TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

Pendekatan Translasi Otomatis Catatan Medis Indonesia untuk Ekstraksi Informasi dan Pemetaan Medis berbasis cTAKES–UMLS

Iwan Kasan[1,*], Lukman Heryawan[2], Ellya Qolina[3], Aliyah[1]

[1] Fakultas Teknik, Program Studi Teknik Informatika, Universitas Cendekia Abditama, Banten, Indonesia
[2] Fakultas Matematika dan Ilmu Pengetahuan Alam, Departemen Ilmu Komputer dan Elektronika, Universitas Gadjah Mada, Yogyakarta, Indonesia
[3] Fakultas Ilmu Keperawatan, Program Studi Keperawatan, Universitas Cendekia Abditama, Banten, Indonesia
Email: [1,*]iwan@uca.ac.id, [2]lukmanh@ugm.ac.id, [3]ellya_qolina@uca.ac.id, [4]aliyah@uca.ac.id
Email Penulis Korespondensi: iwan@uca.ac.id

Abstrak–Catatan medis tidak terstruktur dalam format SOAP merupakan aset krusial bagi analisis klinis, namun pemrosesan otomatisnya dalam bahasa Indonesia masih menghadapi tantangan besar karena keterbatasan dukungan teknologi NLP global. Penelitian ini mengevaluasi integrasi Apache cTAKES dan terminologi medis UMLS untuk mengekstraksi informasi medis dari catatan rekam medis elektronik berbahasa Indonesia. Permasalahan utama terletak pada arsitektur cTAKES yang dioptimalkan untuk bahasa Inggris, sehingga penerapan langsung pada teks berbahasa Indonesia menghasilkan kemampuan deteksi (Recall) yang sangat rendah, yaitu hanya 17,9%. Sebagai solusi pragmatis untuk menjembatani hambatan linguistik tersebut, penelitian ini mengusulkan alur pra-pemrosesan berbasis translasi otomatis menggunakan layanan Google Translate API sebelum dilakukan proses ekstraksi informasi oleh cTAKES. Evaluasi dilakukan terhadap 50 dokumen catatan medis format SOAP yang mencakup 840 entitas medis. Hasil eksperimen menunjukkan bahwa pendekatan translasi otomatis meningkatkan kemampuan deteksi entitas secara signifikan dengan nilai Recall 90,2% dan F1-Score 93,4%. Meskipun terdapat tantangan berupa kehilangan informasi pada singkatan medis lokal dan ambiguitas translasi, penelitian ini membuktikan bahwa translasi otomatis dapat menjadi strategi transisi yang efektif pada lingkungan dengan sumber daya terbatas. Pendekatan ini tidak hanya mendukung ekstraksi informasi klinis tetapi juga memungkinkan pemetaan terminologi medis ke standar internasional seperti ICD-10, SNOMED-CT, dan RxNorm secara otomatis untuk mendukung interoperabilitas data kesehatan nasional.

Kata Kunci: cTAKES; UMLS; Catatan Medis Pasien; SOAP; NLP; Google Translation API

Abstract–Unstructured medical notes in SOAP format are crucial assets for clinical analysis; however, their automated processing in the Indonesian language remains a significant challenge due to limited support from global NLP technologies. This study evaluates the integration of Apache cTAKES and the Unified Medical Language System (UMLS) to extract medical information from Indonesian electronic health records. The primary obstacle lies in the cTAKES architecture, which is optimized for English, causing direct application to Indonesian texts to yield a very low detection rate (Recall) of only 17.9%. As a pragmatic solution to bridge this linguistic barrier, this research proposes a preprocessing pipeline based on automatic translation using the Google Translate API prior to the cTAKES extraction process. The evaluation was conducted on a dataset of 50 SOAP-format medical records identifying 840 medical entities. Experimental results demonstrate that the automatic translation approach significantly improves entity detection, achieving a Recall of 90.2% and an F1-Score of 93.4%. Despite challenges such as information loss from local medical abbreviations and translation ambiguities, this study proves that automatic translation serves as an effective transitional strategy in resource-limited environments. This approach not only supports clinical information extraction but also enables the automatic mapping of medical terminology to international standards such as ICD-10, SNOMED-CT, and RxNorm to foster national health data interoperability.

Keywords: cTAKES; UMLS; Medical Notes; SOAP; NLP; Google Translation API

1. PENDAHULUAN

Catatan medis merupakan dokumen penting yang menyampaikan kondisi kesehatan pasien, hasil pemeriksaan, serta rencana terapi. Meski demikian, isi laporan sering dipenuhi istilah medis yang kompleks, singkatan, dan bahasa teknis yang sulit dipahami oleh orang yang tidak memiliki latar belakang kedokteran (Mariammal et al., 2025). *Electronic Health Record* (EHR) adalah representasi digital dari informasi klinis pasien yang disimpan secara aman oleh penyedia layanan kesehatan sepanjang waktu. EHR biasanya mencakup data demografi, rencana perawatan, diagnosis, resep obat, tanda vital, riwayat kesehatan, detail vaksinasi, serta hasil laboratorium. Sebagian besar informasi klinis dalam EHR tersedia dalam bentuk teks bebas yang tidak terstruktur, dan proses mengubahnya menjadi format terstruktur sering kali memakan waktu serta berisiko tidak menangkap seluruh aspek penting entitas medis (Sophie et al., 2022). Analisis dan visualisasi data terstruktur dari EHR relatif mudah dilakukan. Namun, proses manual untuk mengekstrak informasi penting dari catatan klinis bebas yang tidak terstruktur (Shafqat et al., 2023)sangat melelahkan, rawan kesalahan, dan biasanya memerlukan pengetahuan khusus di bidang klinis (Bai et al., 2021). Salah satu contoh catatan medis tidak terstruktur adalah format SOAP (Hermawan & Erfira, 2024), yang berisi catatan narasi medis dokter.

Penerapan *Electronic Health Record* (EHR) di Indonesia harus mengacu pada standar internasional seperti ICD-10 untuk klasifikasi penyakit, SNOMED-CT untuk terminologi klinis, LOINC untuk hasil laboratorium, serta standar obat yang diakui secara global, guna memastikan integrasi dan pertukaran data yang aman dan efisien antar sistem informasi kesehatan(Cetak Biru Transformasi Teknologi Kesehatan, 2024).

Perkembangan rekam medis elektronik (EHR) telah mendorong peningkatan volume dan kompleksitas data klinis secara signifikan. menyatakan bahwa data klinis saat ini tidak lagi berfungsi semata sebagai arsip administratif, melainkan telah berkembang menjadi aset strategis dalam pengembangan sistem pendukung keputusan klinis (*Clinical Decision Support Systems*) dan analitik kesehatan berbasis data (Bai et al., 2021). Pemanfaatan data EHR secara optimal

TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

memungkinkan institusi kesehatan untuk melakukan analisis tren penyakit, evaluasi mutu pelayanan, serta perumusan kebijakan kesehatan yang lebih berbasis bukti empiris.

Permasalahan tersebut mendorong berkembangnya pendekatan *Natural Language Processing* (NLP) sebagai solusi untuk mengekstraksi informasi klinis dari teks tidak terstruktur. NLP memungkinkan sistem komputer memahami, menginterpretasikan, dan memproses bahasa alami yang digunakan dalam catatan medis. Berbagai penelitian menunjukkan bahwa NLP dapat dimanfaatkan untuk mengekstraksi entitas klinis penting seperti diagnosis, gejala, prosedur medis, serta temuan klinis lainnya (Russel Hossain et al., 2024).

Salah satu platform NLP yang secara khusus dikembangkan untuk domain klinis adalah *Clinical Text Analysis and Knowledge Extraction System* (cTAKES) (Kim et al., 2025). Sistem ini merupakan perangkat lunak *open-source* yang dirancang untuk mengenali dan mengekstraksi konsep medis dari teks klinis secara otomatis (Dávila-García et al., 2026). cTAKES menyediakan berbagai komponen analisis bahasa alami, mulai dari *tokenization*, *part-of-speech tagging*, hingga *named entity recognition* (NER) yang memungkinkan sistem mengidentifikasi berbagai entitas klinis dalam dokumen medis.

Untuk memastikan konsistensi representasi konsep medis, hasil ekstraksi dari cTAKES umumnya dipetakan ke terminologi medis standar menggunakan Unified Medical Language System (UMLS). UMLS adalah sekumpulan berkas dan perangkat lunak yang menggabungkan lebih dari 200 kosakata serta standar kesehatan dan biomedis untuk memungkinkan interoperabilitas antar sistem (Chen et al., 2022) UMLS yang dikembangkan oleh National Library of Medicine (NLM) Amerika Serikat, merupakan basis data terminologi medis yang mengintegrasikan berbagai sistem klasifikasi kesehatan, seperti RxNorm ((Shamimul Hasan et al., 2023)), SNOMED-CT, ICD-10 (Tran et al., 2024).Melalui pemetaan ini, informasi klinis yang semula bersifat naratif dapat dikonversi menjadi data terstruktur yang interoperabel dan dapat dimanfaatkan dalam berbagai sistem informasi kesehatan.

Tantangan ini muncul secara signifikan dalam konteks teks klinis berbahasa Indonesia. Bahasa medis di Indonesia memiliki karakteristik unik berupa pencampuran istilah medis formal, istilah serapan asing, serta penggunaan singkatan lokal yang tidak terstandarisasi. Selain itu, teks tersebut cenderung kompleks secara linguistik, bersifat informal, dan jarang tersedia dalam bentuk korpus beranotasi (Ananda et al., 2025). Kondisi ini menciptakan hambatan linguistik yang nyata penerapan langsung model NLP (Vayadande et al., 2026) berbasis bahasa Inggris sering kali gagal mengenali entitas klinis karena ketidakcocokan leksikon dengan kamus internasional seperti UMLS.

Bagi sebagian besar negara non-berbahasa Inggris, tidak tersedia sumber daya seperti waktu, dana, dan keahlian manusia untuk membangun sistem terminologi terpadu dari awal, sebagaimana yang dilakukan oleh *National Institutes of Health*. Akibatnya, terdapat kebutuhan bagi negara-negara yang belum memiliki integrasi antar kosakata medis untuk mengembangkan metode komputasional yang mampu mewujudkan sistem terminologi terpadu, guna memfasilitasi pemrosesan informasi medis (misalnya di Tiongkok) untuk penelitian dan peningkatan klinis (Chen et al., 2023).

Beberapa pendekatan telah dilakukan mengatasi permasalahan tersebut, seperti membangun model NLP medis khusus bahasa Indonesia atau melakukan *fine-tuning* pada Large Language Models (LLM) dengan fokus pada NER medis berbahasa Indonesia. Misalnya, penelitian Abdillah et al. (2023) menerapkan BioNER, (Kusumawardani & Kusumawati, 2024) dengan Bi-LSTM-CRF model, sedangkan (Ananda et al. (2025) menggunakan BERT-CRF. Namun, pendekatan ini memerlukan ketersediaan korpus medis beranotasi dalam jumlah besar, tenaga ahli untuk memvalidasi setiap anotasi, serta sumber daya komputasi yang tinggi. Hingga saat ini, korpus berbahasa Indonesia yang tervalidasi oleh tenaga medis masih sangat terbatas. Alternatif lain berupa sistem berbasis aturan (*rule-based*) juga memiliki keterbatasan karena cenderung kaku, sulit beradaptasi terhadap variasi penulisan klinis yang dinamis (Purwitasari et al., 2021)), serta membutuhkan waktu dan tenaga untuk merumuskan aturan medis berbahasa Indonesia.

Berdasarkan tantangan tersebut, penelitian ini mengusulkan pendekatan alternatif berupa integrasi translasi otomatis sebagai tahap pra-pemrosesan dalam *pipeline* cTAKES. Melalui pendekatan ini, teks klinis berbahasa Indonesia terlebih dahulu diterjemahkan ke dalam bahasa Inggris sebelum diproses oleh sistem NLP klinis. Strategi ini sejalan dengan penelitian Iza et al. (2022), yang menerapkan translasi dari bahasa Spanyol ke bahasa Inggris, serta Ye et al. (2024) yang melakukan translasi dari bahasa Chinese ke bahasa Inggris. Dengan demikian, pendekatan ini memungkinkan pemanfaatan kamus terminologi medis berbasis UMLS tanpa perlu membangun model NLP medis baru yang khusus ditujukan untuk bahasa Indonesia.

Fokus utama penelitian ini adalah mengevaluasi efektivitas pendekatan translasi otomatis dalam meningkatkan kemampuan ekstraksi informasi medis serta pemetaan terminologi pada catatan SOAP (Subjective, Objective, Assessment, Plan) berbahasa Indonesia. Melalui pendekatan ini diharapkan dapat diperoleh bukti konsep (*proof of concept*) mengenai potensi penggunaan translasi otomatis sebagai solusi pragmatis untuk mendukung interoperabilitas semantik data klinis dalam ekosistem kesehatan digital di Indonesia, termasuk dalam mendukung inisiatif integrasi data kesehatan nasional berstandar internasional.

2. METODOLOGI PENELITIAN

2.1 Kerangka Dasar Penelitian

Penelitian ini menggunakan pendekatan eksperimental untuk mengevaluasi efektivitas ekstraksi informasi medis dari dokumen klinis tidak terstruktur berbahasa Indonesia. Fokus utama penelitian adalah mengatasi keterbatasan linguistik pada sistem NLP klinis berbasis bahasa Inggris melalui integrasi proses translasi otomatis sebagai tahap pra-

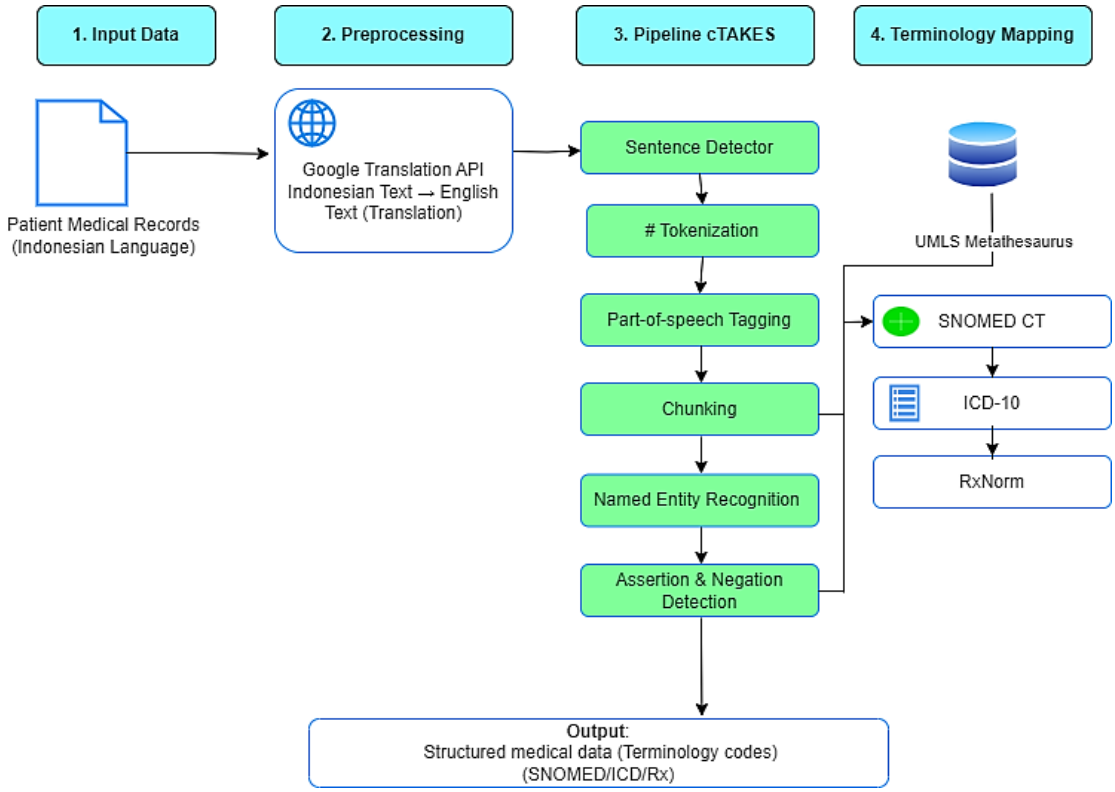
TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

pemrosesan. Pendekatan ini bertujuan untuk mengkaji sejauh mana integritas semantik informasi klinis dapat dipertahankan selama proses translasi serta bagaimana proses tersebut mempengaruhi kinerja ekstraksi entitas medis dan pemetaan terminologi standar menggunakan sistem NLP klinis.

Penelitian ini secara khusus menilai efektivitas *pipeline* yang mengintegrasikan translasi otomatis dengan sistem Apache cTAKES untuk melakukan ekstraksi informasi klinis dan pemetaan terminologi medis ke dalam standar internasional seperti SNOMED-CT, ICD-10, dan RxNorm melalui *metatesaurus* UMLS**.**

2.2 Tahapan Penelitian

Penelitian ini terdiri dari beberapa tahapan yang membentuk *pipeline* ekstraksi informasi medis. Alur kerja penelitian dapat dilihat pada Gambar 1.



Gambar 1. Arsitektur Proses Ekstraksi Informasi Medis

2.2.1 Akuisisi Data Klinis

Data penelitian berupa dokumen rekam medis dalam format SOAP (*Subjective*, *Objective*, *Assessment*, *Plan*) yang ditulis dalam bahasa Indonesia. Dokumen SOAP dipilih karena merupakan format pencatatan klinis yang umum digunakan oleh tenaga medis dalam mendokumentasikan kondisi pasien.
Dataset yang digunakan terdiri dari sejumlah dokumen SOAP yang telah melalui proses anonimisasi untuk menghilangkan informasi identitas pasien guna menjaga privasi dan kerahasiaan data medis.
Karakteristik data meliputi narasi klinis yang mengandung:

1. Istilah medis
2. Singkatan klinis
3. Variasi terminologi lokal
4. Struktur kalimat tidak baku yang umum ditemukan dalam dokumentasi dokter

Dokumen-dokumen tersebut kemudian digunakan sebagai sumber data untuk proses ekstraksi entitas medis.

2.2.2 Pra-pemrosesan: Translasi Otomatis

Karena *Apache* cTAKES dirancang untuk memproses teks medis berbahasa Inggris, tahap pra-pemrosesan dilakukan dengan menerjemahkan dokumen SOAP berbahasa Indonesia ke bahasa Inggris menggunakan Google Translation API (Gambar 1). Proses translasi dilakukan secara otomatis melalui layanan API dengan konfigurasi: *source language* Indonesian (id) dan target *language English* (en).

Tujuan dari tahap ini adalah mentransformasikan istilah medis lokal dan struktur kalimat bahasa Indonesia ke dalam bentuk bahasa Inggris yang kompatibel dengan pipeline analisis cTAKES.
Tahap translasi ini berfungsi sebagai mekanisme untuk mengatasi hambatan linguistik yang menjadi kendala utama dalam penerapan sistem NLP klinis berbasis bahasa Inggris pada data medis berbahasa Indonesia.

TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

2.2.3 Pipeline Analisis Menggunakan cTAKES

Teks hasil translasi selanjutnya diproses menggunakan *Apache Clinical Text Analysis and Knowledge Extraction System* (cTAKES) yang dibangun di atas *framework Unstructured Information Management Architecture* (UIMA) (Gambar 1). Pipeline analisis dalam cTAKES terdiri dari beberapa tahap anotasi linguistik, yaitu:

1. *Sentence Boundary Detection*
   Mengidentifikasi batas kalimat dalam dokumen klinis.
2. *Tokenization*
   Memecah teks menjadi unit token seperti kata, angka, atau simbol.
3. *Part-of-Speech* (POS) Tagging
   Memberikan label gramatikal pada setiap token.
4. *Shallow Parsing*
   Mengidentifikasi struktur frasa dalam kalimat untuk membantu analisis semantik.

Tahapan ini menghasilkan representasi linguistik yang digunakan pada proses ekstraksi entitas medis.

2.2.4 Named Entity Recognition dan Pemetaan Terminologi Medis

Setelah proses analisis linguistik, sistem melakukan Named Entity Recognition (NER) untuk mengidentifikasi entitas medis dalam teks. Proses NER pada cTAKES menggunakan mekanisme *Fast Dictionary Lookup* yang memanfaatkan kamus terminologi dari *Unified Medical Language System* (UMLS) *Metathesaurus*.
Entitas medis yang berhasil dikenali kemudian dipetakan ke dalam kode terminologi standar, antara lain:

1. SNOMED-CT untuk konsep klinis dan temuan medis
2. ICD-10 untuk klasifikasi diagnosis
3. RxNorm untuk terminologi farmasi dan obat-obatan

Pemetaan terminologi ini bertujuan untuk mengubah informasi klinis yang tidak terstruktur menjadi data terstandarisasi yang dapat digunakan untuk analisis lebih lanjut dalam sistem informasi kesehatan.

2.2.5 Analisis Konteks Klinis

Untuk meningkatkan akurasi interpretasi informasi medis, pipeline cTAKES juga menggunakan modul analisis konteks yang terdiri dari:

1. *Assertion Detection*
   Komponen ini menentukan status klinis dari suatu entitas, misalnya apakah kondisi tersebut: benar-benar dialami pasien, merupakan kondisi yang disangkal, hanya berupa kemungkinan atau rencana medis
2. *Negation Detection*
   Modul ini menggunakan algoritma NegEx untuk mendeteksi ekspresi negasi dalam teks klinis. Hal ini penting untuk membedakan antara temuan klinis yang benar-benar ada dengan kondisi yang secara eksplisit dinyatakan tidak dialami oleh pasien.

2.3 Validasi Hasil Ekstraksi

Validasi hasil ekstraksi dalam penelitian ini dilakukan dengan menyusun kumpulan data acuan (*ground truth*) yang dikurasi secara manual dari 50 dokumen rekam medis format SOAP dan dapat mengindentifikasi 840 entitas medis. Penyusunan data acuan ini mengacu pada terminologi medis standar yang terintegrasi dalam *Unified Medical Language System* (UMLS). Penggunaan UMLS memungkinkan pemetaan konsep klinis secara konsisten dan diakui secara internasional, sehingga berfungsi sebagai standar pembanding yang objektif.
Dalam proses validasi, entitas medis yang berhasil diekstraksi oleh sistem dipetakan dan diverifikasi ke dalam beberapa sistem terminologi standar yang tersedia di dalam UMLS, yaitu:

1. SNOMED CT untuk validasi konsep klinis dan temuan medis.
2. ICD-10 untuk validasi klasifikasi diagnosis penyakit.
3. RxNorm untuk validasi terminologi obat dan farmasi.

Proses verifikasi dilakukan dengan mencocokkan setiap entitas medis yang terdapat pada dokumen SOAP asli dengan konsep terminologi yang relevan dalam basis data UMLS. Data acuan ini dikurasi secara teliti untuk memastikan bahwa setiap anotasi mencerminkan kondisi klinis yang sebenarnya**.** Pendekatan ini digunakan sebagai referensi utama dalam mengevaluasi metrik *Precision*, *Recall*, dan *F1-Score* guna mengukur efektivitas arsitektur translasi otomatis yang diusulkan dalam mengatasi hambatan linguistik pada teks medis Indonesia.

2.4 Evaluasi Model

Untuk mengukur kinerja sistem, dilakukan perbandingan antara hasil ekstraksi otomatis dengan terminilogi medis standar pada UMLS. Metrik evaluasi yang digunakan adalah *Precision* (P), *Recall* (R), dan *F1-Score* (F1), yang dihitung dengan rumus berikut:

1. *Precision* (Presisi)
   Metrik ini mengukur sejauh mana entitas medis yang diekstraksi oleh sistem benar-benar relevan dan tepat.

TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

$$Precision = \frac{Entitas\ Terdeteksi\ yang\ Benar}{Total\ Seluruh\ Entitas\ yang\ Terdeteksi} \quad (1)$$

2. *Recall* (Sensitivitas)
   Metrik ini mengukur kemampuan sistem dalam menemukan seluruh entitas medis yang seharusnya ada di dalam dokumen asli. Berdasarkan hasil pengujian.

$$Recall = \frac{Entitas\ Terdeteksi\ yang\ Benar}{Total\ Entitas\ Medis\ yang\ Ada\ di\ Dokumen} \quad (2)$$

3. *F1-Score*
   F1-Score digunakan sebagai metrik keseimbangan antara *Precision* dan *Recall* untuk memberikan gambaran performa sistem secara keseluruhan.

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (3)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pemprosesan Catatan Medis

Pengujian sistem dilakukan terhadap 50 dokumen rekam medis dalam format SOAP yang digunakan sebagai referensi evaluasi dan entitas medis yang terindentifikasi sebanyak 840. Entitas tersebut terdiri dari diagnosis, prosedur medis, dan obat-obatan yang muncul dalam narasi klinis. Untuk memberikan gambaran proses ekstraksi informasi medis yang dilakukan oleh sistem, bagian ini menampilkan contoh hasil pemrosesan dari salah satu catatan medis. Contoh ini digunakan sebagai ilustrasi mekanisme ekstraksi entitas medis pada dua skenario pemrosesan, yaitu pemrosesan langsung pada teks berbahasa Indonesia dan pemrosesan setelah dilakukan translasi otomatis ke bahasa Inggris.

Pada pemrosesan langsung terhadap teks rekam medis berbahasa Indonesia, sistem hanya mampu mengenali sebagian kecil entitas medis yang terdapat dalam dokumen. Hal ini disebabkan oleh keterbatasan dukungan bahasa pada sistem Apache cTAKES, yang dirancang untuk memproses teks medis berbahasa Inggris. Sebaliknya, setelah dokumen melalui tahap translasi otomatis ke bahasa Inggris, sistem mampu mengenali lebih banyak entitas medis karena istilah medis telah diterjemahkan ke dalam bentuk yang kompatibel dengan pipeline analisis cTAKES.

Contoh pada bagian ini hanya merupakan ilustrasi dari satu catatan medis, sedangkan proses evaluasi dilakukan terhadap seluruh 50 dokumen SOAP yang terdapat dalam dataset penelitian. Hasil evaluasi agregat terhadap seluruh entitas tersebut kemudian disajikan pada bagian analisis kuantitatif untuk memberikan gambaran performa sistem secara keseluruhan.

3.1.1 Hasil Ekstraksi Informasi Medis

Berdasarkan hasil ekstraksi dari 50 dokumen SOAP (Tabel 1), sistem berhasil mengidentifikasi total 840 entitas medis dengan 305 entitas unik. Kategori Diseases & Disorders menjadi temuan paling dominan dengan frekuensi 315 kali muncul, disusul oleh Signs & Symptoms sebanyak 210 entitas medis. Tingginya jumlah frekuensi dibandingkan temuan unik menunjukkan adanya konsistensi pola keluhan dan diagnosis pada dataset yang diuji. Penggunaan CUI sebagai identitas unik memastikan bahwa variasi penulisan klinis dapat dipetakan ke dalam satu konsep standar yang sama dalam sistem UMLS.

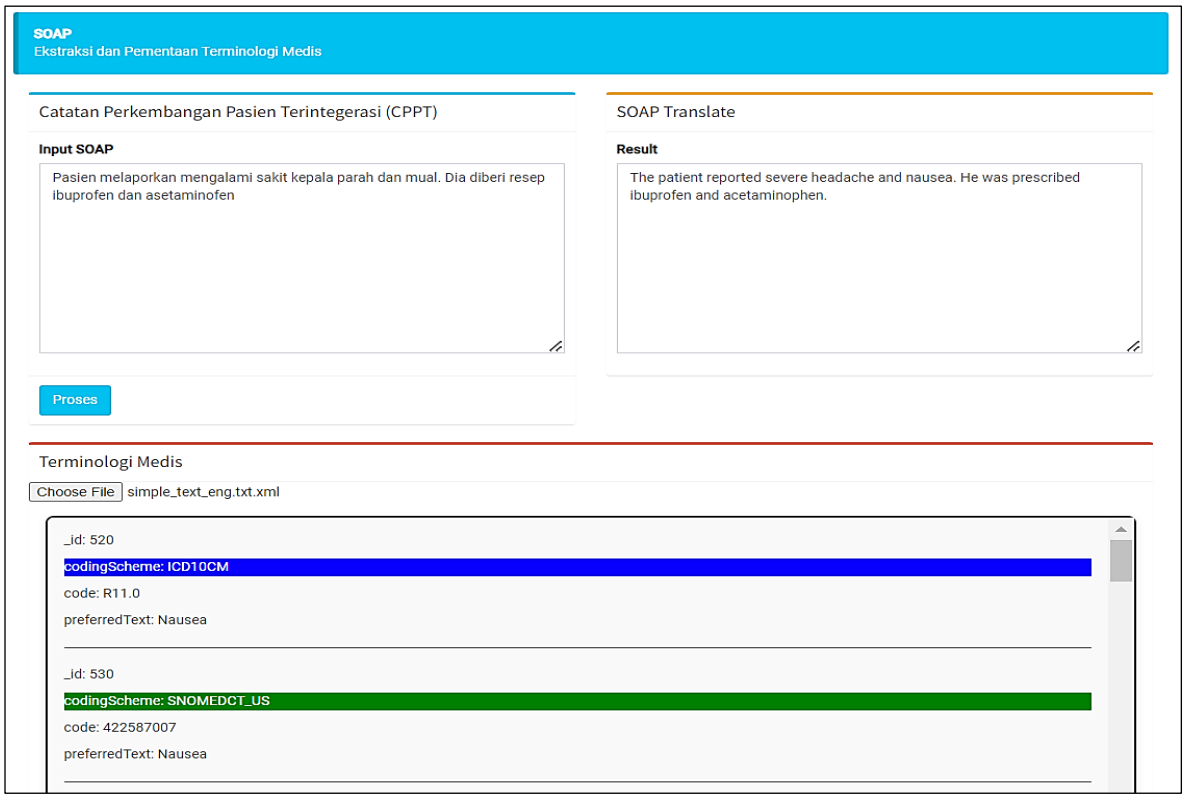
Tabel 1. Distribusi Entitas Klinis Berdasarkan Tipe Semantik UMLS

Kategori Entitas	Jumlah Temuan (Unique)	Frekuensi Total	Contoh Top CUI
Diseases & Disorders	124	315	C0018810 (Hypertension)
Signs & Symptoms	86	210	C0009806 (Chest Pain)
Anatomical Sites	45	185	C0000726 (Abdomen)
Medications/Drugs	32	90	C0003015 (Amlodipine)
Procedures	18	40	C0009924 (Chest X-Ray)
Total	305	840	

3.1.2 Ekstraksi Ekstraksi

Pengujian dilakukan dengan membandingkan efektivitas ekstraksi pada dua kondisi teks yang berbeda yang dapat dilihat pada Gambar 2, yang memberikan simulasi catatan medis berbahasa Indonesia lalu dilakukan tranlasi bahasa Inggris menggunakan Google Translate API.

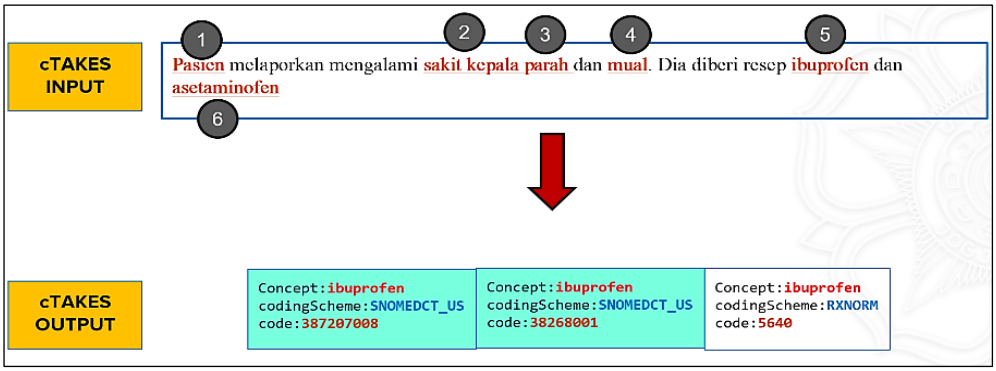
TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471



Gambar 2. Contoh Hasil Ekstraksi Teks Bahasa Indonesia

a. Ekstraksi Teks Asli (Bahasa Indonesia)
Pada simulasi pengujian menggunakan teks asli bahasa Indonesia, sistem hanya mampu mengidentifikasi satu entitas medis, yaitu "ibuprofen". Entitas lain seperti "sakit kepala", "mual", dan "asetaminofen" gagal dikenali karena keterbatasan kamus internal cTAKES terhadap istilah klinis non-Inggris, hasil ekstraksi teks bahasa Indonesia dapat dilihat pada Gambar 3.

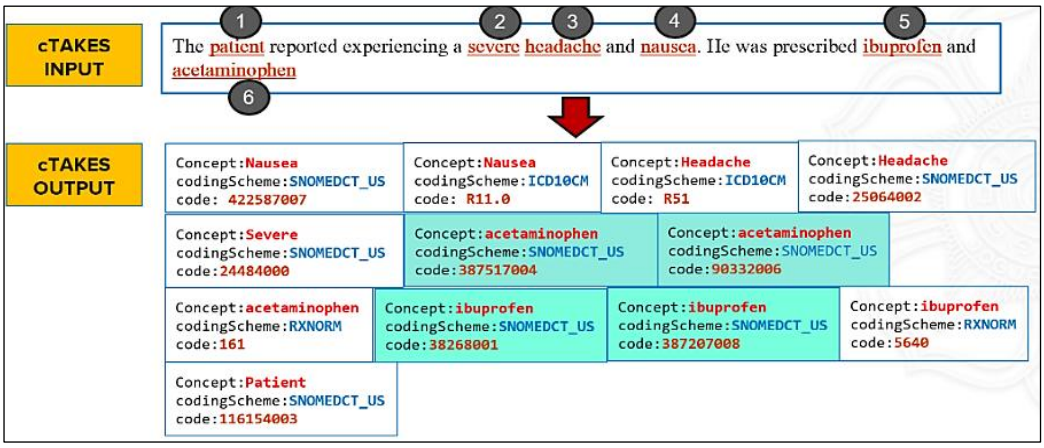
Meskipun integrasi translasi otomatis meningkatkan deteksi secara drastis, analisis lebih mendalam terhadap 50 dokumen SOAP mengungkap adanya hambatan pada singkatan medis lokal yang sangat spesifik. Sebagai contoh, singkatan seperti 'ddc' (*detur duo cap*) atau 'k/p' (*kalau perlu*) sering kali gagal diterjemahkan dengan benar oleh API, sehingga menyebabkan hilangnya informasi dosis atau aturan pakai dalam *pipeline* cTAKES. Selain itu, ditemukan pula munculnya entitas tambahan yang tidak relevan (*false positive*) akibat ambiguitas translasi; misalnya, kata 'suspek' yang terkadang diterjemahkan sebagai konteks hukum (*suspect*) alih-alih dugaan klinis, yang memicu pemetaan kode UMLS yang tidak sesuai dengan kondisi medis pasien.



Gambar 3. Contoh Ekstraksi Teks Bahasa Indonesia

b. Ekstraksi Teks Hasil Translasi (Bahasa Inggris)
Setelah teks diterjemahkan ke bahasa Inggris, performa sistem meningkat secara signifikan. Sistem berhasil mengidentifikasi enam entitas medis utama: *patient*, *severe*, *headache*, *nausea*, *ibuprofen*, dan *acetaminophen*. Seluruh entitas tersebut kemudian dipetakan secara otomatis ke dalam terminologi standar UMLS. Hasil ektraksi teks bahasa Inggris dapat dilihat pada Gambar 4.

TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471



Gambar 4. Contoh Ekstraksi teks bahasa Inggris

3.1.3 Analisis Pemetaan Terminologi (UMLS)

Seluruh entitas klinis yang berhasil diekstraksi dipetakan ke dalam berbagai skema pengkodean medis standar melalui integrasi pustaka Unified Medical Language System (UMLS), yang mencakup terminologi SNOMED-CT, klasifikasi penyakit ICD-10-CM, serta standar farmakoterapi RxNorm dapat dilihat pada Tabel 2.

Tabel 2. Contoh Representatif Hasil Pemetaan Entitas Medis pada Terminologi Standar (UMLS)

Entitas Klinis	*Semantic Type* (UMLS)	Kode Target	Sistem Terminologi	Peran dalam Sistem
Nausea	*Sign or Symptom*	422587007	SNOMED-CT	Dokumentasi Klinis
Headache	*Sign or Symptom*	R51	ICD-10-CM	Pelaporan Diagnostik
Ibuprofen	*Pharmacologic Substance*	5648	RxNorm	Keamanan Obat (*Drug Safety*)
Acetaminophen	*Pharmacologic Substance*	161	RxNorm	Keamanan Obat (*Drug Safety*)
Severe	*Qualifier Value*	24484000	SNOMED-CT	Penilaian Risiko/*Triase*

3.1.4 Evaluasi Metrik

Berdasarkan pengujian terhadap 840 entitas target yang terdapat dalam 50 dokumen SOAP, dilakukan perhitungan metrik evaluasi untuk membandingkan performa sistem sebelum dan sesudah tahap translasi otomatis. Hasil evaluasi disajikan pada Tabel 3.

Tabel 3. Perbandingan Peforma Eksraksi Informasi Medis

Skenario Pengujian	Target Entitas	Terdeteksi (TP)	Terlewat (FN)	Precision	Recall	F1-Score
Teks Bahasa Indonesia	840	151	689	100%	17,9%	30,3%
Teks Hasil Translasi	840	758	82	96,9%	90,2%	93,4%

3.2 Pembahasan

Hasil ekstraksi menunjukkan bahwa integrasi translasi otomatis efektif dalam mentransformasi narasi klinis lokal menjadi format yang dapat dikenali oleh standar global. Frekuensi total sebanyak 840 temuan entitas medis membuktikan bahwa hambatan leksikal bahasa Indonesia dapat diatasi tanpa perlu membangun kamus medis baru. Hal ini memberikan solusi pragmatis bagi sistem kesehatan di Indonesia untuk mencapai interoperabilitas semantik dengan standar internasional seperti SNOMED-CT, ICD-10, RxNorm melalui perantara bahasa Inggris. Penting untuk ditekankan bahwa kontribusi utama penelitian ini tidak hanya terletak pada kemampuan sistem dalam melakukan ekstraksi entitas medis (*Named Entity Recognition*), tetapi juga pada efektivitas pemetaan terminologi medis secara otomatis ke dalam standar internasional. Sebagaimana diberikan contoh pada Tabel 2, pendekatan translasi otomatis memberikan keunggulan kritis karena memungkinkan sinkronisasi langsung dengan *metatesaurus* UMLS yang sudah mapan. Berbeda dengan pendekatan kamus khusus (*rule-based*) yang memerlukan pemetaan manual yang repetitif dan kaku, penggunaan jalur translasi memungkinkan sistem secara dinamis memetakan variasi narasi klinis lokal ke dalam kode SNOMED-CT untuk temuan klinis, ICD-10 untuk diagnosis, dan RxNorm untuk farmasi. Kemampuan pemetaan ini sangat krusial bagi ekosistem entitas medis Indoneisa, di mana standarisasi antar-fasilitas kesehatan memerlukan pertukaran data berbasis kode dengan makna klinis seragam secara global.

Evaluasi menyeluruh yang disajikan pada Tabel 3 (Matriks Evaluasi) mengungkap perbedaan performa yang signifikan. Pemrosesan teks asli bahasa Indonesia menunjukkan nilai *Recall* yang sangat rendah (17,9%). Fenomena ini mengindikasikan terjadinya kehilangan informasi klinis (*clinical information loss*) secara sistemik jika sistem dipaksakan memproses bahasa lokal tanpa pra-pemrosesan. Sebaliknya, dengan mentransformasi teks melalui mesin penerjemah sebelum memasuki *pipeline* cTAKES, jumlah entitas yang terdeteksi (*True Positive*) meningkat drastis hingga mencapai nilai Recall 90,2% dan F1-Score 93,4%. Keberhasilan ini memberikan peluang besar bagi akselerasi

TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

integrasi data nasional, di mana otomatisasi pengkodean dapat mereduksi beban administratif dokter serta meminimalisir variabilitas interpretasi data.

Meski hasilnya impresif, analisis kritis terhadap translasi mengungkap beberapa catatan penting, seperti penurunan Precision (96,9%) akibat munculnya entitas tambahan yang tidak relevan (noise) karena ambiguitas kata. Misalnya, istilah "suspek" yang kadang diterjemahkan dalam konteks hukum (suspect) alih-alih konteks klinis. Selain itu, hambatan berupa singkatan lokal juga muncul, di mana beberapa entitas tetap terlewat (False Negative) akibat penggunaan singkatan medis spesifik (seperti "ddc" atau "k/p") yang gagal dikonversi oleh mesin penerjemah.

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa integrasi mesin penerjemah otomatis sebagai tahap pra-pemrosesan merupakan strategi yang sangat efektif untuk mengatasi hambatan linguistik dalam pemanfaatan teknologi NLP klinis global di Indonesia. Data eksperimen membuktikan bahwa penggunaan *Apache* cTAKES secara langsung pada teks medis berbahasa Indonesia menghasilkan tingkat deteksi yang sangat rendah dengan nilai Recall hanya 17,9%, yang menegaskan adanya risiko kehilangan informasi klinis yang signifikan jika sistem tidak diadaptasi secara linguistik. Namun, dengan penerapan jalur translasi otomatis, performa sistem meningkat secara drastis hingga mencapai F1-Score 93,4%, yang menunjukkan bahwa konteks semantik dan entitas klinis dari catatan SOAP lokal dapat dipertahankan dan dipetakan secara akurat ke dalam standar internasional seperti SNOMED-CT, ICD-10, dan RxNorm. Secara strategis, temuan ini memberikan solusi praktis bagi percepatan interoperabilitas data kesehatan nasional dalam ekosistem SATUSEHAT tanpa harus membangun model bahasa medis bahasa Indonesia dari nol yang membutuhkan sumber daya besar. Meskipun terdapat limitasi pada akurasi singkatan medis lokal yang spesifik, kerangka kerja (framework) ini mampu mentransformasi rekam medis tidak terstruktur menjadi data terstruktur yang diakui secara internasional. Dengan demikian, pendekatan ini berpotensi menjadi katalisator bagi transformasi digital kesehatan di Indonesia, memungkinkan pemanfaatan analisis data besar (big data analytics) untuk mendukung pengambilan keputusan medis yang lebih presisi serta perumusan kebijakan kesehatan publik yang berbasis data yang lebih akurat di masa depan.

REFERENCES

Abdillah, A. F., Purwitasari, D., Juniat, S., & Purnomo, H. H. (2023). Pengenalan entitas biomedis dalam teks konsultasi kesehatan online berbahasa Indonesia berbasis arsitektur transformers. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 10(1), 131–140

Ananda, N., Haryadi, D., & Fathiyana, R. Z. (2025). NER for Medical Component Classification in Doctor's Responses Using BERT-CRF. *2025 8th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 1109–1114. https://doi.org/10.1109/ISRITI68345.2025.11393439

Bai, L., Mulvenna, M. D., Wang, Z., & Bond, R. (2021, June 10). Clinical Entity Extraction: Comparison between MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical. *2021 32nd Irish Signals and Systems Conference, ISSC 2021*. https://doi.org/10.1109/ISSC52156.2021.9467856

Chen, L., Qi, Y., Wu, A., Deng, L., & Jiang, T. (2022). Enhancing Cross-lingual Medical Concept Alignment by Leveraging Synonyms and Translations of the Unified Medical Language System. *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, 2078–2083. https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00309

Chen, L., Qi, Y., Wu, A., Deng, L., & Jiang, T. (2023). Mapping Chinese Medical Entities to the Unified Medical Language System. *Health Data Science*, *3*. https://doi.org/10.34133/hds.0011

Dávila-García, D. M., Schuelke, M. J., & Wilcox, A. B. (2026). Lightweight open-source large language models versus cTAKES for information extraction from discharge summaries: tobacco smoking status test case. *JAMIA Open*, *9*(1). https://doi.org/10.1093/jamiaopen/ooaf182

Hermawan, E. (n.d.). Assessment of Medical Record Documentation and SOAP Completeness in Outpatient Services at a Primary Health Facility. In *International Journal of Health and Pharmaceutical*. Retrieved https://ijhp.net

Iza, J., Morejon, S., & Uyaguari, A. (2022). Automated Web Annotator of Biomedical Entities in Spanish Language. *Proceedings - 3rd International Conference on Information Systems and Software Technologies, ICI2ST 2022*, 72–78. https://doi.org/10.1109/ICI2ST57350.2022.00018

Kementerian Kesehatan Republik Indonesia. (2021). *Cetak Biru Strategi Transformasi Digital Kesehatan 2024*. Jakarta

Kim, M. H., Miramontes, S., Mehta, S., Schwartz, G. L., Kim, Y. J., Yang, Y., Hill-Jarrett, T. G., Cevallos, N., Chen, R., Glymour, M. M., Ferguson, E. L., Zimmerman, S. C., Choi, M., & Sims, K. D. (2025). Extracting Housing and Food Insecurity Information From Clinical Notes Using cTAKES. *Health Services Research*, *60*(S3). https://doi.org/10.1111/1475-6773.14440

Kusumawardani, R. P., & Kusumawati, K. N. (2024). Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstmcrf algorithm. *Procedia Computer Science*, *245*, 1146–1156. https://doi.org/10.1016/j.procs.2024.10.344

TIN: Terapan Informatika Nusantara
Vol 6, No 10, March 2026, page 1885-1893
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i10.9471

Mariammal, G., Swetha, K., & Samuel Jimrys, S. (2025). Medical Report Simplification System: Enhancing Healthcare Accessibility using NLP-based Extraction and AI-Driven Explanation. *Proceedings of 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2025*, 541–545. https://doi.org/10.1109/ICICV64824.2025.11085700

Purwitasari, D., Abdillah, A. F., Juanita, S., & Purnomo, M. H. (2021). Transfer Learning Approaches for Indonesian Biomedical Entity Recognition. *Proceedings of 2021 13th International Conference on Information and Communication Technology and System, ICTS 2021*, 348–353. https://doi.org/10.1109/ICTS52701.2021.9608496

Russel Hossain, M., Mahabub, S., Al Masum, A., & Jahan, I. (2024). *Natural Language Processing (NLP) in Analyzing Electronic Health Records for Better Decision Making*. https://doi.org/10.32996/jcsts

Shafqat, S., Anwar, Z., Javaid, Q., & Ahmad, H. F. (2023). *NER Sequence Embedding of Unified Medical Corpora to incorporate Semantic Intelligence in Big Data Healthcare Diagnostics*. https://doi.org/10.21203/rs.3.rs-3148503/v1

Shamimul Hasan, S. M., Agasthya, G., Santel, D., Bhatnagar, S., Goethert, I., Glauser, T., & Pestian, J. (2023). Application of Unified Medical Language System (UMLS) to Standardize Pediatric Drug Data. *Proceedings - 2023 IEEE 11th International Conference on Healthcare Informatics, ICHI 2023*, 753–755. https://doi.org/10.1109/ICHI57859.2023.00138

Sophie, S. L. M., Sathya, S. S., & Deepesh, C. (2022). Analyzing the Performance of Information Extraction System for Annotation of Patient Discharge Summary. *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2022*. https://doi.org/10.1109/IATMSI56455.2022.10119418

Tran, H. V., Tran, L. Q., Bui, T. P., Nguyen, V. V., & Nguyen, P. T. (2024). An Approach for Standardized Medical Terminology Machine Translation Using Pre-Trained Large Language Models. *Proceedings - International Conference on Knowledge and Systems Engineering, KSE*, 274–278. https://doi.org/10.1109/KSE63888.2024.11063657

Vayadande, K., Shinde, R., Bende, S., Sathe, H., Walunj, S., & Jha, S. (2026). *Specialized Large Language Models for Hindi Medical Natural Language Processing: A Clinical Entity in a Multi-Modal Framework Recognition and Semantic Understanding*. 1–8. https://doi.org/10.1109/ictbig68706.2025.11323575

Ye, Q., Yao, Z., Hu, P., Ji, X., Ruan, T., & Hou, R. (2024). Alignment of Chinese-English Medical Terminology in Small-Sample Scenarios: A Two-Stage Approach. *Proceedings - 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024*, 3908–3911. https://doi.org/10.1109/BIBM62325.2024.10821920