MULTIMODAL TEKS-NUMERIK: TINJAUAN SISTEMATIS TENTANG TEKNIK FUSI, KOMPATIBILITAS ENCODER, DAN KINERJA MODEL

Dipo Oktavian, Bayu Anugerah Putra
Teknik Informatika, Universitas Muhammadiyah Riau
Jalan KH. Ahmad Dahlan No.88, Sukajadi, Kota Pekanbaru, Riau, Indonesia
220401076@student.umri.ac.id

ABSTRAK

Integrasi data multimodal teks dan numerik menawarkan nilai prediktif yang lebih tinggi dibandingkan penggunaan modalitas tunggal. Namun, perbedaan karakteristik struktural antar-modalitas dan kompleksitas penyelarasan representasi sering menghasilkan performa suboptimal jika desain encoder dan mekanisme fusi tidak ditangani secara adaptif.Penelitian ini bertujuan mengkaji pengaruh arsitektur encoder dan strategi fusi terhadap kinerja model dalam skenario integrasi teks-numerik.Kajian dilakukan melalui pendekatan Systematic Literature Review (SLR) terhadap 28 studi terpilih dari rentang tahun 2019 hingga 2025.Temuan menunjukkan bahwa intermediate fusion mendominasi arsitektur multimodal, di mana mekanisme attention dan gating memberikan peningkatan performa yang lebih signifikan dibandingkan teknik concatenation. Analisis juga mengungkapkan bahwa kontribusi modalitas bersifat domain-specific; teks mendominasi tugas semantik, sementara numerik unggul pada data sinyal fisik, sehingga kompatibilitas antara encoder dan strategi fusi menjadi kunci efektivitas model.

Kata Kunci :*Multimodal learning, fusi data, teks dan numerik, Systematic Literature Review, deep learning.*

1. PENDAHULUAN

Dalam era digital saat ini, banyak fenomena data melibatkan integrasi data teks dan numerik secara bersamaan. Contohnya meliputi laporan finansial yang menggabungkan deskripsi naratif dan metrik keuangan [1], data medis berupa catatan klinis yang dipadukan dengan hasil tes laboratorium [2], serta data sensor IoT yang mencakup log deskriptif dan pengukuran kuantitatif [3]. Integrasi kedua modalitas ini terbukti memberikan nilai prediktif yang lebih baik dibandingkan penggunaan modalitas tunggal, karena teks menyediakan konteks semantik yang kaya, sementara data numerik menangkap pola kuantitatif yang stabil.

Peningkatan volume dan keberagaman data tersebut mendorong kebutuhan akan sistem kecerdasan buatan yang mampu memproses informasi multimodal secara simultan, khususnya pada sektor krusial seperti prediksi keuanga n dan diagnosis medis [4]. Namun, penggabungan data teks dan numerik menghadapi tantangan struktural yang signifikan. Teks bersifat tidak terstruktur dengan ketergantungan semantik yang kompleks, sedangkan data numerik umumnya terstruktur dalam bentuk tabular atau deret waktu dengan karakteristik distribusi yang berbeda. Perbedaan mendasar ini menjadikan penyelarasan representasi antar-modalitas sebagai tantangan utama, yang semakin diperumit oleh kondisi dunia nyata seperti ketidakseimbangan modalitas dan kehilangan data, yang dapat menurunkan keandalan dan akurasi model jika tidak ditangani secara adaptif [5].

Guna mengatasi kesenjangan representasi antara kedua jenis data tersebut, solusi utamanya terletak pada penggunaan arsitektur encoder yang mampu mengubah input mentah menjadi representasi laten yang kompatibel. Untuk menyelesaikan tantangan semantik pada teks, pendekatan berbasis Transformer telah diadopsi secara luas sebagai standar dominan karena kemampuannya yang superior dibandingkan metode tradisional [6]. Sementara itu, sebagai solusi untuk menangkap interaksi antar-fitur pada data numerik yang kompleks, penelitian kini beralih dari penggunaan Multi-Layer Perceptron menuju penerapan *deep tabular models* seperti FT-Transformer yang terbukti lebih efektif [7]. Meskipun encoder yang tepat merupakan langkah awal pemecahan masalah, keberhasilan integrasi sistem secara keseluruhan sangat bergantung pada strategi fusi yang diterapkan untuk menjembatani kedua representasi tersebut. Hal ini penting karena hubungan antara desain encoder dan mekanisme fusi sering kali belum terpetakan dengan baik, sehingga pendekatan yang tepat diperlukan untuk menghindari konfigurasi statis yang menghasilkan performa suboptimal [8].

Selanjutnya, sebagai mekanisme untuk menyatukan informasi dari kedua modalitas secara koheren, berbagai strategi fusi diterapkan sebagai solusi integrasi yang adaptif. Strategi ini secara sistematis dikelompokkan menjadi *early fusion*, *intermediate (joint) fusion*, *late fusion*, serta pendekatan *hybrid* untuk menangani interaksi data pada tingkat kedalaman yang berbeda [9]. Teknik-teknik spesifik seperti *concatenation*, *attention-based fusion*, hingga *Gating* dikembangkan sebagai jawaban teknis untuk memaksimalkan sinergi antar-modalitas. Penerapan teknik-teknik ini bertujuan untuk menyeimbangkan *trade-off* antara kapasitas ekspresif dan efisiensi komputasi, sekaligus memberikan ketahanan terhadap *noise*. Oleh karena itu, pemilihan strategi fusi yang adaptif menjadi solusi krusial dalam merancang model yang tidak hanya akurat tetapi juga

robust dalam menghadapi tantangan *multimodal learning* teks–numerik [10].

Berdasarkan konteks tersebut, studi ini bertujuan untuk mengkaji secara sistematis bagaimana pilihan arsitektur encoder multimodal, level fusi, dan teknik fusi memengaruhi kinerja model pada skenario integrasi data teks dan numerik di berbagai domain aplikasi. Kajian ini menjadi dasar penyusunan studi literatur yang diharapkan dapat memberikan pemahaman komprehensif serta panduan desain model multimodal yang lebih efektif.

2. TINJAUAN PUSTAKA

2.1. Multimodal DeepLearning

Multimodal Deep Learning merepresentasikan paradigma komputasi lanjutan yang dirancang untuk mengorkestrasi pemrosesan, penyelarasan (*alignment*), dan sintesis informasi dari berbagai modalitas data yang heterogen secara simultan. Mengingat setiap modalitas memiliki topologi struktural yang unik, tantangan utama dalam domain ini terletak pada konstruksi mekanisme *encoding* yang mampu menjembatani kesenjangan distribusi antar-data [10]. Literatur terkini mendikotomikan strategi representasi menjadi dua pendekatan fundamental: modality-specific representation dan shared representation. Pendekatan pertama menekankan preservasi fitur unik melalui encoder terpisah yang disesuaikan dengan karakteristik intrinsik data, sementara pendekatan kedua bertujuan memproyeksikan modalitas ke dalam ruang laten isomorfik untuk menangkap korelasi lintas-modal. Riset terbaru menunjukkan bahwa memodelkan kedua komponen ini secara eksplisit—memisahkan fitur spesifik (private) dan fitur bersama (shared)—menghasilkan representasi yang lebih robust dan diskriminatif dibandingkan hanya menggunakan salah satunya[11]. Dalam konteks spesifik integrasi data tekstual dan tabular, sinergi ini bermanifestasi pada penggabungan narasi tidak terstruktur (seperti laporan finansial, clinical notes, atau log operasional) dengan metrik kuantitatif terstruktur (seperti rasio keuangan, data laboratorium EHR, atau time-series sensor IoT). Studi empiris mengonfirmasi bahwa integrasi kedua modalitas ini menghasilkan performa prediktif yang superior dibandingkan model unimodal, di mana informasi tekstual menyediakan konteks semantik yang kaya, sementara data numerik memvalidasi pola kuantitatif yang stabil dan objektif[12].

2.2. Arsitektur Multimodal

Arsitektur multimodal dapat dikategorikan menjadi beberapa paradigma utama dalam proses integrasi modalitas:

Single-steam Architecture Pada pendekatan ini, seluruh modalitas digabungkan menjadi satu representasi awal yang kemudian diproses oleh sebuah model tunggal. Meskipun sederhana, pendekatan ini rentan terhadap ketidaksesuaian struktur antar modalitas. Contoh implementasinya secara terbatas dapat ditemukan pada model sederhana penggabungan teks dan tabular berbasis concatenation awal atau Transformer yang menerima embedding gabungan, sebagaimana dibuktikan dalam studi TabLLM yang menunjukkan potensi sekaligus batasan dari penyatuan paksa kedua modalitas ini [12].

Dual-stream Architecture memiliki perbedaan Berbeda dengan pendekatan aliran tunggal, Dual-stream (atau sering disebut Two-tower) Architecture mengadopsi prinsip architectural decoupling, di mana setiap modalitas diproses melalui jalur encoder independen yang terisolasi sebelum dilakukan fusi representasi pada lapisan intermediate atau tahap akhir jaringan. Strategi paralelisasi ini memungkinkan model untuk menangkap topologi feature manifold yang unik dari masing-masing modalitas—seperti dependensi sekuensial pada teks dan korelasi statistik pada data tabular—tanpa risiko distorsi fitur akibat pencampuran dini[11]. Desain ini telah menjadi paradigma dominan dalam literatur terkini karena fleksibilitasnya; contoh konkret terlihat pada model prediksi risiko finansial modern yang menggunakan LSTM untuk aliran numerik dan BERT untuk aliran teks berita secara terpisah [1], serta dalam sistem Rekam Medis Elektronik (EHR) multimodal di mana catatan klinis tidak terstruktur dan data *time-series* vital diproses secara paralel untuk memaksimalkan akurasi prognostik [13]

Multi-Branch Architecture Arsitektur ini merupakan pengembangan dari desain dual-stream yang dirancang untuk menangani berbagai jenis data dengan karakteristik berbeda. Dalam sistem ini, input tidak hanya terbatas pada dua jenis data melainkan mencakup berbagai aliran informasi yang beragam, seperti teks naratif, data angka yang tidak berubah, sinyal berupa data waktu dari proses fisiologis, serta informasi demografis. Dalam struktur ini, setiap cabang memiliki fungsi khusus sebagai ekstraktor fitur yang memproses satu jenis data secara terpisah, sehingga memungkinkan model menangkap ciri khas dari setiap jenis data tanpa mengganggu proses lainnya. Pendekatan ini memiliki banyak manfaat dan menjadi pilihan utama dalam pembuatan sistem pendukung pengambilan keputusan medis, seperti yang terlihat dari perkembangan model prediksi risiko berbasis catatan kesehatan elektronik (EHR), mulai dari model MEDFuse hingga arsitektur modern berbasis jaringan saraf graf (seperti MINGLE dan MEME). Penelitian terbaru menunjukkan bahwa arsitektur multi-cabang yang dilengkapi dengan mekanisme penggabungan data secara hierarkis mampu memberikan gambaran pasien yang lebih lengkap dan stabil dibandingkan dengan metode yang menggabungkan data pada tahap awal proses [14].

2.3. Level Fusi

Level fusion merujuk pada di tahap mana modalitas digabungkan di dalam pipeline model. Literatur umum membedakan paling tidak tiga level:

early fusion (signal/feature-level), intermediate atau joint fusion, dan late/decision-level fusion; beberapa studi juga memperkenalkan desain hybrid yang menggabungkan lebih dari satu level sekaligus.

Early Fusion adalah jenis fusi yang berlangsung pada tingkat data dan biasanya terjadi dalam tahap input di setiap cabang. Data mentah dipetakan ke suatu ruang yang sama melalui teknologi penyelarasan dan translasi data, lalu data multi-modal mengalami difusi untuk menghasilkan bentuk data yang lebih kaya dan ekspresif. Keunggulan dari pendekatan ini dapat langsung mempelajari interaksi antar modalitas, namun rentang terhadap ketidakseimbangan skala dan struktur fitur [15].

Intermediate / joint Fusion menggabungkan representasi laten yang sudah diproses encoder per-modalitas, biasanya di lapisan menengah, sehingga encoder masih bisa belajar fitur spesifik modalitas sebelum berinteraksi. Pendekatan ini adalah strategi yang paling umum karena menyediakan keseimbangan antara fleksibilitas dan expressiveness. Contohnya termasuk attention-based fusion, bilinear fusion, atau sketch-based projection.[8], [15]

Late Fusion menggabungkan output prediksi atau skor dari model per-modalitas (misalnya rata-rata atau voting berbobot), lebih sederhana namun sering kali kurang menangkap interaksi halus antar fitur. Pendekatan ini mengurangi risiko overfitting tetapi mengurangi kemampuan model menangkap interaksi mendalam antar fitur lintas modalitas.[15]

Hybrid fusion mengombinasikan lebih dari satu titik penggabungan, misalnya early fusion untuk subset fitur numerik dan intermediate fusion dengan teks, atau penggabungan berlapis (hierarchical fusion), Pendekatan ini memungkinkan model untuk mengeksploitasi keuntungan ganda: menangkap korelasi tingkat rendah melalui fusi awal pada subset modalitas yang kompatibel, sekaligus memodelkan abstraksi semantik tingkat tinggi melalui fusi menengah atau akhir [15].

2.4. Teknik Fusi

Di dalam tiap level fusion, berbagai teknik telah diusulkan untuk menggabungkan embedding teks dan numerik. Pendekatan paling sederhana adalah concatenation, yaitu menggabungkan vektor representasi dari tiap modalitas kemudian meneruskannya ke lapisan MLP atau classifier; teknik ini banyak dipakai sebagai baseline dan sering tetap kompetitif. Di luar concatenation, banyak studi menggunakan attention-based fusion, termasuk self-attention dan cross-attention untuk memodelkan interaksi antar fitur lintas modalitas secara lebih selektif dan kontekstual [8].

Concatenation Sebagai teknik fusi yang paling fundamental, *concatenation* beroperasi dengan menyatukan vektor fitur dari *encoder* tiap modalitas menjadi satu representasi gabungan sebelum diteruskan ke lapisan pemrosesan akhir. Dalam literatur terkini, metode ini sering diklasifikasikan sebagai bagian dari strategi *Early Fusion* atau *feature-level fusion*, di mana interaksi lintas modalitas (*cross-modal interaction*) dimodelkan sejak tahap awal untuk menghasilkan *embedding* terpadu yang kaya konteks. Meskipun pendekatan ini menawarkan efisiensi komputasi dan sering dijadikan *baseline* yang stabil, penerapannya menuntut penyelarasan dimensi yang cermat untuk menghindari distorsi representasi akibat inkompatibilitas struktur antar data yang berbeda[16].

Attention-based Fusion Berbeda dengan metode penggabungan statis, Attention-based Fusion menerapkan mekanisme seleksi dinamis yang memungkinkan model untuk memusatkan fokus pada fitur paling informatif dari setiap modalitas secara adaptif. Teknik ini sering diimplementasikan melalui mekanisme seperti Gated Fusion atau modul atensi spasial-kanal (seperti CBAM), yang bekerja dengan menghitung bobot signifikansi untuk mengatur besaran kontribusi sinyal—misalnya memprioritaskan pola visual grafik dibandingkan data numerik saat volatilitas pasar meningkat[17].

Self-Attention Mechanism Sebagai evolusi dari pemrosesan sekuensial statis, mekanisme *Self-Attention* memberdayakan model untuk secara otomatis mengidentifikasi dan memprioritaskan segmen data yang paling relevan dalam setiap modalitas sebelum integrasi dilakukan[18]. Pada dasarnya Self-Attention merupakan mekanisme yang memungkinkan model untuk secara otomatis memilah dan memfokuskan perhatian pada bagian terpenting dari datanya sendiri sebelum digabungkan dengan data lainCross-Attention Mechanism Sebagai metode fusi tingkat lanjut, Cross-Attention dirancang untuk memfasilitasi penyelarasan langsung antara dua modalitas yang memiliki karakteristik struktural berbeda, seperti teks dan citra[19]. Cross-Attention adalah mekanisme fusi yang menjembatani dua modalitas berbeda dengan cara memetakan hubungan langsung antara elemen spesifik dari satu data ke data lainnya.

2.5. Penelitian Terdahulu

Beberapa penelitian terdahulu telah memanfaatkan integrasi data teks tidak terstruktur dan data numerik terstruktur untuk meningkatkan performa prediksi pada berbagai domain. Tavakoli et al [1] melakukan prediksi risiko kredit dan pergerakan saham dengan menggabungkan *earnings call transcripts* sebagai data teks dan data numerik pasar keuangan. Penelitian ini menggunakan encoder Transformer (BERT) untuk teks dan CNN untuk data numerik, serta menerapkan strategi *hybrid fusion* yang mengombinasikan *early fusion* pada data numerik dan *intermediate fusion* melalui konkatenasi dan *cross-attention*. Hasilnya menunjukkan bahwa model multimodal secara konsisten mengungguli model unimodal, dengan kontribusi teks yang lebih dominan dalam menangkap sentimen dan konteks ekonomi yang tidak tercermin pada data numerik historis. Temuan ini juga menunjukkan bahwa pemilihan

encoder numerik yang kompatibel, seperti CNN, berperan penting dalam efektivitas fusi dengan representasi teks berbasis Transformer.

Pendekatan multimodal serupa juga diterapkan oleh Mou et al [20] dalam peramalan harga valuta asing dan emas, dengan mengintegrasikan teks berita keuangan dan data time-series harga. Studi ini menggunakan FinBERT sebagai encoder teks dan iTransformer untuk data numerik, lalu melakukan *intermediate fusion* dengan memasukkan embedding teks sebagai token variabel tambahan dalam mekanisme *self-attention* dan *cross-attention*. Hasil eksperimen menunjukkan penurunan nilai MSE yang signifikan dibandingkan model iTransformer unimodal, serta menegaskan pentingnya penyelarasan temporal antara teks dan data numerik. Pada domain kesehatan, Lee et al [21] mengusulkan pendekatan yang berbeda dengan mengonversi data tabular EHR menjadi *pseudo-notes* sehingga seluruh modalitas dapat diproses oleh encoder teks yang sama (MedBERT). Representasi dari masing-masing modalitas kemudian difusikan pada level embedding menggunakan *self-attention*, yang terbukti lebih efektif dibandingkan *early fusion* langsung pada level input.

Meskipun penelitian-penelitian tersebut telah menunjukkan keberhasilan integrasi teks dan numerik melalui berbagai kombinasi encoder dan strategi fusi, masing-masing studi masih berfokus pada skenario dan konfigurasi tertentu. Perbedaan dalam pemilihan level fusi (early, intermediate, hybrid), teknik fusi (konkatenasi, *cross-attention*, gating), serta jenis encoder yang digunakan menunjukkan belum adanya pemetaan yang terstruktur mengenai kecenderungan dan efektivitas pendekatan multimodal teks–numerik. Oleh karena itu, artikel ini bertujuan untuk meninjau dan menganalisis secara sistematis penelitian-penelitian terdahulu terkait teknik fusi, level fusi, dan arsitektur encoder pada multimodalitas teks dan numerik, guna memberikan gambaran komprehensif sebagai landasan pengembangan penelitian selanjutnya.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan Systematic Literature Review (SLR) untuk meninjau dan mengevaluasi penelitian-penelitian mengenai Multimodal teks dan numerik. Pendekatan SLR dipilih karena memberikan cara yang terstruktur dalam mengidentifikasi, mengevaluasi, dan mensintesis hasil penelitian sebelumnya guna menjawab pertanyaan penelitian yang telah dirumuskan.Penelitian ini bertujuan untuk membantu menjawab tiga pertanyaan penelitian yang relevan

PERTANYAAN PENELITIAN

RQ1: Bagaimana desain level fusion (early, intermediate, late, hybrid) dan teknik fusion (concatenation, attention-based, dsb.) digunakan pada model multimodal untuk data teks dan modalitas berbeda, dan bagaimana dampaknya terhadap performa model?

RQ2: Bagaimana kontribusi relatif modality teks dan modalitas lain terhadap performa model multimodal, termasuk dalam kondisi khusus seperti data imbalanced, krisis, atau noise pada salah satu modality?

RQ3: Bagaimana kombinasi arsitektur encoder teks dan modalitas berbeda (misalnya CNN, RNN, transformer, MLP) mempengaruhi efektivitas strategi fusion pada berbagai domain aplikasi modalitas Teks dan modalitas berbeda?

3.1. Protokol Peninjauan

Artikel dicari melalui basis data bereputasi seperti Scopus dan portal jurnal besar seperti ScientDirect, ReserchGate, ACL, MDPI, JMIR, dan arXiv menggunakan kombinasi kata kunci dalam bahasa Inggris, misalnya: “multimodal”, “text and tabular”, “text and numeric”, “fusion”, “deep learning”, “risk prediction”, “financial forecasting”, “multimodal EHR”, dan sejenisnya. Selain itu, dilakukan penelusuran lanjutan dari artikel kunci yang relevan seperti studi multimodal EHR, multimodal time series forecasting, dan survei teknik data fusion.

3.2. Kriteria Inklusi dan Eksklusi

Kriteria inklusi dan eksklusi ditetapkan untuk memastikan bahwa artikel yang di analisis relevan dengan tujuan dan pertanyaan penelitian. Adapun kriterian iklusi dan ekslusi yang digunakan sebagai berikut :

a. Terbit dalam 2019 - 2025 dengan beberapa pengecualian untuk karya dasar yang sering dirujuk.
b. Mengusulkan atau mengevaluasi model multimodal yang melibatkan teks + numerik/tabular, dan hanya sebagian kecil yang berbasis teks + citra bila memberikan insight ke desain fusion
c. Menggunakan metode berbasis deep learning atau transformer/hypergraph/LLM sehingga relevan dengan pembahasan arsitektur encoder dan fusion tingkat lanjut
d. Artikel berbahasa Inggris yang terbit di jurnal atau prosiding terindeks (Scopus/ISI) atau repository preprint bereputasi

Artikel di keluarkan bila :

a. Hanya menggunakan satu modalitas (murni teks atau murni numerik) tanpa fusion eksplisit.
b. Memakai multimodal tetapi di luar fokus teks–numerik (misalnya murni audio–visual) dan tidak membahas fusion yang dapat digeneralisasi ke kasus teks–tabular.
c. Berupa tulisan non-ilmiah seperti editorial singkat atau artikel tanpa detail metodologis.

3.3. Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan melalui penelusuran literatur secara sistematis terhadap artikel ilmiah yang membahas pemodelan multimodal dengan integrasi data teks dan numerik. Pencarian artikel dilakukan pada beberapa basis data bereputasi seperti yang sudah dijelaskan dalam Protokol peninjauan. Proses penyaringan selanjutnya dilakukan berdasarkan kriteria inklusi dan eksklusi yang telah ditetapkan, mencakup tahun publikasi, jenis modalitas data, dan pendekatan metodologis yang digunakan. Setelah seluruh tahapan penyaringan dilakukan, diperoleh 28 artikel yang dinyatakan relevan dan layak untuk dianalisis lebih lanjut, sebagaimana ditampilkan pada Table 1 yang menjadi dasar pembahasan dalam studi literatur ini.

Tabel 1. Daftar jurnal relevan

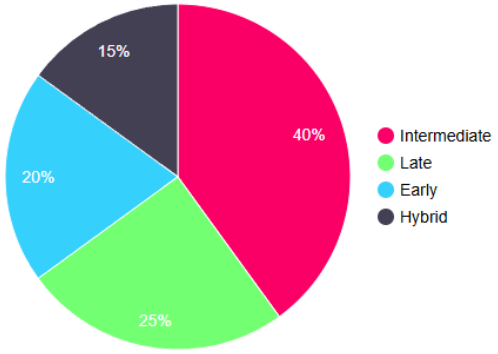
Artikel	RQ1	RQ2	RQ3
[1]	✓	✓	✓
[22]	✓		✓
[20]	✓	✓	✓
[23]	✓	✓	✓
[24]	✓	✓	✓
[25]	✓	✓	✓
[26]	✓	✓	✓
[21]	✓	✓	✓
[27]	✓	✓	✓
[28]	✓	✓	✓
[29]	✓	✓	✓
[30]	✓	✓	✓
[31]	✓	✓	✓
[32]	✓	✓	✓
[33]	✓	✓	✓
[34]	✓		✓
[35]	✓	✓	
[36]	✓		✓
[37]	✓	✓	✓
[38]	✓	✓	✓
[39]	✓	✓	✓
[40]	✓	✓	
[41]	✓	✓	✓
[42]	✓	✓	✓
[43]	✓	✓	
[44]		✓	
[45]			✓
[46]			✓

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan sintesis komprehensif dari 28 studi terpilih yang telah melalui proses penyaringan ketat berdasarkan kriteria inklusi dan eksklusi yang ditetapkan sebelumnya.Artikel-artikel yang digunakan Adalah artikel dari rentang waktu 2019-2025 yang merepresentasikan periode perkembangan pesat dalam arsitektur *deep learning* multimodal. Fokus utama dari tinjauan ini adalah pada model yang mengintegrasikan data teks tidak terstruktur dengan data numerik atau tabular terstruktur, sebuah kombinasi yang sangat relevan untuk domain dengan kompleksitas tinggi seperti prediksi finansial dan diagnosis medis

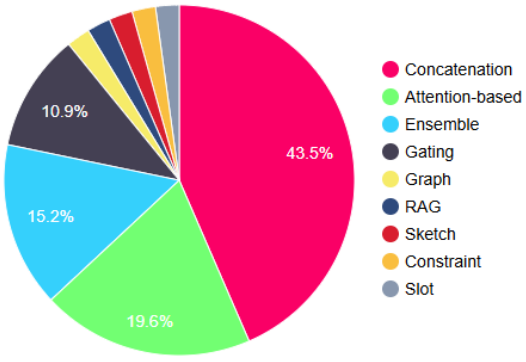
4.1. Level dan Teknik Fusion serta Dampaknya terhadap Performa Model Multimodal

Berdasarkan analisis terhadap 26 studi terpilih, *Intermediate (feature-level) fusion* muncul sebagai pendekatan yang paling dominan dalam arsitektur multimodal karena kemampuannya menangkap korelasi non-linear antar-modalitas di ruang representasi laten. Selain itu, arsitektur *Hybrid Fusion* juga menunjukkan kinerja kompetitif, khususnya pada prediksi kredit dan data medis kompleks [1], [23], [27], [35]. *Early Fusion* lebih efektif pada konteks dengan keterikatan temporal yang kuat atau ketika data tabular diserialisasi menjadi teks untuk LLM [22], [29], [36], sementara *Late Fusion* unggul ketika satu modalitas dominan atau pada dataset tidak seimbang, serta lebih efisien secara memori [30], [35], [40], [42].



Gambar 1. Level fusi yang sering digunakan dalam artikel temuan

Pada tingkat teknik, *concatenation* tetap menjadi metode paling umum karena kesederhanaannya dan performa yang stabil, terutama saat dikombinasikan dengan CNN atau MLP [1], [21], [35], [38], [39]. Namun, peningkatan kinerja terbesar umumnya diperoleh melalui mekanisme *attention* dan *gating*, seperti *Cross-Attention*, *Overall Attention*, *Slot Attention*, dan *Multimodal Adaptation Gate* [23], [25], [26], [28]. Peningkatan performa yang dilaporkan bervariasi, mulai dari kenaikan moderat hingga lonjakan signifikan, termasuk peningkatan AUROC sebesar +0.231 pada prediksi risiko rumah sakit dengan *Knowledge-Augmented Reasoning* [30]. Selain itu, *disentangled representation* terbukti krusial dalam menjaga kinerja model [28].



Gambar 2. Teknik fusi yang sering digunakan dalam artikel temuan

Secara umum, pendekatan multimodal secara konsisten mengungguli unimodal di berbagai domain. Modalitas teks sering menjadi kontributor prediktif utama, khususnya pada tugas keuangan dan medis [1], [28], meskipun kontribusi modalitas lain tetap signifikan, seperti data gambar pada crowdfunding dan time series numerik pada prediksi fisiologis [24], [39]. Namun, penambahan modalitas yang tidak terkelola dengan baik berpotensi menurunkan kinerja akibat noise tambahan [38].

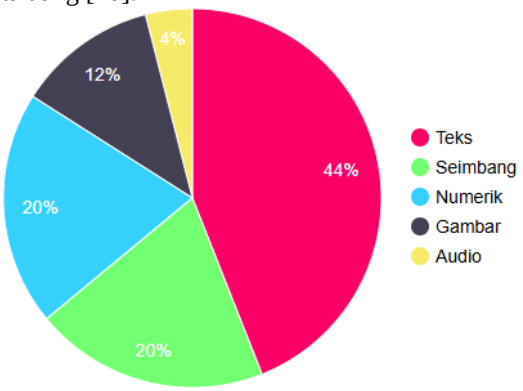
4.2. Kontribusi Relatif Modalitas Teks dan Kondisi Khusus

Secara umum, pendekatan multimodal secara konsisten mengungguli model unimodal, namun kontribusi modalitas teks sangat bergantung pada domain dan karakteristik data. Pada tugas yang menuntut pemahaman semantik dan penalaran kualitatif, teks sering menjadi kontributor dominan, seperti pada prediksi risiko kredit dan analisis narasi kecelakaan [1], [22]. Dominasi ini juga terlihat pada sebagian domain EHR, di mana penghapusan catatan klinis menyebabkan penurunan performa signifikan, termasuk penurunan presisi hingga 17%, menunjukkan bahwa data laboratorium saja tidak cukup merepresentasikan kondisi klinis pasien [28]. Pola serupa ditemukan pada dataset ulasan produk dan film, di mana teks secara konsisten mengungguli modalitas lain.

Sebaliknya, pada tugas yang sangat bergantung pada pemantauan sinyal fisiologis atau kondisi fisik yang terukur secara objektif, modalitas non-teks seperti *time series*, gambar, atau data Sebaliknya, pada tugas berbasis sinyal fisik objektif, seperti prediksi mortalitas ICU, ilmu material, dan pertanian, modalitas non-teks justru memegang peranan utama, sementara teks bersifat pelengkap [24], [38], [44]. Dalam prediksi mortalitas ICU, kontribusi time series tanda vital (≈43%) melampaui teks (≈33%) [24]. Dominasi modalitas juga bersifat temporal; pada fase awal rawat inap, data tabular lebih informatif, sementara teks menjadi lebih dominan pada fase akhir ketika ringkasan klinis tersedia [45].

Peran teks menjadi sangat krusial pada kondisi krisis atau ketidakpastian eksternal. Dalam pasar yang stabil, te ks berita dapat bertindak sebagai noise, tetapi saat terjadi gejolak pasar atau krisis global, teks berubah menjadi sinyal prediktif yang kuat karena mampu menangkap sentimen dan persepsi risiko yang tidak tercermin dalam data numerik historis [29]. Hal ini menunjukkan fungsi teks sebagai sensor kontekstual terhadap kejadian langka.

Pada kondisi data tidak seimbang atau langka, integrasi teks—termasuk melalui LLM secara few-shot—terbukti meningkatkan performa pada kelas minoritas, baik pada data kecelakaan maupun prediksi medis dengan kejadian rendah [21]. Namun, kontribusi ini sangat bergantung pada relevansi dan keselarasan temporal; teks yang tidak relevan justru meningkatkan error, sebagaimana terlihat pada prediksi nilai tukar mata uang [20].

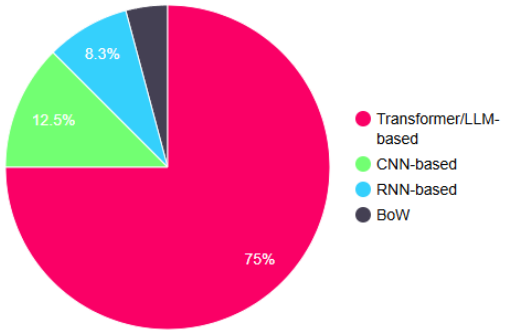


Gambar 3. Modalitas dengan pengaruh kuat dalam artikel temuan

Tantangan utama muncul ketika teks mengandung noise atau strategi fusi yang tidak tepat. *Early fusion* pada teks yang noisy dapat menurunkan kinerja dibandingkan model unimodal, sementara integrasi modalitas tambahan yang tidak informatif dapat mendilusi sinyal utama [24], [40]. Oleh karena itu, desain arsitektur menjadi faktor penentu.

4.3. Kompatibilitas Arsitektur Encoder dan Strategi Fusi Multimodal

Secara umum, literatur menunjukkan dominasi arsitektur berbasis Transformer (BERT, BioBERT, FinBERT, hingga LLM) sebagai encoder utama untuk modalitas teks, yang umumnya dipadukan dengan arsitektur spesifik domain untuk modalitas lain. Pola yang konsisten memperlihatkan kombinasi encoder teks dengan CNN atau ResNet untuk data visual, serta RNN (GRU, LSTM, mTAND) untuk data deret waktu atau sinyal fisiologis. Strategi fusi juga mengalami pergeseran dari *early concatenation* menuju pendekatan *intermediate* atau *hybrid* yang memanfaatkan mekanisme attention dan alignment fitur, karena terbukti lebih efektif dalam menangani perbedaan representasi semantik antar-modalitas dibandingkan penggunaan fitur beku [24].



Gambar 4. Arsitektur yang sering digunakan dalam artikel temuan

Dalam domain medis dan Electronic Health Records (EHR), strategi fusi dipengaruhi oleh

kompleksitas, heterogenitas, dan sparsity data klinis. Pendekatan hibrida, seperti integrasi inisialisasi LLM ke dalam Hypergraph Neural Network, efektif dalam menangani struktur graf medis yang jarang melalui mekanisme *message passing* [27]. Untuk data time series yang tidak teratur, kombinasi mTAND dan Slot Attention memungkinkan penangkapan pola temporal tingkat tinggi [26]. Pendekatan homogenisasi data, seperti konversi data tabular menjadi *pseudo-notes* agar dapat diproses oleh encoder yang sama (MedBERT), juga terbukti mampu menghindari inkompatibilitas arsitektur dan masalah truncation [21]. Selain itu, pemisahan fitur *modality-specific* dan *modality-shared* menggunakan Disentangled Transformer efektif dalam mereduksi noise pada data laboratorium [28].

Di sektor keuangan, pemilihan arsitektur sangat dipengaruhi oleh kebutuhan penyelarasan temporal dan panjang dokumen. *Timeline-based fusion* menjadi krusial untuk menyelaraskan berita dengan data harga saham secara ketat dan mencegah noise temporal [29]. Menariknya, pada prediksi peringkat kredit, CNN mengungguli BERT karena keterbatasan panjang token Transformer dalam memproses transkrip keuangan yang sangat panjang, sehingga fusi hibrida berbasis CNN menjadi solusi yang lebih efektif [1]. Sebaliknya, pada domain crowdfunding, *late fusion* sederhana antara CNN (teks) dan VGG-16 (gambar) sudah cukup untuk mengungguli model unimodal karena sifat modalitas yang saling melengkapi [39].

Analisis komparatif menunjukkan bahwa kompleksitas arsitektur harus sebanding dengan kompleksitas interaksi antar-modalitas. Strategi fusi berbasis attention, hypergraph, atau *joint learning* secara konsisten mengungguli *simple fusion* ketika terdapat ketergantungan semantik atau temporal yang kuat. Namun, *late fusion* tetap kompetitif ketika satu modalitas dominan atau ketika modalitas bersifat relatif independen. Hal ini menegaskan bahwa peningkatan kompleksitas arsitektur tidak selalu menjamin peningkatan performa.

5. KESIMPULAN DAN SARAN

Berdasarkan analisis terhadap studi terpilih, disimpulkan bahwa kinerja model multimodal dengan teks sebagai modalitas utama sangat ditentukan oleh interaksi antara strategi fusi, kontribusi dinamis antar-modalitas, dan kompatibilitas arsitektur *encoder*. Pendekatan *intermediate fusion* mendominasi sebagai metode yang paling efektif, diikuti oleh *hybrid fusion* untuk data kompleks, di mana mekanisme berbasis *attention* memberikan peningkatan performa yang lebih signifikan dibandingkan teknik *concatenation* sederhana. Kontribusi modalitas teks terbukti tidak statis melainkan bergantung pada domain, di mana teks memegang peran dominan pada tugas berbasis semantik serta berfungsi sebagai sensor kontekstual krusial dalam kondisi anomali atau data tidak seimbang. Meskipun Transformer menjadi standar dominan untuk memproses teks, efektivitasnya dibatasi oleh panjang sekuens, sehingga arsitektur alternatif seperti CNN atau pendekatan hierarkis tetap diperlukan untuk menangani dokumen yang sangat panjang guna menghindari risiko *truncation*

DAFTAR PUSTAKA

[1] M. Tavakoli, R. Chandra, F. Tian, and C. Bravo, "Multi-modal deep learning for credit rating prediction using text and numerical data streams," *Appl. Soft Comput.*, vol. 171, Mar. 2025, doi: 10.1016/j.asoc.2025.112771.

[2] S. Cui, J. Wang, Y. Zhong, H. Liu, T. Wang, and F. Ma, "Automated Fusion of Multimodal Electronic Health Records for Better Medical Predictions," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.11252

[3] Z. Yu, S. Yang, Z. Li, L. Li, H. Luo, and F. Yang, "LogMS: a multi-stage log anomaly detection method based on multi-source information fusion and probability label estimation," *Front. Phys.*, vol. 12, 2024, doi: 10.3389/fphy.2024.1401857.

[4] R. Koval, N. Andrews, and X. Yan, "Financial Forecasting from Textual and Tabular Time Series," 2024.

[5] L. Bi, Y. Zhang, L. Wang, Y. Niu, and H. Zhao, "Two Challenges, One Solution: Robust Multimodal Learning through Dynamic Modality Recognition and Enhancement," 2025.

[6] S. Mahmoud Sajjadi Mohammadabadi, B. Cem Kara, C. Eyupoglu, C. Uzay, M. Serkan Tosun, and O. Karaku, "A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications," 2025, doi: 10.3390/electronics.

[7] V. Borisov, T. Leemann, K. Sebler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.

[8] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal Learning with Transformers: A Survey," May 2023, [Online]. Available: http://arxiv.org/abs/2206.06488

[9] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," Apr. 2020, doi: 10.1109/JSTSP.2020.2987728.

[10] S. Li and H. Tang, "Multimodal Alignment and Fusion: A Survey," Oct. 2025, [Online]. Available: http://arxiv.org/abs/2411.17040

[11] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2307.14126

[12] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "TabLLM:

Few-shot Classification of Tabular Data with Large Language Models," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2210.10723

[13] Y. Li *et al.*, "BEHRT: Transformer for Electronic Health Records," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-62922-y.

[14] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," Dec. 01, 2022, *Nature Research*. doi: 10.1038/s41746-022-00712-8.

[15] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications," 2024, *Tech Science Press*. doi: 10.32604/cmc.2024.053204.

[16] S. T. Erukude, S. Reddy Veluru, and V. Chaitanya Marella, "Multimodal Deep Learning: A Survey of Models, Fusion Strategies, Applications, and Research Challenges," 2025.

[17] L. Chen and X. Fan, "TIC-FusionNet: A multimodal deep learning framework with temporal decomposition and attention-based fusion for time series forecasting," *PLoS One*, vol. 20, no. 10 October, Oct. 2025, doi: 10.1371/journal.pone.0333379.

[18] H. Bhin and J. Choi, "Multimodal Personality Recognition Using Self-Attention-Based Fusion of Audio, Visual, and Text Features," *Electronics (Switzerland)*, vol. 14, no. 14, Jul. 2025, doi: 10.3390/electronics14142837.

[19] D. Kim, "Cross attention for Text and Image Multimodal data fusion Stanford CS224N Custom Project."

[20] S. Mou, Q. Xue, J. Chen, T. Takiguchi, and Y. Ariki, "MM-iTransformer: A Multimodal Approach to Economic Time Series Forecasting with Textual Data," *Applied Sciences (Switzerland)*, vol. 15, no. 3, Feb. 2025, doi: 10.3390/app15031241.

[21] Y. Wang, C. Yin, and P. Zhang, "Multimodal risk prediction with physiological signals, medical images and clinical notes," *Heliyon*, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26772.

[22] H. Cui, X. Fang, R. Xu, X. Kan, J. C. Ho, and C. Yang, "Multimodal Fusion of EHR in Structures and Semantics: Integrating Clinical Records and Notes with Hypergraph and LLM," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2403.08818

[23] S. A. Lee *et al.*, "Multimodal Clinical Pseudo-notes for Emergency Department Prediction Tasks using Multiple Embedding Model for EHR (MEME)," Apr. 2024, doi: 10.1038/s41746-025-01777-x.

[24] T. Minh *et al.*, "MEDFuse: Multimodal EHR Data Fusion with Masked Lab-Test Modeling and Large Language Models," 2024. doi: XXXXXXX.XXXXXXX.

[25] S. Jaradat, M. Elhenawy, R. Nayak, A. Paz, H. I. Ashqar, and S. Glaser, "Multimodal Data Fusion for Tabular and Textual Data: Zero-Shot, Few-Shot, and Fine-Tuning of Generative Pre-Trained Transformer Models," *AI (Switzerland)*, vol. 6, no. 4, Apr. 2025, doi: 10.3390/ai6040072.

[26] M. Ma, M. Wang, B. Gao, Y. Li, J. Huang, and H. Chen, "Research on Multimodal Fusion of Temporal Electronic Medical Records," *Bioengineering*, vol. 11, no. 1, Jan. 2024, doi: 10.3390/bioengineering11010094.

[27] T. Bonnier, "Revisiting Multimodal Transformers for Tabular Data with Text Fields," 2024. [Online]. Available: https://github.com/thomas-bonnier/

[28] F. Wang, F. Wu, Y. Tang, and L. Yu, "CTPD: Cross-Modal Temporal Pattern Discovery for Enhanced Multimodal Electronic Health Records Analysis," 2025.

[29] F. Karadaş, B. Eravcı, and A. M. Özbayoğlu, "Multimodal Stock Price Prediction," 2025.

[30] R. Datta *et al.*, "Improving Hospital Risk Prediction with Knowledge-Augmented Multimodal EHR Modeling," Aug. 2025, [Online]. Available: http://arxiv.org/abs/2508.01970

[31] J. Ye, J. Hai, J. Song, and Z. Wang, "Multimodal Data Hybrid Fusion and Natural Language Processing for Clinical Prediction Models."

[32] Z. Gao *et al.*, "Improving the Prognostic Evaluation Precision of Hospital Outcomes for Heart Failure Using Admission Notes and Clinical Tabular Data: Multimodal Deep Learning Model," *J. Med. Internet Res.*, vol. 26, 2024, doi: 10.2196/54363.

[33] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12911-020-01297-6.

[34] S. Kumar, S. Sharma, and K. T. Megra, "Transformer enabled multi-modal medical diagnosis for tuberculosis classification," *J. Big Data*, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40537-024-01054-w.

[35] C. A. Kotula *et al.*, "Comparison of Multimodal Deep Learning Approaches for Predicting Clinical Deterioration in Ward Patients: Observational Cohort Study," *J. Med. Internet Res.*, vol. 27, 2025, doi: 10.2196/75340.

[36] L. Yan, X. Wu, and Y. Wang, "Student engagement assessment using multimodal deep learning," *PLoS One*, vol. 20, no. 6 June, Jun. 2025, doi: 10.1371/journal.pone.0325377.

[37] Y. Cui, S. Liang, and Y. Y. Zhang, "Multimodal representation learning for tourism recommendation with two-tower architecture," *PLoS One*, vol. 19, no. 2 February, Feb. 2024, doi: 10.1371/journal.pone.0299370.

[38] V. Costa, J. M. Oliveira, and P. Ramos, "Deep Learning-Driven Integration of Multimodal Data for Material Property Predictions," *Computation*, vol. 13, no. 12, p. 282, Dec. 2025, doi: 10.3390/computation13120282.

[39] C. Cheng, F. Tan, X. Hou, and Z. Wei, "Success Prediction on Crowdfunding with Multimodal Deep Learning," 2019. [Online]. Available: www.kickstarter.com/projects/573995669/rebuild-a-better-

[40] J. Moon, F. Ke, Z. Sokolikj, and S. Chakraborty, "Applying multimodal data fusion to track autistic adolescents' representational flexibility development during virtual reality-based training," *Computers and Education: X Reality*, vol. 4, Jan. 2024, doi: 10.1016/j.cexr.2024.100063.

[41] J. Zhao *et al.*, "Multimodal Feature Fusion Method for Unbalanced Sample Data in Social Network Public Opinion," *Sensors*, vol. 22, no. 15, Aug. 2022, doi: 10.3390/s22155528.

[42] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective Techniques for Multimodal Data Fusion: A Comparative Analysis," *Sensors*, vol. 23, no. 5, Mar. 2023, doi: 10.3390/s23052381.

[43] S. Singh, E. Saber, P. P. Markopoulos, and J. Heard, "Regulating Modality Utilization within Multimodal Fusion Networks," *Sensors*, vol. 24, no. 18, Sep. 2024, doi: 10.3390/s24186054.

[44] J. Yao, Y. Wu, J. Liu, and H. Wang, "Multimodal deep learning-based drought monitoring research for winter wheat during critical growth stages," *PLoS One*, vol. 19, no. 5 May, May 2024, doi: 10.1371/journal.pone.0300746.

[45] C. S.-T. Huang, C. B. L. Ng, and M. Rei, "Predictive Multimodal Modeling of Diagnoses and Treatments in EHR," Aug. 2025, [Online]. Available: http://arxiv.org/abs/2508.11092

[46] K.-H. Lu *et al.*, "DeSTA2: Developing Instruction-Following Speech Language Model Without Speech Instruction-Tuning Data," Jan. 2025, doi: 10.1109/ICASSP49660.2025.10889444