan kode ICD yang benar (ground truth) pada data ke-i
Pi = himpunan kode ICD hasil rekomendasi model pada data ke-i
K = jumlah rekomendasi teratas yang dievaluasi
N = jumlah total data uji

Gambar 10. Rumus dan Interpretasi Mean Reciprocal Rank (MRR)

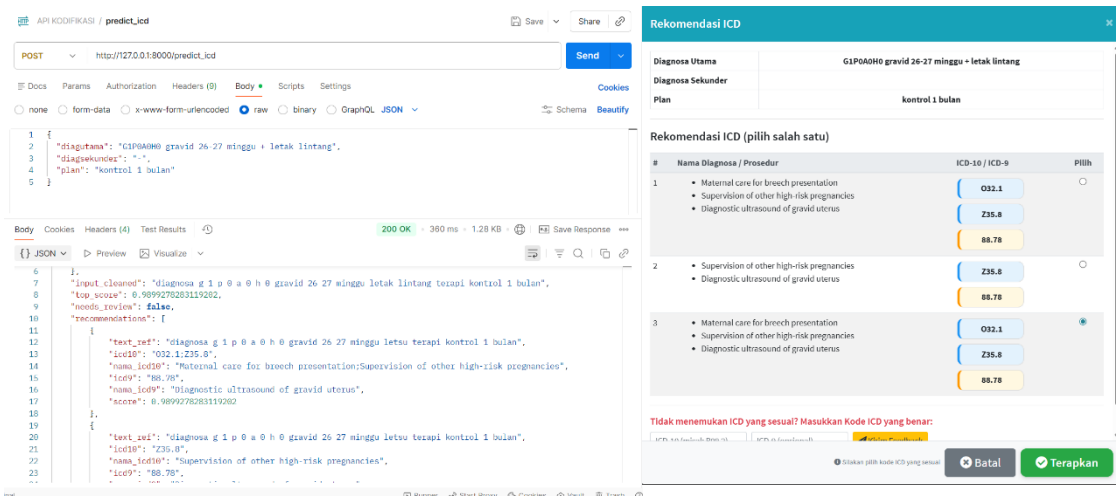
Rentang Nilai	Interpretasi Performa	Keterangan
≥ 0.50	Sangat Baik	Kode relevan umumnya muncul di peringkat 1–2
0.40 – 0.49	Baik	Kode relevan rata-rata muncul di peringkat 2–3
0.30 – 0.39	Cukup	Kode relevan sering muncul di posisi bawah Top-3
< 0.30	Rendah	Kode relevan jarang muncul di posisi awal

Data diperoleh secara sekunder dari database rumah sakit dan telah melalui proses verifikasi internal unit casemix. Seluruh data dianonimkan dengan menghilangkan identitas pasien sebelum pemrosesan. Penelitian dilakukan sesuai dengan prinsip kerahasiaan data pasien dan regulasi yang berlaku.

3. HASIL DAN PEMBAHASAN

Untuk memastikan model dapat digunakan dalam lingkungan operasional rumah sakit, sistem diimplementasikan dalam bentuk RESTful API yang dapat diakses oleh sistem informasi rumah sakit (SIMRS). API menerima input berupa teks diagnosis yang terdiri dari diagnosis utama, diagnosis sekunder, dan rencana tindakan, kemudian mengembalikan rekomendasi kode ICD-10 dan ICD-9 beserta skor kemiripan (cosine similarity) dalam skema Top-3

Pada pengujian kasus rawat inap, kode relevan umumnya muncul pada peringkat pertama atau kedua, selaras dengan nilai MRR yang berada di atas 0,4. Sementara itu, pada rawat jalan, sistem tetap mampu menampilkan kandidat kode yang relevan meskipun tingkat kecocokan lebih rendah dibandingkan rawat inap.



Gambar 11. Respons API Model dan Tampilan di User

3.1 Evaluasi Perbulan

Evaluasi model dilakukan menggunakan pendekatan temporal hold-out pada data Desember 2025 dan Januari 2026, hasil menunjukkan bahwa model memiliki performa yang stabil pada dua periode pengujian berbeda. Nilai Hit-k di atas 0,89 mengindikasikan bahwa pada lebih dari 89% kasus, minimal satu kode ICD yang benar muncul dalam tiga rekomendasi teratas. Recall-k yang konsisten di atas 0,78 menunjukkan bahwa sebagian besar kode relevan berhasil direkomendasikan oleh sistem, nilai Precision-3 berada pada kisaran 0,51–0,54, yang berarti sekitar setengah dari tiga kandidat rekomendasi merupakan kode yang benar. Sementara itu, nilai Mean Reciprocal Rank (MRR) di atas 0,4 menunjukkan bahwa kode yang relevan cenderung muncul pada peringkat awal rekomendasi.

Stabilitas performa antarperiode menunjukkan bahwa pendekatan retrieval berbasis semantic similarity cukup adaptif terhadap variasi data temporal tanpa memerlukan proses retraining model.

Tabel 5. Pengujian Model Per Periode Bulan

Bulan	Total Data	Hit-K	Recall	Precission	MRR
Des 2025	857	0,80775	0,7088	0,42055	0,4192
Jan 2026	878	0,77545	0,68735	0,40585	0,42965

3.2 Evaluasi Perbulan Perunit

Untuk mengevaluasi konsistensi performa pada karakteristik layanan yang berbeda, dilakukan analisis terpisah antara data rawat inap (ranap) dan rawat jalan (rajal), Hasil menunjukkan bahwa performa model pada kasus rawat inap secara konsisten lebih tinggi dibandingkan rawat jalan. Pada rawat inap, Hit-k mencapai lebih dari 0,89 pada kedua periode, sedangkan pada rawat jalan berkisar di sekitar 0,66.

Perbedaan ini dapat dijelaskan oleh karakteristik klinis masing-masing unit layanan. Diagnosis rawat inap umumnya lebih panjang, komprehensif, serta mencakup komorbiditas dan rencana tindakan yang lebih jelas. Kondisi ini menghasilkan representasi embedding yang lebih kaya secara semantik, sehingga meningkatkan akurasi retrieval., sebaliknya, diagnosis rawat jalan cenderung lebih singkat, sering kali berbentuk singkatan atau gejala utama saja, serta memiliki variasi penulisan yang lebih ekstrem. Representasi semantik pada teks yang lebih pendek menjadi kurang informatif, sehingga memengaruhi nilai similarity dan akurasi rekomendasi.

Meskipun demikian, nilai Hit-k pada rawat jalan yang berada di kisaran 0,65 menunjukkan bahwa sistem masih mampu memberikan minimal satu rekomendasi yang relevan pada sekitar dua pertiga

kasus, sehingga tetap memiliki nilai praktis sebagai alat bantu koding, Karena satu diagnosis dapat memiliki lebih dari satu kode ICD, evaluasi dilakukan menggunakan skema multi-label retrieval. Nilai Recall-k yang relatif tinggi menunjukkan bahwa sistem mampu menangkap sebagian besar kode relevan dalam tiga kandidat teratas, namun nilai Precision-k yang berada di kisaran 0,5 (ranap) dan 0,3 (rajal) menunjukkan bahwa masih terdapat kandidat yang tidak relevan dalam daftar rekomendasi. Hal ini wajar dalam sistem berbasis similarity ranking, terutama ketika beberapa kode memiliki konteks klinis yang berdekatan.

Nilai MRR yang berada di atas 0,4 menunjukkan bahwa kode yang benar umumnya muncul pada peringkat awal, sehingga dalam konteks Clinical Decision Support Tool (CDSS), sistem telah memenuhi fungsi utamanya sebagai pemberi rekomendasi awal yang cepat dan relevan.

Tabel 6. Pengujian Model Per Periode Bulan Per Unit

Bulan	Total Data	Hit-K	Recall	Precission	MRR
Rawat Jalan Des 2025	719	0,6662	0,6005	0,2976	0,4133
Rawat Inap Des 2025	138	0,9493	0,8171	0,5435	0,4251
Rawat Jalan Jan 2026	766	0,658	0,5929	0,2998	0,398
Rawat Inap Jan 2026	112	0,8929	0,7818	0,5119	0,4613

1)

4. KESIMPULAN

Hasil penelitian menunjukkan bahwa pendekatan NLP ringan berbasis sentence embeddings dan cosine similarity efektif digunakan sebagai sistem rekomendasi kode ICD dalam lingkungan operasional rumah sakit. Pada evaluasi temporal, model mencapai nilai Hit-3 sebesar 94,9% dan 89,3% pada kasus rawat inap periode Desember 2025 dan Januari 2026, dengan Recall-3 masing-masing 81,7% dan 78,1%, serta MRR di atas 0,42. Sebaliknya, pada rawat jalan diperoleh Hit-3 sekitar 66% dengan Recall-3 sekitar 59–60% dan MRR mendekati 0,40.

Perbedaan performa tersebut menunjukkan bahwa model bekerja lebih optimal pada dokumentasi klinis yang lebih lengkap, seperti pada kasus rawat inap. Sementara itu, performa pada rawat jalan membuka peluang pengembangan lanjutan, antara lain melalui pengayaan kamus normalisasi singkatan klinis, fine-tuning model untuk teks pendek, serta perbaikan standar operasional penulisan diagnosis dan rencana tindakan.

Sebagai sistem pendukung keputusan klinis (Clinical Decision Support System/CDSS), model ini tidak dimaksudkan untuk menggantikan perekam medis, melainkan untuk mempercepat proses seleksi kode awal dan meningkatkan konsistensi antarpetugas. Dengan pendekatan non-training inference, sistem dapat diperbarui secara berkala melalui regenerasi embedding tanpa perlu melakukan retraining model besar, sehingga lebih praktis bagi fasilitas kesehatan dengan sumber daya terbatas.

Secara keseluruhan, temuan ini memperkuat bahwa pendekatan retrieval-based semantic similarity merupakan solusi yang realistis, adaptif, dan implementatif untuk mendukung percepatan proses pengkodean serta klaim INA-CBG di fasilitas kesehatan Indonesia

Referensi

1. Agus Teguh Riadi, F. I. (2025). Cross-Temporal Generalization of IndoBERT for Indonesian Hoax News Classification. *Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 5, October 2025*, 5291-5304.
2. Belni Puspilasari, S. J. (2022). Analisis Perbedaan Kode Diagnosis ICD-10 Antara Rumah Sakit

Dengan Verifikator Bpjs Kesehatan. *Jurnal Keperawatan Priority, Vol 5, No 2, Juli 2022*, 25-36.
3. Chraibi, A., Delerue, David, Taillard, Julien, Draa, Ismat Chaib, Beuscart, Régis, & Hansske, Arnaud. (2021, 7). A deep learning framework for automated ICD-10 coding. *Public Health and Informatics: Proceedings of MIE 2021*, 347-351. doi:10.3233/SHTI210178
4. Kemenkes. (2021). Pedoman Indonesia Case Base Groups (INA-CBG). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 26 Tahun 2021*, 1-65.
5. Laksono, A. S. (2025). Studi Eksplorasi Metode Pengenalan Emosi Untuk Deteksi Multi-Emosi Dalam Entri Jurnal. *Skripsi Program Studi Informatika Fakultas Teknologi Industri Universitas Islam Indonesia*.
6. Maisharoh, D. S. (2025). Pembangunan Ulang Aplikasi Kecerdasan Buatan (Ai) Untuk Diagnosis, Kodifikasi, Dan Tindakan Kasus Kebidanan Di Rumah Sakit. *Journal of Information Technology and Computer Science (INTECOMS)Volume 8 Nomor 6*, 1767 - 1774.
7. Masud, J. H. (2023, 7 1). Applying Deep Learning Model to Predict Diagnosis Code of Medical Records. (S. Antani, Ed.) *Diagnostics, 13*. doi:10.3390/diagnostics13132297
8. Muhammad Abdul Hafizh Fathuddin, E. P. (2025). Penerapan Sentence-Bert dan Cosine Similarity untuk Pencarian Semantik Dokumen Skripsi dalam Format PDF. *Ranah Research Vol. 8, No. 1 (2025)* , 322 -337.
9. Muhammad Andika Fadilla, M. F. (2024). Pengembangan Sistem Klasifikasi Diagnosa Medis Menggunakan Progressive Web Application Terintegrasi Machine Learning. *Jurnal Syntax Admiration 5356 Vol. 5, No. 12*, 5488 - 5503.
10. Mulyana, S. (2021). Input Text Processing Using NLP In Case Based Reasoning To Help Diagnose Types Of Schizophrenic Disorder And Their Management. *Disertation Doctoral Faculty of Mathematics and Natural Sciences Universitas Gadjah Mada*.
11. Parjono, & Sri Kusumadewi. (2023, 9 16). Pemodelan Text Mining dalam Pengkodean Penyakit Pasien Berdasar Kode ICD 10. *Jurnal Nasional Teknologi dan Sistem Informasi, 9*, 200-207. doi:10.25077/teknosi.v9i2.2023.200-207
12. Priwibowo, A. (2025). Pembuatan Model Klasifikasi Teks Dengan BERT dan GRU untuk Pengkodean ICD-10 Diagnosis Utama Penyakit pada Proses Klaim BPJS di RSUP Persahabatan. *Tesis Magister Komputer Konsentrasi Sains Data Fakultas Teknologi Industri Universitas Islam Indonesia*.
13. Xin Zhang, Y. Z. (2024). mGTE:Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. *arXiv:2407.19669v1 29 Jul 2024*, 1-20.