Implementasi Hierarchical Reasoning Model (HRM) Pada Kasus Klasifikasi

Arwansyah*, Suryani, Hasyrif SY, Nurdiansah, Ahyuna
Universitas Dipa Makassar; Jl Perintis Kemerdekaan Km. 9 Makassar, Telp. (0411) 587194
e-mail: *[1]arwansyah@undipa.ac.id, [2]suryani187@undipa.ac.id, [3]hasyrif@gmail.com, [4]nurdiansah@undipa.ac.id, [5]ahyuna@undipa.ac.id

Abstrak

Penelitian ini mengeksplorasi penerapan Hierarchical Reasoning Model (HRM) tersimulasi sebagai arsitektur pembelajaran mendalam untuk tugas klasifikasi pada berbagai karakteristik data. HRM dirancang untuk meniru proses kognitif dengan memisahkan "perencanaan" tingkat tinggi dari "eksekusi" tingkat rendah dalam siklus penalaran berulang. Model ini diimplementasikan menggunakan TensorFlow/Keras dan diuji pada tiga dataset benchmark: Breast Cancer Wisconsin (BCD), Diabetes Pima Indian, dan Iris. Hasil eksperimen menunjukkan perbedaan kinerja yang signifikan berdasarkan dimensi fitur dan kompleksitas data. HRM menunjukkan efektivitas superior pada dataset BCD (30 fitur) dengan mencapai akurasi pengujian sebesar 96%, membuktikan bahwa mekanisme penyempurnaan state laten sangat menguntungkan pada data berdimensi tinggi. Sebaliknya, pada dataset Diabetes (8 fitur), model mengalami overfitting ekstrem dengan akurasi 71%, sementara pada dataset Iris (4 fitur), model mengalami mode collapse dengan akurasi hanya 33,3%. Temuan ini menyimpulkan bahwa meskipun arsitektur penalaran hierarkis unggul dalam mengekstraksi informasi dari data yang kompleks dan kaya fitur, model ini cenderung tidak stabil dan over-engineered untuk dataset berdimensi rendah. Penelitian ini memberikan wawasan penting mengenai batasan struktural HRM dan perlunya adaptasi arsitektur berdasarkan karakteristik input data untuk mencapai generalisasi yang optimal.

Kata kunci: *Klasifikasi, Hierarchical Reasoning Model.*

Abstract

This study explores the application of a simulated Hierarchical Reasoning Model (HRM) as a deep learning architecture for classification tasks on various data characteristics. The HRM is designed to mimic cognitive processes by separating high-level "planning" from low-level "execution" in an iterative reasoning cycle. The model is implemented using TensorFlow/Keras and tested on three benchmark datasets: Breast Cancer Wisconsin (BCD), Diabetes Pima Indian, and Iris. Experimental results show significant performance differences based on feature dimensionality and data complexity. The HRM demonstrates superior effectiveness on the BCD dataset (30 features) by achieving a test accuracy of 96%, proving that the latent state refinement mechanism is particularly advantageous on high-dimensional data. In contrast, on the Diabetes dataset (8 features), the model suffers from extreme overfitting with an accuracy of 71%, while on the Iris dataset (4 features), the model experiences mode collapse with an accuracy of only 33.3%. These findings suggest that while hierarchical reasoning architectures excel at extracting information from complex, feature-rich data, they tend to be unstable and over-engineered for low-dimensional datasets. This research provides important insights into the structural limitations of HRM and the need for architecture adaptation based on input data characteristics to achieve optimal generalization.

Keywords*: Classification, Hierarchical Reasoning Model.*

1. PENDAHULUAN

Konsep penalaran berulang telah menjadi area fokus utama dalam penelitian deep learning. Penelitian awal yang mendasari ide ini adalah Jaringan Saraf Berulang (RNN), yang menunjukkan kapabilitas memproses data sekuensial [1]. Namun, untuk tugas-tugas yang memerlukan penalaran kompleks, model harus meniru struktur kognitif yang lebih maju. Hierarchical Reasoning Model (HRM) muncul sebagai arsitektur yang menjanjikan dalam domain ini. Secara konseptual, HRM adalah mekanisme komputasi yang secara eksplisit memisahkan perencanaan tingkat tinggi yang abstrak dari

eksekusi tingkat rendah yang detail. Ide inti HRM dipopulerkan dalam konteks Visual Question Answering (VQA).

Model yang menggunakan penalaran berulang dan perhatian adaptif untuk memecah tugas VQA menjadi langkah-langkah logis, secara efektif mengintegrasikan penalaran hierarkis diperkenalkan oleh [2], [3]. Penelitian lain memperluas konsep ini dengan menggunakan mekanisme fiksasi dan penyempurnaan (fixation and refinement) untuk meningkatkan akurasi penalaran [4]. Pemanfaatan mekanisme iteratif untuk menyempurnakan representasi data adalah elemen kunci dari HRM. Penelitian lainnya mengeksplorasi penggunaan mekanisme recurrent refinement untuk meningkatkan kualitas hasil dalam tugas sintesis gambar, yang mendasari ide penyempurnaan state laten seperti yang dilakukan dalam implementasi ini [5]. Selain itu, konsep Iterative Refinement Networks telah dipelajari untuk meningkatkan prediksi dengan menjalankan proses inference secara berulang [6]. Dalam konteks aplikasi medis dan klasifikasi, deep learning telah menjadi standar emas. Penelitian lainnya meninjau secara komprehensif efektivitas berbagai arsitektur DL untuk diagnostik kanker payudara, menegaskan akurasi tinggi yang dapat dicapai, namun menyoroti tantangan dalam interpretasi dan penalaran mendalam [7], [8]. Data diagnostik kanker payudara dari Wisconsin (Wisconsin Breast Cancer Diagnostic dataset) yang digunakan dalam studi ini merupakan benchmark standar dan telah digunakan secara luas oleh para peneliti untuk menguji algoritma machine learning dan deep learning [9], [10].

Salah satu arsitektur yang menjanjikan adalah Hierarchical Reasoning Model (HRM), Studi ini memanfaatkan kemampuan penalaran mendalam HRM untuk menyediakan model yang mampu menyaring fitur diagnostik dengan lebih efektif melalui siklus penyempurnaan internal. Pendekatan ini selaras dengan tren yang lebih luas dalam deep learning menuju model yang sadar-alasan (reasoning-aware) [11], seperti yang terlihat pada penggunaan model berulang dan arsitektur modular dalam tugas penalaran kompleks [12], [13]. Implementasi ini juga mendukung hipotesis bahwa struktur berulang dapat meningkatkan efisiensi parameter dan kedalaman komputasi model [14], [15]. Ide utama dari penelitian ini didukung oleh premis bahwa tugas klasifikasi yang kompleks memerlukan lebih dari sekadar pemetaan input-output tunggal serta keinginan dalam mengimplementasikan sebuah model yang dikembangkan oleh Guang et al [16] dari sebuah makalah penelitian yang berjudul Hierarchical Reasoning Model diimana untuk memaksimalkan informasi dari data, model harus mampu "merenung" atau "memikirkan" representasi internalnya.

HRM menawarkan mekanisme untuk mencapai penalaran yang dalam (deep reasoning) melalui komputasi yang efisien. Daripada membangun jaringan yang sangat dalam, HRM mencapai kedalaman komputasi dengan menjalankan lapisan yang sama dalam loop waktu (time loop) yang terstruktur secara hierarkis. Modul tingkat tinggi membuat keputusan makro (perencanaan), sementara modul tingkat rendah membuat keputusan mikro (eksekusi/penyempurnaan) berdasarkan rencana tersebut. Siklus perbaikan berulang ini (sebagaimana terlihat dalam kode dengan 5 siklus tingkat tinggi, masing-masing dengan 3 langkah tingkat rendah) memungkinkan model untuk memadatkan informasi penting dan menghilangkan kebisingan, menghasilkan state laten yang lebih diskriminatif untuk klasifikasi akhir. Meskipun model pembelajaran mendalam konvensional telah mencapai kesuksesan besar dalam tugas klasifikasi, terdapat beberapa celah penting yang melandasi urgensi penelitian ini diantaranya arsitektur deep learning standar umumnya beroperasi dalam jalur feed-forward satu arah yang kaku, Sebagian besar penelitian klasifikasi menggabungkan ekstraksi fitur dan logika pengambilan keputusan dalam satu tumpukan lapisan yang homogen, Implementasi model penalaran hierarkis sejauh ini masih terbatas pada domain spesifik seperti pemrosesan bahasa alami atau pengenalan gambar tingkat lanjut. Oleh karena itu, penelitian ini bertujuan untuk memvalidasi hipotesis bahwa struktur penalaran hierarkis dan iteratif ini secara inheren lebih efektif untuk tugas klasifikasi yang membutuhkan ekstraksi fitur yang cermat dan akurat, seperti diagnostik medis.

Kontribusi utama dari penelitian ini adalah:

1. Implementasi Konseptual HRM Tersimulasi
   Menyajikan implementasi Hierarchical Reasoning Model (HRM) yang disederhanakan dan end-to-end menggunakan TensorFlow/Keras, yang secara eksplisit memodelkan siklus Perencanaan (High-Level) dan Eksekusi (Low-Level). Implementasi ini berfungsi sebagai proof-of-concept untuk menunjukkan viabilitas arsitektur HRM di luar domain yang biasa.
2. Penerapan dalam Klasifikasi Biomedis
   Menerapkan model HRM ini pada tugas klasifikasi, sebuah aplikasi di mana akurasi dan keandalan keputusan sangat penting. Hal ini menunjukkan potensi HRM sebagai alat diagnostik.
3. Analisis Mekanisme Penalaran Berulang
   Mendemonstrasikan bagaimana penalaran berulang yang diatur secara hierarkis dapat digunakan untuk secara iteratif menyempurnakan latent state data, yang pada akhirnya mengarah pada kinerja klasifikasi yang kuat.

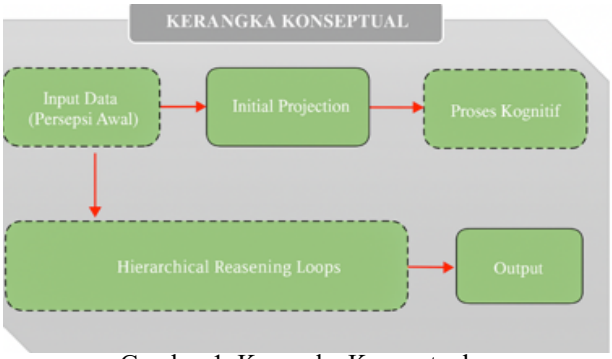
2. METODOLOGI PENELITIAN

Bagian ini menguraikan metodologi yang digunakan dalam penelitian ini untuk mengimplementasikan dan mengevaluasi kinerja Hierarchical Reasoning Model (HRM) tersimulasi pada tugas klasifikasi. Metode yang disajikan meliputi deskripsi dataset benchmark yang digunakan, rincian arsitektur jaringan saraf kustom yang memodelkan penalaran hierarkis, penjelasan cara kerja mekanisme perencanaan (tingkat tinggi) dan eksekusi (tingkat rendah), serta spesifikasi hyperparameter yang dikonfigurasi selama proses pelatihan. Secara keseluruhan, bagian ini memberikan kerangka kerja yang transparan untuk mereplikasi eksperimen dan memvalidasi hasil yang diperoleh.

2.1. Kerangka Konseptual

Kerangka konseptual ini menjelaskan hubungan antara data input, proses penalaran berulang oleh Hierarchical Reasoning Model (HRM), dan hasil klasifikasi. Kerangka ini bertujuan untuk memodelkan proses kognitif dalam HRM, yang membagi tugas klasifikasi biner menjadi siklus Perencanaan (Tingkat Tinggi) dan Eksekusi (Tingkat Rendah) untuk menyempurnakan representasi data laten secara iteratif.

1. Input Data (Persepsi Awal) : Fitur-fitur mentah dari dataset. Data input ini pertama kali diproses oleh Initial Projection Layer
2. Proses Kognitif : Ini adalah inti dari HRM, tempat state laten disempurnakan. Proses ini terjadi dalam siklus hierarkis yang tertutup.
3. Output : Keputusan klasifikasi.



Gambar 1. Kerangka Konseptual

2.2. Dataset

Penelitian ini menggunakan tiga dataset benchmark klasik dalam bidang machine learning untuk menguji generalisasi dan efektivitas Hierarchical Reasoning Model (HRM) dalam tugas klasifikasi.

1. Breast Cancer
   Dataset biner (klasifikasi Benign/Malignant) yang berasal dari hasil digitasi citra aspirasi jarum halus (Fine Needle Aspiration - FNA) massa payudara. Terdiri dari 30 fitur kontinu berdimensi tinggi yang berasal dari sepuluh karakteristik inti sel (misalnya, radius, tekstur, perimeter, area, dan simetri) yang dihitung rata-rata, standard error, dan nilai terburuk.
2. Diabetes
   Digunakan untuk memprediksi apakah seorang pasien menderita diabetes berdasarkan beberapa pengukuran diagnostic. Terdiri dari 8 fitur numerik yang mencakup informasi seperti jumlah kehamilan, konsentrasi glukosa plasma, tekanan darah diastolik, Body Mass Index (BMI), dan usia.
3. Iris
   Dataset multikelas (3 kelas) yang terkenal dalam pengujian algoritma klasifikasi. Terdiri dari 4 fitur numerik yang mengukur dimensi kelopak (sepal) dan mahkota (petal) (panjang dan lebar) dari tiga spesies bunga Iris.

2.3. Pra-Pemrosesan Data

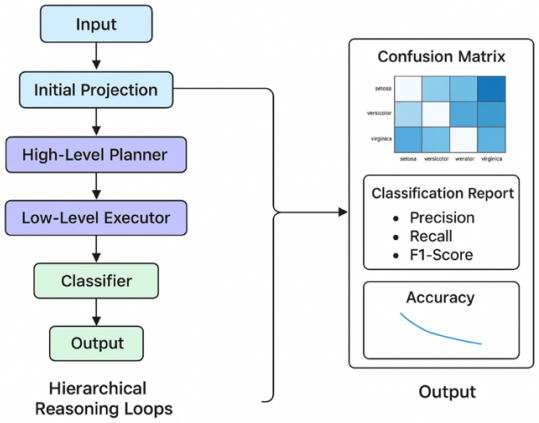
Sebelum dataset digunakan pada proses pelatihan model, terdapat beberapa poin penting yang dilakukan yakni :

1. Pembagian Data
   Data dibagi menjadi set pelatihan dan pengujian dengan rasio 80:20.
2. Standardisasi Fitur

Fitur-fitur kontinu distandardisasi menggunakan StandardScaler dari scikit-learn untuk menghasilkan mean nol dan varian satu.

2.4. Arsitektur Model

Hierarchical Reasoning Model (HRM) merupakan pendekatan pembelajaran bertingkat yang meniru cara manusia melakukan penalaran kompleks melalui dua tingkatan pemrosesan utama, yaitu High-Level Planner dan Low-Level Executor. Model ini dirancang untuk menyelesaikan permasalahan yang memerlukan pemahaman mendalam dan proses berpikir berulang, seperti klasifikasi, perencanaan, atau pengambilan keputusan.



Gambar 2. Arsitektur Model

1. Input dan Initial Projection
   Tahapan pertama menerima data mentah. Fitur tersebut kemudian diproyeksikan ke ruang laten melalui Initial Projection Layer. Lapisan ini mengubah representasi input menjadi vektor berdimensi tinggi agar lebih mudah diolah oleh modul reasoning.
2. High-Level Planner
   Modul ini berfungsi sebagai perencana abstrak yang menyusun strategi atau representasi global terhadap data. Ia menghasilkan rencana awal berdasarkan input yang telah diproyeksikan, sebelum diberikan kepada tingkat yang lebih rendah.
3. Low-Level Executor
   Setiap siklus reasoning melibatkan Low-Level Executor, yaitu modul yang melakukan perhitungan detail dan pembaruan cepat terhadap rencana yang dibuat oleh High-Level Planner. Eksekutor ini beroperasi beberapa kali dalam setiap iterasi reasoning untuk menyempurnakan hasil.
4. Hierarchical Reasoning Loops
   Bagian inti HRM adalah loop hierarkis antara High-Level Planner dan Low-Level Executor. Dalam implementasi Anda, siklus ini dilakukan beberapa kali. Setiap iterasi memperbaiki rencana berdasarkan hasil eksekusi sebelumnya, sehingga menghasilkan representasi yang semakin matang.
5. Classifier dan Output
   Setelah proses reasoning selesai, hasil akhir dikirim ke Classifier Layer dengan aktivasi softmax untuk menghasilkan probabilitas setiap kelas. Output ini menjadi prediksi akhir dari model.

2.5. Arsitektur Model

Beberapa parameter yang digunakan meliputi:

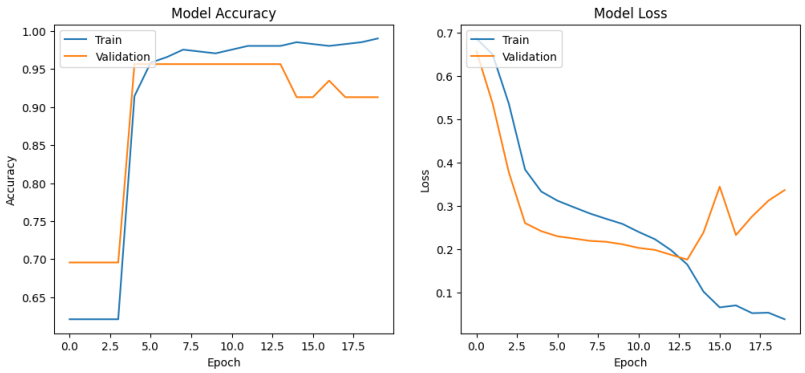
1. High Level Unit : 32
2. Low Level Unit : 3
3. Loss Function : Sparse Categorical Crossentropy
4. Optimizer : ADAM
5. Batch size : 32
6. Learning rate : 0.001
7. Epoch : 50
8. Jumlah neuron : 32

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan dan menganalisis temuan eksperimental yang diperoleh dari implementasi Hierarchical Reasoning Model (HRM) tersimulasi. Hasil evaluasi kinerja model disajikan secara kuantitatif untuk ketiga dataset yang diuji meliputi Kanker Payudara Wisconsin (BCD), Diabetes Pima Indian, dan Iris. Analisis ini mencakup metrik akurasi, loss, classification report, dan confusion matrix untuk setiap kasus. Pembahasan akan berfokus pada interpretasi hasil, membandingkan efektivitas HRM dalam tugas klasifikasi biner (BCD dan Diabetes) dan multikelas (Iris). Selanjutnya, bagian ini membahas implikasi mekanisme penalaran hierarkis dan iteratif terhadap kualitas state laten, kemampuan diskriminatif model, dan generalisasi kinerja HRM di berbagai jenis data.

3.1. Dataset Kanker Payudara Wisconsin (BCD)

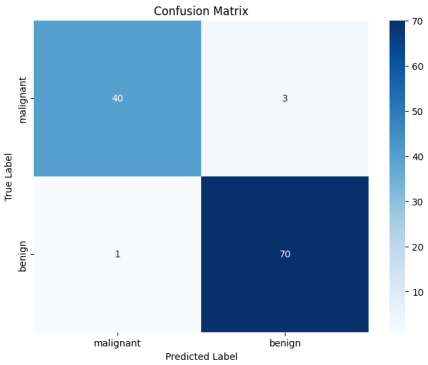
1. Analisis Riwayat Pelatihan (Accuracy and Loss)



Gambar 3. Model Accuracy dan Loss Dataset Kanker Payudara Wisconsin (BCD)

Gambar 3 menunjukkan Akurasi pelatihan dan validasi meningkat tajam pada epoch 5 dan 6. Akurasi pelatihan (Train) mencapai dan mempertahankan tingkat yang sangat tinggi (approx 0.98), sementara Akurasi validasi (Validation) stabil di sekitar 0.91-0.95. Sementara Loss pelatihan (Train) dan validasi (Validation) menunjukkan penurunan yang stabil hingga sekitar epoch 13. Setelah itu, loss pelatihan terus menurun, tetapi loss validasi mulai menunjukkan fluktuasi signifikan (terutama pada epoch 16), yang dapat mengindikasikan awal dari overfitting (model menjadi terlalu spesifik pada data pelatihan).

2. Analisis Confusion Matrix



Gambar 4. Confusion Matrix Dataset Kanker Payudara Wisconsin (BCD)

Gambar 4 menjelaskan bahwa Tiga kasus yang sebenarnya malignant (ganas) diklasifikasikan salah sebagai benign (jinak). Ini adalah jenis kesalahan yang paling kritis dalam diagnostik medis (Type II Error), karena berpotensi menunda pengobatan. Hanya satu kasus yang sebenarnya benign (jinak) diklasifikasikan salah sebagai malignant (ganas). Ini adalah kesalahan yang kurang kritis (Type I Error), meskipun dapat menyebabkan kecemasan pasien dan pengujian yang tidak perlu. Tingkat False Positive (FP) yang rendah (3 kasus) dan False Negative (FN) yang sangat rendah (1 kasus) sangat diinginkan. Kesalahan FN adalah yang paling harus dihindari, dan model hanya memiliki satu kasus FN.

3. Classification Report

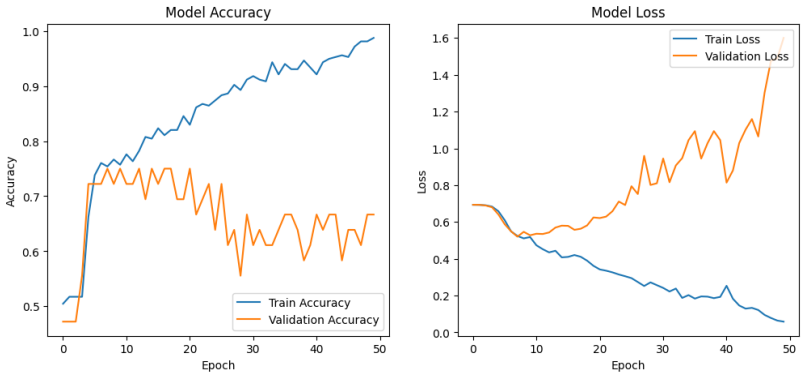
```
--- Classification Report ---
              precision    recall  f1-score   support

   malignant       0.98      0.93      0.95        43
      benign       0.96      0.99      0.97        71

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```

Gambar 5. Classification Report Dataset Kanker Payudara Wisconsin (BCD)

Gambar 5 menjelaskan bahwa Model menunjukkan Precision yang sangat tinggi untuk kelas malignant (0.98), menunjukkan bahwa dari semua kasus yang diprediksi sebagai ganas, 98% di antaranya benar. Model mencapai Recall yang luar biasa untuk kelas benign (0.99), yang berarti hampir semua kasus benign yang sebenarnya diidentifikasi dengan benar. Sementara F1-Score (rata-rata harmonis dari precision dan recall) menunjukkan keseimbangan kinerja yang kuat pada kedua kelas, yaitu 0.95 untuk malignant dan 0.97 untuk benign. Kinerja model yang tinggi (Akurasi 96%, F1-Score Tertimbang 0.96) memberikan bukti kuat akan efektivitas Hierarchical Reasoning Model (HRM) dalam tugas klasifikasi biner diagnostik.

3.2. Dataset Diabetes Pima Indian

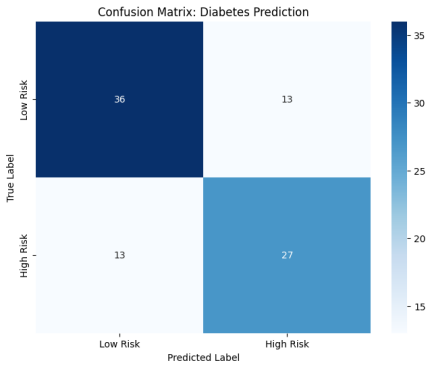
1. Analisis Riwayat Pelatihan (Accuracy and Loss)



Gambar 6. Model Accuracy dan Loss Dataset Diabetes Pima Indian

Gambar 6 menunjukkan Plot riwayat pelatihan menunjukkan perilaku yang sangat berbeda dibandingkan hasil pada dataset kanker. Akurasi pelatihan (Train Accuracy) meningkat dengan cepat dan mencapai hampir 1.00 pada epoch akhir. Namun, Akurasi Validasi (Validation Accuracy) segera mendatar di sekitar 0.65-0.75 sejak epoch 10 dan tidak menunjukkan peningkatan signifikan, bahkan berfluktuasi secara ekstrem. Sementara Loss pelatihan (Train Loss) menurun secara stabil hingga mendekati nol, sedangkan Loss Validasi (Validation Loss) meningkat tajam dan tidak stabil setelah epoch 10, mencapai nilai di atas 1.5.

2. Analisis Confusion Matrix



Gambar 7. Confusion Matrix Dataset Diabetes Pima Indian

Gambar 7 menjelaskan bahwa Kasus sebenarnya risiko rendah salah diklasifikasikan sebagai risiko tinggi. Ini dapat menyebabkan intervensi medis yang tidak perlu. Sementara Kasus sebenarnya risiko tinggi salah diklasifikasikan sebagai risiko rendah. Ini adalah jenis kesalahan yang lebih berisiko dalam konteks medis, karena dapat menunda diagnosis dan pengobatan diabetes yang diperlukan. Terdapat keseimbangan jumlah kesalahan yang signifikan (FP=13 dan FN=13), yang konsisten dengan metrik Precision dan Recall yang seimbang (tetapi moderat) untuk kedua kelas. Dataset Diabetes Pima Indian memiliki jumlah fitur yang jauh lebih sedikit (8 fitur) dan kompleksitas data yang berbeda dari BCD (30 fitur).

Kemungkinan ruang fitur yang terbatas atau noise yang tinggi dalam data Diabetes membuat proses penalaran iteratif yang dilakukan oleh HRM menjadi kontraproduktif, di mana 5 siklus perulangan memungkinkan model terlalu cepat overfit pada training set yang kecil. Tingkat kesalahan yang tinggi dan seimbang pada klasifikasi risiko tinggi dan rendah ($FP=13, FN=13$) menunjukkan bahwa model ini belum dapat diandalkan untuk aplikasi diagnostik diabetes yang sesungguhnya. Dalam konteks ini, False Negatives (kasus High Risk yang terlewat) sangat berbahaya, dan akurasi 71% terlalu rendah untuk aplikasi klinis. Mekanisme HRM yang kompleks (5 siklus perencanaan, 3 langkah eksekusi) mungkin terlalu rumit (over-engineered) untuk dataset dengan dimensi fitur yang rendah (8 fitur). Model klasifikasi yang lebih sederhana, seperti Feedforward Network non-recurrent atau model Machine Learning klasik, mungkin lebih efisien dan stabil pada dataset ini.

3. Classification Report

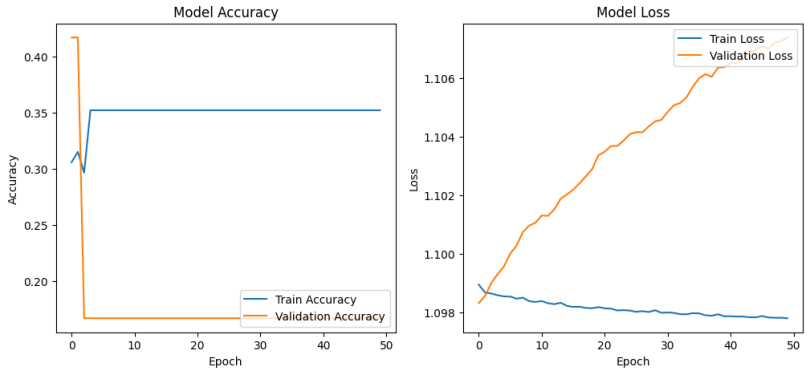
```
Evaluating the model on the test data...
Test Accuracy: 0.7079
Test Loss: 1.1788

--- Classification Report ---
              precision    recall  f1-score   support

    Low Risk       0.73      0.73      0.73        49
   High Risk       0.68      0.68      0.68        40

    accuracy                           0.71        89
   macro avg       0.70      0.70      0.70        89
weighted avg       0.71      0.71      0.71        89
```

Gambar 8. Classification Report Dataset Diabetes Pima Indian

Gambar 8 menunjukkan bahwa Metrik Precision, Recall, dan F1-Score berada pada tingkat yang serupa untuk kedua kelas (sekitar 0.70). Hal ini mengindikasikan bahwa model tidak secara signifikan memihak salah satu kelas. Sementara Kinerja model sedikit lebih baik dalam mengklasifikasikan risiko rendah (Low Risk) (F1-Score: 0.73) dibandingkan risiko tinggi (High Risk) (F1-Score: 0.68).

3.3. Dataset Iris

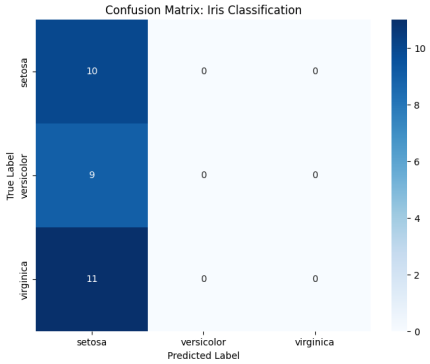
1. Analisis Riwayat Pelatihan (Accuracy and Loss)



Gambar 9. Model Accuracy dan Loss Dataset Iris

Gambar 9 menjelaskan bahwa Train Accuracy stabil di sekitar 0.35 sejak epoch awal. Validation Accuracy segera jatuh ke 0.17 dan tetap di sana. Tidak ada peningkatan yang terlihat, menunjukkan model berhenti belajar secara efektif setelah forward pass pertama. Train Loss sedikit menurun, namun Validation Loss terus meningkat secara linier sepanjang 50 epochs. Ini adalah sinyal terburuk, mengindikasikan bahwa setiap siklus HRM (perulangan) semakin memperburuk kinerja model pada data validasi, menjauhkannya dari solusi yang optimal.

2. Analisis Confusion Matrix



Gambar 10. Confusion Matrix Dataset Iris

Gambar 10 menunjukkan bahwa semua 30 sampel pengujian diprediksi sebagai 'setosa', Hanya 10 kasus Setosa yang benar serta 20 sampel (semua Versicolor dan Virginica) salah diklasifikasikan sebagai Setosa. Selain itu, Model mengalami Mode Collapse, di mana ia gagal untuk membedakan antara kelas dan hanya menghasilkan satu output yang sama (Setosa) untuk semua input.

3. Classification Report

```
Evaluating the model on the test data...
Test Accuracy: 0.3333
Test Loss: 1.1000

--- Classification Report ---
              precision    recall  f1-score   support

      setosa       0.33      1.00      0.50        10
  versicolor       0.00      0.00      0.00         9
   virginica       0.00      0.00      0.00        11

    accuracy                           0.33        30
   macro avg       0.11      0.33      0.17        30
weighted avg       0.11      0.33      0.17        30
```

Gambar 11. Classification Report Dataset Iris

Gambar 11 menjelaskan bahwa Akurasi 33.3% hampir setara dengan peluang acak (1/3) dalam klasifikasi tiga kelas, mengindikasikan model tidak mempelajari pola diskriminatif yang signifikan. Semua kasus Setosa terdeteksi dengan benar (True Positives), tetapi ini juga terjadi karena model memprediksi semua sampel sebagai Setosa. Dari semua yang diprediksi sebagai Setosa (yaitu, 30 sampel), hanya 33% (10 sampel) yang benar-benar Setosa. Model tidak pernah memprediksi kedua kelas ini dengan benar, karena model tidak pernah memprediksinya sama sekali.

3.4. Perbandingan Performa Model

Tabel 1. Perbandingan Performa Model

Dataset	HRM	SVM	Random Forest	MLP
Iris	≈33.00%	≈96.00% [17]	≈96.10% [18]	≈96.00% [17]
Breast Cancer	≈96.00%	≈97.80% [19]	≈98.50% [20]	≈95.00% [17]
Diabetes	≈71.00%	≈86.00% [21]	≈77.00% [22]	≈77.54% [21]

4. KESIMPULAN

Penelitian ini mengimplementasikan dan mengevaluasi Hierarchical Reasoning Model (HRM) tersimulasi sebagai arsitektur deep learning untuk tugas klasifikasi, dengan fokus pada pemanfaatan mekanisme penalaran berulang hierarkis. Evaluasi dilakukan pada tiga dataset dengan karakteristik dimensi dan kompleksitas yang beragam: Kanker Payudara Wisconsin (BCD), Diabetes Pima Indian, dan Iris. HRM menunjukkan kinerja yang superior dan stabil pada Dataset Kanker Payudara (BCD) yang memiliki dimensi fitur tinggi (30 fitur), mencapai Akurasi Pengujian sekitar 96 dan F1-Score yang seimbang. Ini memvalidasi hipotesis bahwa mekanisme perencanaan dan eksekusi iteratif HRM efektif dalam menyaring representasi data laten dan mengekstraksi fitur diskriminatif yang kompleks, yang sangat berharga untuk aplikasi

diagnostik. Berdasarkan perbandingan performa pada hasil experiment, dapat disimpulkan bahwa Hierarchical Reasoning Model (HRM) lebih tepat dan efektif untuk diimplementasikan pada dataset yang memiliki karakteristik berupa dimensi fitur tinggi (High-Dimensional Data), hubungan fitur yang kompleks, tugas klasifikasi biner yang membutuhkan presisi tinggi, serta dataset dengan volume sampel yang cukup.

5. SARAN

Untuk meningkatkan generalisasi dan stabilitas HRM, penelitian di masa depan disarankan untuk:

1. Mengembangkan mekanisme HRM adaptif di mana jumlah siklus penalaran (N_HL) atau langkah eksekusi (N_LL) ditentukan secara dinamis atau dihentikan berdasarkan konvergensi state laten.
2. Menguji arsitektur HRM pada tugas klasifikasi multikelas berdimensi tinggi yang lebih kompleks, seperti klasifikasi citra medis, untuk menguji batas kemampuan penalaran hierarkisnya.

DAFTAR PUSTAKA

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[2] J. Johnson, A. Khosla, F. Lu, S. Yeung, T. Ren, L. Fei-Fei, and W. Li, “Inferring and executing programs for visual reasoning,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2989–2998.

[3] A. Goyal, Z. Zhang, T. Khot, A. Vahdat, D. Li, H. Lee, and Y. Bengio, “The essential role of language in fine-grained visual classification,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[4] A. Khosla, J. Johnson, F. Lu, T. Ren, W. Li, and L. Fei-Fei, “Hierarchical reasoning with fixation and refinement,” arXiv preprint arXiv:1804.09849, 2018.

[5] F. Li, T. Yao, and T. Mei, “Recurrent refinement network for image completion,” in Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, 2018.

[6] T. D. Kulkarni, R. Grosse, and J. B. Tenenbaum, “Deep convolutional inverse graphics network,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 28, 2015.

[7] M. Ahmad, H. Al-Ghamdi, and B. Al-Thubiti, “Deep learning models for breast cancer detection and classification: A review,” J. Healthcare Eng., vol. 2020, Art. no. 8833541, 2020.

[8] R. Das, H. Maity, and S. Hore, “A comparative study on breast cancer prediction using machine learning and deep learning techniques,” in Proc. Int. Conf. Comput. Sci., Eng. Appl. (ICCSEA), 2020, pp. 1–6.

[9] W. H. Wolberg and O. L. Mangasarian, “Multisurface method of pattern separation for medical diagnosis applied to breast cytology,” Proc. Natl. Acad. Sci. USA, vol. 87, no. 23, pp. 9193–9196, 1990.

[10] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” in Proc. IS&T/SPIE Int. Symp. Electron. Imaging: Sci. Technol., 1993.

[11] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019.

[12] A. Graves, G. Wayne, and I. Danihelka, “Neural Turing machines,” arXiv preprint arXiv:1410.5401, 2014.

[13] S. Sukhbaatar, J. Weston, M. Ranzato, and R. Collobert, “End-to-end memory networks,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 28, 2015.

[14] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 13, 2000.

[15] D. Ha and J. Schmidhuber, “Recurrent world models facilitate policy evolution,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 29, 2016.

[16] G. Wang et al., “Hierarchical reasoning model,” arXiv:2506.21734v3 [cs.AI], 2025.

[17] E. R. Susanto and D. Misdiantoro, “Optimasi akurasi prediksi penyakit kanker payudara menggunakan metode random forest,” J. Pendidikan dan Teknologi Indonesia, vol. 5, no. 5, pp. 1407–1416, 2025.

[18] I. Irawansyah, K. Adiwijaya, and W. Astuti, “Comparative analysis of support vector machine (SVM) and random forest (RF) classification for cancer detection using microarray,” in Proc. 9th Int. Conf. Inf. Commun. Technol. (ICoICT), 2021, pp. 1–6, doi: 10.1109/ICoICT52021.2021.9527458.

[19] D. Alfiani, M. P. Putri, and W. Widayanti, “Perbandingan algoritma support vector machine (SVM) dan logistic regression dalam klasifikasi kanker payudara,” J. Kecerdasan Buatan dan Teknologi

Informasi, 2025.
[20] M. F. Aryansyah, “Perbandingan algoritma random forest, decision tree, dan support vector machine (SVM) dalam klasifikasi tingkat keganasan kanker payudara,” Repository UBSI, 2025.
[21] Sutrisno and J. Jupron, “Analisa klasifikasi penyakit diabetes dengan algoritma neural network,” eJournal: Komunitas Dosen Indonesia, vol. 6, no. 3, pp. 304–308, 2024.
[22] G. Abdurrahman, H. Oktavianto, and M. Sintawati, “Optimasi algoritma XGBoost classifier menggunakan hyperparameter grid search dan random search pada klasifikasi penyakit diabetes,” INFORMAL Informatics J., vol. 7, no. 3, 2022.