A Hierarchical Multi-Label Classification Approach for the Automated Interpretation of Spinal MRI Series

David Cahyadi[1], Edwin Pramana[2], Rudi Limantara[3], I Gusti Lanang Ngurah Agung Artha Wiguna[4], Maria Florencia Deslivia[5] and Ivan Alexander Liando[4]
[1]Department of Information Technology, Faculty of Science and Technology, Institut Sains dan Teknologi Terpadu Surabaya, Indonesia
[2]Department of Informatics, Faculty of Science and Technology, Institut Sains dan Teknologi Terpadu Surabaya, Indonesia
[3]School of Information Technology, Informatics Department Universitas Ciputra Surabaya, Indonesia
[4]Rumah Sakit Umum Pusat Sanglah Denpasar (Sanglah Central Public Hospital Denpasar), Indonesia
[5]Department of Orthopedic Surgery, St Carolus Hospital, Jakarta, Indonesia

Corresponding author: David Cahyadi (e-mail: liu.david.chd@gmail.com).

ABSTRACT Manually selecting MRI slices is a significant bottleneck in clinical workflows. This issue is worsened by inconsistent naming conventions and variable acquisition protocols across institutions and radiologists, often leading to redundant efforts and potential oversights during medical image data preprocessing. This study introduces a fully automated, four-level hierarchical classification system specifically designed to intelligently filter and select clinically relevant spinal MRI slices directly from raw DICOM series. Our primary objective is to streamline the initial stages of radiological assessment, ensuring that only pertinent images are presented for subsequent analysis and review. We thoroughly evaluated the performance of modern, efficient deep learning architectures, including EfficientViT, MobileNetV4, and RepViT, benchmarking them against a robust ResNet-18 baseline. The proposed pipeline systematically refines its analysis through a structured hierarchy: it first broadly identifies the anatomical region, then precisely classifies the spine location and specific view (axial, sagittal, or coronal). Subsequently, it categorizes the imaging contrast, and finally, confirms the presence of the spinal cord. Our comprehensive experimental results reveal that the EfficientViT-based model achieved the highest end-to-end F1-score of 0.8357, demonstrating robust accuracy across all classification levels. Furthermore, its average inference speed of 9.17 ms per image highlights its computational efficiency. This automated pipeline offers an effective and computationally efficient solution for speeding up initial medical image preprocessing, ensuring subsequent analytical tasks are performed on accurately selected, clinically relevant data.

KEYWORDS Deep Learning, Hierarchical Classification, Medical Image Analysis, Transfer Learning

I. INTRODUCTION

MRI is excellent for soft tissue visualization, but its adoption in automated systems faces significant technical hurdles. A primary challenge is the inconsistent naming of MRI series across different institutions. This issue complicates automated medical data processing, especially when different series augmentation techniques lead to misaligned data labels. Such inconsistencies slow down data analysis, a problem often seen in large datasets. Current approaches haven't fully addressed the need for a system that can operate without human intervention from the initial image selection stage. This study tackles this gap by developing an automated, end-to-end expert system. It specifically focuses on the initial, yet critical, task of selecting the correct MRI series for analysis. We introduce an innovative expert system that uses a hierarchical classification structure to progressively filter and identify the desired spinal MRI slices.

The key contributions of this work include:

- A novel hierarchical inference structure that breaks down the complex task of MRI series identification into four manageable stages, effectively addressing data requirements at each level.
- A comprehensive performance evaluation of several state-of-the-art efficient vision models (EfficientViT[1], MobileNet V4[2], RepVit[3]) against a ResNet-18[4] baseline to identify the optimal architecture for this task.

- A demonstration of a computationally efficient system that can significantly accelerate the pre-analytical phase of spinal MRI review, making it feasible for real-time clinical applications.

By automating this foundational step, our system has the potential to reduce data processing time, support the radiologist's workflow, and increase the accessibility and standardization of spinal cord image analysis.

II. THEORY

A. SPINE ANATOMY

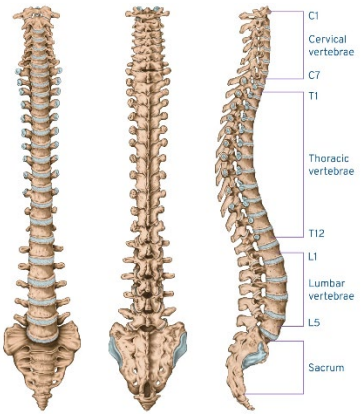

Figure 1. **Spine Anatomy.**

The human spine, also known as the vertebral column, is a vital and complex structure providing support, flexibility, and protection for the spinal cord, composed of a series of stacked bones called vertebrae. As illustrated, it is divided into distinct regions: the cervical vertebrae (C1-C7) in the neck support the head and allow for its extensive movement; the thoracic vertebrae (T1-T12) in the upper and mid-back articulate with the ribs to form the protective rib cage, limiting mobility; the lumbar vertebrae (L1-L5), located in the lower back, are the largest and strongest, bearing most of the body's weight and allowing for trunk flexibility. Below these, the sacrum comprises five fused vertebrae, connecting the spine to the pelvis for stability, and finally, the coccyx, or tailbone, consists of several small, fused vertebrae at the very bottom, serving as an attachment point for pelvic muscles.

B. MRI SEQUENCE

Magnetic Resonance Imaging (MRI) is a powerful diagnostic tool that utilizes strong magnetic fields and radio waves to generate detailed images of organs and soft tissues. Different MRI sequences are employed to highlight specific tissue characteristics, providing varied contrasts crucial for diagnosis.

- **T1-weighted (T1w) Images:** T1-weighted images are produced by optimizing sequence parameters to emphasize the longitudinal (T1) relaxation time of tissues. On T1w images, fluids (like cerebrospinal fluid, CSF) appear dark, while fat and areas with protein or hemorrhage appear bright. T1w sequences are excellent for anatomical detail, visualizing brain anatomy, spinal cord structures, and detecting certain types of lesions, especially post-contrast administration where enhanced tissues appear bright.
- **T2-weighted (T2w) Images:** T2-weighted images are optimized to emphasize the transverse (T2) relaxation time. In contrast to T1w, fluids (like CSF, edema, or inflammation) appear bright on T2w images, while most normal solid tissues appear intermediate to dark. T2w sequences are highly sensitive to pathology, making them invaluable for detecting edema, inflammation, tumors, and demyelination, as these conditions often lead to increased water content.
- **Dixon Sequence:** The Dixon sequence is a specialized MRI technique designed to separate fat and water signals within the same acquisition. This is particularly useful in areas where fat can obscure pathology or interfere with signal analysis. By acquiring data at different echo times, the Dixon method can generate multiple image sets: water-only, fat-only, in-phase, and out-of-phase images.
- **Myelography:** MR Myelography is a non-invasive MRI technique used to visualize the spinal cord and the surrounding cerebrospinal fluid (CSF) within the subarachnoid space. Unlike conventional X-ray myelography which requires injecting contrast directly into the CSF, MR myelography typically relies on heavily T2-weighted sequences. These sequences suppress the signal from surrounding tissues, making the bright CSF appear prominent, thus outlining the spinal cord and nerve roots. It is particularly useful for detecting spinal canal stenosis, nerve root compression, disc herniations, spinal tumors, and other conditions that impinge upon the CSF space.

C. MRI VIEW POINT

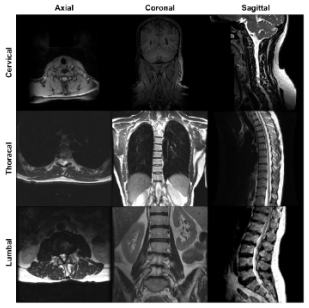

Figure 2. **Spine Location and View.**

In medical imaging, especially when looking at the spine with MRI, we use three main ways to view the anatomy, helping us understand its complex 3D structure. The **axial view** takes horizontal slices, like looking down at a cross-section of the spine at a specific level, which is great for seeing how the spinal cord and nerves fit inside and checking for any pressure in Figure 2 the first column. The **coronal view** cuts

the spine from front to back, showing us how the vertebrae line up from a front or back perspective, useful for spotting sideways curves like scoliosis in the Figure 2 the second column. Finally, the **sagittal view** slices the spine from side to side, giving us a clear profile view that's excellent for seeing the spine's natural curves, disc health, and any problems like slipped discs or narrowing of the spinal canal along its length in Figure 2 the third column.

D. TRANSFER LEARNING

Transfer learning is a machine learning technique where a model trained on a large dataset for one task (the source task) is repurposed or fine-tuned for a different but related task (the target task). In medical imaging, this is exceptionally beneficial because acquiring and annotating vast, diverse datasets for specific medical conditions is challenging and resource-intensive. By starting with a model pre-trained on a massive dataset like ImageNet (or in our case, RadImageNet), the model has already learned a rich hierarchy of features, from basic edges and textures to more complex patterns. These learned features, representing general image characteristics, can then be transferred and adapted to new medical imaging tasks, such as classifying spinal cord pathologies, with significantly less training data and computational resources than training a model from scratch. This approach accelerates convergence, improves generalization, and often leads to superior performance

E. ADAMW OPTIMIZER

AdamW[5] is an optimization algorithm commonly used for training deep neural networks. It is a variant of the popular Adam (Adaptive Moment Estimation) optimizer, but with a crucial modification regarding weight decay. In standard Adam, weight decay (a regularization technique to prevent overfitting) is coupled with the adaptive learning rates, which can sometimes lead to suboptimal regularization. AdamW decouples weight decay from the adaptive learning rates, treating it as a distinct regularization term. This separation ensures that weight decay is applied more effectively and consistently, often leading to better generalization performance and improved training stability compared to the original Adam optimizer, especially in models with a large number of parameters like those used in deep learning for vision tasks.

III. PREVIOUS RESEARCH

The core concept of this work, Hierarchical Multi-Label Classification (HMC), has been extensively studied within the machine learning literature, particularly for domains with inherent hierarchical relationships[6], [7], [8], [9]. In the context of visual recognition, researchers have successfully employed hierarchical deep convolutional neural networks (HD-CNNs) for large-scale image classification[10], showing that breaking down a problem into a hierarchy of parent and child classes can improve performance. The fundamental principle is that by decomposing a complex problem into smaller, more manageable sub-tasks, the data requirements at each decision point are effectively reduced while maintaining high classification accuracy.

This approach is particularly valuable in the medical imaging domain, where large-scale, expertly labeled datasets are often challenging and resource-intensive to obtain. Kowsari et al[8]. have demonstrated the use of a Hierarchical Medical Image Classification (HMIC) framework, underscoring the potential of this technique in the medical field. These methods provide a robust foundation for building specialized clinical tools.

Furthermore, our research group has previously leveraged the same private dataset utilized in this study for other critical spinal analysis tasks. Our prior investigations successfully applied deep learning models for the purpose of automated lesion segmentation, as well as for a basic classification of axial slice scores[11]. This foundational work not only validated the richness and utility of our dataset for deep learning applications but also highlighted the need for a more efficient and structured method for initial series identification, a bottleneck that our current hierarchical framework is designed to address.

In parallel to the application-specific research, there has been a growing recognition of the importance of the pre-training datasets used to initialize deep learning models. While ImageNet, a dataset of natural images, has been the de facto standard for pre-training, its domain divergence from medical imaging can limit performance. The introduction of RadImageNet, a large-scale dataset of radiological images, has offered a compelling alternative.

Multiple studies have demonstrated that models pre-trained on RadImageNet[12], [13], [14], [15] can achieve better performance on a variety of medical imaging tasks. This advantage is particularly pronounced when the target application involves radiological images, such as MRI, and when the available training data for the specific task is limited. The features learned from RadImageNet are inherently more relevant to the anatomical structures and tissue contrasts found in medical scans, leading to more effective transfer learning. For spinal MRI analysis, leveraging a RadImageNet-pretrained model can be particularly beneficial, as it provides the network with a foundational understanding of radiological imaging that is more closely aligned with the target domain. This can lead to faster convergence during training and improved overall accuracy of the final model, whether for classification or segmentation.

IV. DATASET AND METHOD

A. DATASET ACQUISITION

The development and validation of our framework were conducted using a diverse collection of six public and two private datasets to ensure robust model training and

evaluation. The primary resource for training our foundation classification model was RadImageNet[16], which contains over 1.4 million images from multiple anatomical regions and imaging modalities. For spinal cord-specific imaging, we used the Spine Generic dataset (T1w and T2w MRI data from 267 subjects) and the SPIDER dataset.

To address specific research needs, two private datasets were incorporated. The Siloam Dataset contained T1w and T2w spinal MRI scans from 258 patients. The Sanglah Dataset contained spinal MRI scans from 93 patients, with a focus on cervical spine cases.

The distribution of these datasets across the various classification tasks is detailed in Table 1.

TABLE I
DATA DISTRIBUTION ON EVERY TASKS

Task	Dataset Name	Train	Val	Test
Anatomy Classification	RadImageNet[16]	1,354,913	0	0
	M4Raw[17] (brain)	0	0	1,000
	Siloam (spine)	0	0	1,000
	Sanglah (spine)	0	0	1,000
	CHAOS[18] (abdomen)	0	0	718
	Spine Generic[19] (spine)	0	0	1,000
	SPIDER (spine)	0	0	1,000
	FastMRI[20] Prostate (abdomen)	0	0	1,000
	FastMRI[20] Knee (knee)	0	0	1,000
	FastMRI[20] Brain (brain)	0	0	1,000
Spine Location and View	RadImageNet[16]	71,515	0	0
	Spine Generic[19]	101,782	6,798	26,507
	SPIDER[21]	5,667	857	1,066
Contrast Classification	Sanglah	2,782	226	695
	Siloam	7,386	493	1,413
	Spine Generic[19]	42,161	2,580	11,006
Spinal Cord	Sanglah & Siloam	7,352	530	1,467
	Spine Generic[19]	12,906	768	3,068

B. DATA PREPARATION

All datasets underwent rigorous preparation to ensure consistency. Images were converted to PNG format for unified data handling, and min-max normalization was applied to standardize intensity variations across scanners and protocols. To enhance dataset diversity and model robustness, we employed data augmentation techniques including rotation, flipping, blurring, and brightness/contrast adjustments.

We also implemented a Human-in-the-Loop (HITL) approach with active learning to address the critical lack of anatomical labels in the acquired RadImageNet data. This strategy focused on maximizing labeling speed without compromising quality, especially for the "Spine Location and View" task. The workflow began with a small, manually annotated dataset for initial model training. The model then inferred labels on new batches, which human experts reviewed and corrected. This efficient process, where humans only corrected model errors, allowed us to progressively improve the model's performance with each verified batch. This synergistic loop enabled us to efficiently scale labeling across the entire RadImageNet dataset, generating the high-fidelity labels required for our study.

C. HIERARCHICAL CLASSIFICATION FRAMEWORK

We developed a hierarchical classification system that decomposes the complex challenge of analyzing spinal MRI series into smaller, more manageable sub-tasks. This approach progressively refines the classification from broad anatomical regions to specific spinal cord slices, which addresses the common challenge of dataset imbalance and reduces the volume of training data needed at each stage. The classification process flows through four distinct levels, as illustrated in Figure 1 and detailed in Table 2.

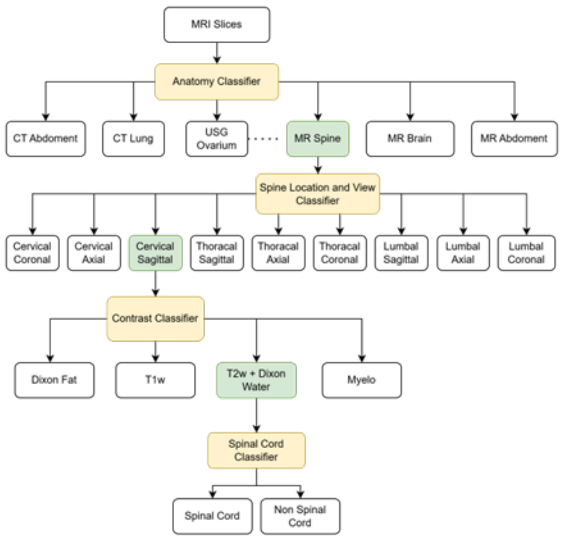

Figure 3. **Hierarchical Classification Structure.**

This hierarchical classification system in Figure 3, processes medical imaging data through four distinct levels, each progressively refining the identification of relevant MRI spine images. The hierarchical classification flow works step-by-step:

1. **Level 1:** Anatomy Classification, acts as the initial and broadest filter, sifting through a diverse collection of MRI slices from various body parts and other imaging modalities to isolate only MRI scans of the spine. The Anatomy Classifier is trained to differentiate between major anatomical regions, and scanning tools to be used (CT, MRI and USG) with total of 24 different group labels, ultimately passing only confirmed MR Spine images to the next stage.

2. **Level 2**: Spine Location and View Classification, the system categorizes the identified spine scans by specific vertebral column region and imaging plane. This level takes MRI spine slices confirmed by Level 1 and performs a detailed analysis to determine if the slice is from the Cervical, Thoracic, or Lumbar region, and whether the view is Sagittal, Axial, or Coronal. The output is a precisely labeled slice, such as Cervical Sagittal or Thoracal Axial, ensuring the system understands the exact spine part and its orientation.
3. **Level 3:** Spine Sequence Classification, identifies the specific MRI physics sequence used to generate the image. The Contrast Classifier analyzes image properties like pixel intensity, texture, and tissue appearance to determine the underlying MRI sequence, such as T1w, T2w + Dixon Water, Dixon Fat, or Myelo. This level outputs an image slice identified by its anatomy, view, and technical sequence (e.g., a T2w + Dixon Water slice of the sagittal cervical spine), selecting the most appropriate sequences for the final task.
4. **Level 4:** Spinal Cord Classification, determines if the individual, fully-characterized image slice actually contains the spinal cord. This level receives a pre-filtered and fully identified slice from Level 3 and performs a binary classification. Leveraging all contextual information from previous levels, the Spinal Cord Classifier makes a binary decision, resulting in the final classification of the slice as either containing the Spinal Cord or Non Spinal Cord.

TABLE 2
TASK DESCRIPTIONS

Level	Task	Description	MRI Input
1	Anatomy Classification	The primary classification of anatomical structures to identify spine-related images.	MRI Slices
2	Spine Location and View Classification	Classification of anatomical structures to identify specific spine views (e.g., Cervical-Sagittal).	MRI Spine Slice
3	Spine Sequence Classification	MRI sequence identification (e.g., T2w) for proper image interpretation.	MRI Sagittal Cervical Spine Slice
4	Spinal Cord Classification	Detailed classification to identify slices containing the spinal cord, utilizing contextual information from previous levels.	MRI Sequence Identified Slice

This hierarchical structure enhances processing efficiency by filtering out irrelevant images at each level, ensuring that only a subset of images reaches the more specialized classification stages. The specific design of this hierarchy is a direct response to challenges in data availability and distribution, particularly the need to mitigate severe label imbalance. In a raw collection of medical scans, images containing the actual spinal cord are exceedingly rare compared to the vast number of other images (e.g., brain, abdomen, or even spine slices adjacent to the cord).

By structuring the problem this way, we strategically group the data at each step. The output of one classifier becomes a large, curated, and more balanced training set for the next. For instance, the Anatomy Classifier first creates a substantial pool of "MR Spine" images. The Spine Location and View Classifier then draw from this pool to create robust subsets, like "Cervical Sagittal." This ensures that even the final, most specialized classifier has sufficient and relevant data to train on, dramatically reducing the likelihood of label imbalance and enabling each model in the chain to be trained effectively.

D. MODEL ARCHITECTURES AND TRAINING

At each node in the hierarchy, we evaluated four distinct model architectures:

- **EfficientViT**[1]**:** EfficientViT is a family of high-speed vision transformers designed to overcome the computational costs and memory inefficiencies of existing transformer models, making them suitable for real-time applications. It features a "sandwich layout" block, which places a single memory-bound Multi-Head Self-Attention (MHSA) layer between more memory-efficient Feed Forward Network (FFN) layers to reduce memory access time. To combat computational redundancy in attention maps, EfficientViT introduces a "cascaded group attention" (CGA) module. This module feeds attention heads with different splits of the full feature, enhancing diversity and saving computation by reducing redundant attention calculations.
- **MobileNetV4**[2]**:** MobileNetV4 (MNv4) represents the latest generation of MobileNets, featuring architectures that are universally efficient for mobile devices. A core innovation is the Universal Inverted Bottleneck (UIB) search block, a flexible structure that unifies existing micro-architectures like Inverted Bottleneck (IB), ConvNext, and Feed Forward Network (FFN), and introduces a new variant called Extra Depthwise (ExtraDW). MNv4 also introduces Mobile Multi-Query Attention (MQA), an attention block specifically optimized for mobile accelerators, achieving over a 39% inference speedup compared to standard Multi-Head Attention. The models are developed using a refined two-phase Neural Architecture Search (NAS) approach that separates coarse and fine-grained searches, which improves search efficiency and allows for the creation of significantly larger models.
- **RepViT**[3]**:** RepViT is a new family of pure lightweight Convolutional Neural Networks (CNNs) that revisits the efficient design of lightweight CNNs from a Vision Transformer[22] (ViT) perspective, emphasizing their potential for mobile devices. RepViT incrementally enhances the mobile-friendliness of MobileNetV3[23]

by integrating efficient architectural designs typically found in lightweight ViTs.

- **ResNet**[4]**:** The core idea is to reformulate layers to learn residual functions with reference to the layer inputs, rather than unreferenced functions. This approach makes residual networks easier to optimize and allows them to gain accuracy from considerably increased depth. For the ResNet variant used on this research was the architecture with 18-layer "plain" network. The model demonstrates that residual learning effectively addresses the degradation problem (where deeper networks have higher training error) and enables accuracy gains from increased depth.

Our anatomy model was built using the complete RadImageNet dataset as its foundation. For other classification tasks, we employed a transfer learning strategy, fine-tuning the weights from this pre-trained anatomy model. The anatomy models underwent 50 epochs of training, utilizing the AdamW optimizer, a learning rate of 1e−3 with cosine decay scheduling, and Cross-Entropy Loss. To enhance computational efficiency, we implemented mixed-precision (16-bit) training. Subsequent models were trained for 25 epochs, as fine-tuning offers greater efficiency compared to training from scratch.

E. EVALUATION METRICS

Model performance for all classification tasks was assessed using a triad of standard metrics: Precision, Recall, and F1-Score.

Precision, as shown in (1) measures the proportion of true positive predictions among all positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}} \quad (1)$$

Recall (or Sensitivity), shown in (2) measures the proportion of actual positive instances that were correctly identified:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}} \quad (2)$$

The **F1-Score** is the harmonic mean of Precision and Recall as shown in (3), providing a balanced measure of performance, which is especially useful for imbalanced classes:

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

All metrics were specifically calculated for their respective target classes. For anatomy, metrics were compared against the MR-spine target. Location and view metrics were focused on the cervical-sagittal label. For sequence, the target was T2w images, and for spinal cord, metrics were aligned with the spinal cord image itself.

V. RESULTS

The hierarchical classification system was evaluated based on individual task performance and end-to-end system accuracy and efficiency.

A. INDIVIDUAL CLASSIFICATION TASK PERFORMANCE

The performance of the four architectures was evaluated across the four distinct classification tasks. In the Anatomy Classification task (identifying spine images), MobileNet V4 achieved the highest F1-score of 0.8175. For Spine Location and View Classification (identifying sagittal cervical views), EfficientViT was the top performer with an F1-score of 0.9857 and the highest precision of 0.9862. In the Cervical Sequence Classification task (identifying T2-weighted images), all models performed exceptionally well, with ResNet achieving the highest F1-score of 0.9905 and EfficientViT achieving a perfect recall of 1.0000. Finally, in the Spinal Cord Classification task, ResNet showed the best overall performance with an F1-score of 0.9226. The complete results are shown in Table 3.

TABLE 3
CLASSIFICATION RESULT PER TASK

Task	Metric	ResNet	Efficient ViT	MN V4	RepViT
Anatomy Classification	Precision	0.9837	0.9930	**0.9950**	0.9858
	Recall	0.6778	0.6713	**0.6938**	0.6768
	F1 Score	0.7912	0.8010	**0.8175**	0.8026
Spine Location	Precision	0.9425	**0.9862**	0.9574	0.9843
	Recall	**0.9927**	0.9853	0.9839	0.9778
	F1 Score	0.9669	**0.9857**	0.9705	0.9810
Cervical Sequence	Precision	**0.9822**	0.9767	0.9765	0.9758
	Recall	0.9993	**1.0000**	0.9962	0.9993
	F1 Score	**0.9905**	0.9879	0.9860	0.9871
Spinal Cord	Precision	**0.9131**	0.9129	0.8997	0.8534
	Recall	0.9324	0.9300	**0.9420**	0.9701
	F1 Score	**0.9226**	0.9213	0.9203	0.9080

Note: Bold values indicate the best performance for each metric within each task.

B. END-TO-END CLASSIFICATION PERFORMANCE

The most critical evaluation is the end-to-end performance of the entire pipeline in identifying the final target: a T2-weighted, sagittal, cervical slice containing the spinal cord. For series-level decisions (anatomy, location, sequence), a majority voting mechanism was used. The end-to-end results revealed that **EfficientViT** delivered the strongest overall performance, achieving the highest F1-score (0.8357) and recall (0.7905). ResNet achieved the highest precision (0.9060) but with a lower recall. The performance of MobileNet V4 and RepVit dropped significantly in the end-to-end evaluation, suggesting that errors compounded through the hierarchical stages for these architectures. The end-to-end performance metrics are presented in Table 4.

TABLE 4
END-TO-END CLASSIFICATION RESULT

Metric	ResNet	Efficient ViT	Mobile Net V4	RepViT
Precision	**0.9060**	0.8864	0.7339	0.4939
Recall	0.7162	**0.7905**	0.5405	0.5473
F1 Score	0.8000	**0.8357**	0.6226	0.5192

Note: Bold values indicate the best performance for each metric.

C. MODEL PERFORMANCE ANALYSIS

Computational efficiency is crucial for this applicability. Our analysis, conducted with training on an RTX 3060 12GB and inference on a CPU with a Ryzen 5600x, showed that EfficientViT was the fastest architecture, with an inference time of 9.1744 ms per image. This represents a 21.81% speed improvement over the ResNet baseline (11.1748 ms). MobileNet V4 (15.5738 ms) and RepVit (17.5459 ms) showed minimal to no speed improvement over ResNet. These findings establish EfficientViT as the optimal choice, offering the best balance of classification accuracy and processing speed.

TABLE 5
MODEL PERFORMANCE COMPARISON

Model	Training Time	Avg Inference (image / second)	Avg Inference Time in ms
Resnet	20h 10m 22s	89.4874	11.1748
EfficientVit MSRA	**18h 1m 32s**	**108.9994**	**9.1744**
MobileNet V4	36h 3s	64.2105	15.5738
RepVit	33h 45m 23s	56.9935	17.5459

Note: Bold values indicate the best performance for each metric.

Regarding training time, this metric reflects the duration required for each model to learn from the dataset and optimize its internal parameters on the specified hardware. A shorter training time indicates higher computational efficiency during the model development phase. As shown in Table 5, EfficientVit MSRA had the shortest training time at 18 hours, 1 minute, and 32 seconds. This is significantly faster than MobileNet V4 (36h 3s) and RepVit (33h 45m 23s), and also outperforms the ResNet baseline (20h 10m 22s). EfficientVit's shorter training time suggests a more efficient architecture for learning, enabling faster iteration during model development and potentially reducing computational costs.

VI. DISCUSSION

This study aimed to evaluate and identify the optimal deep learning architecture for a hierarchical classification system designed to pinpoint specific T2-weighted, sagittal, cervical MRI slices that include the spinal cord. The evaluation was based on a comprehensive analysis of individual task performance, end-to-end system accuracy, and computational efficiency. The results indicate that **EfficientViT** provides the most compelling balance of high accuracy and operational speed, establishing it as the superior model for this specific application.

A key finding from the individual task evaluation (Table 3) is that no single architecture universally outperformed others across all four distinct classification tasks. This suggests that the architectural biases of different models make them better suited for recognizing different types of features.

MobileNet V4's success in **Anatomy Classification** (F1-score of 0.8175) may be attributed to its design, which is optimized for identifying general object categories efficiently. This task, which involves distinguishing spine images from a broader set of anatomical scans, aligns well with its strengths. However, a notable observation across all models in this task was the significant gap between high precision and lower recall. This pattern suggests that while the models are highly reliable when they identify a spine image (few false positives), they tend to miss a considerable number of true spine images (more false negatives).

EfficientViT excelled in the more granular **Spine Location and View Classification** task (F1-score of 0.9857). This task requires the model to identify specific sagittal cervical views, a process that likely benefits from the attention mechanisms inherent in Vision Transformer architectures, which can better focus on fine-grained spatial relationships and details.

The near-perfect scores achieved by all models in the **Cervical Sequence Classification** task, where **ResNet** narrowly led with a 0.9905 F1-score, indicate that differentiating MRI sequences (like T2-weighted images) is a relatively straightforward task for modern deep learning models, likely due to distinct and consistent texture and contrast features.

For the final **Spinal Cord Classification**, **ResNet** demonstrated the best performance (F1-score of 0.9226). This task involves detecting the presence of a specific, often subtle, anatomical structure within an already correctly-oriented slice. The deep residual connections in ResNet may be particularly effective at capturing the complex and nuanced features required for this level of detailed identification.

The most critical evaluation was the end-to-end performance (Table 4), as it reflects the real-world utility of the entire system. Here, the hierarchical nature of the pipeline brought the issue of compounding errors to the forefront. An error in an early classification stage (e.g., Anatomy) cannot be corrected and guarantees failure for that series.

The superior end-to-end performance of **EfficientViT** (F1-score of 0.8357) is a testament to its robustness across the entire pipeline. Its high recall (0.7905) is particularly important, as it indicates the system is successful at identifying a high percentage of the target series. In contrast, while **ResNet** achieved the highest precision (0.9060), its significantly lower recall (0.7162) suggests a more conservative model that, while rarely wrong, misses many correct instances. For a clinical or research screening tool, higher recall is often preferable to ensure that relevant images are not overlooked.

The most striking result is the sharp decline in performance for **MobileNet V4** and **RepViT** in the end-to-end evaluation. Despite respectable individual task scores, their final F1-scores of 0.6226 and 0.5192, respectively, demonstrate that

minor weaknesses in one or more stages created a cascade of failures, rendering them unsuitable for the complete task.

Beyond predictive accuracy, the practical application of such a system hinges on its computational efficiency. The analysis in Table 5 solidifies **EfficientViT**'s position as the optimal choice. It not only delivered the best end-to-end classification results but was also the most efficient architecture.

With the fastest inference speed (9.1744 ms per image) and the shortest training time (approximately 18 hours), EfficientViT is exceptionally well-suited for deployment. Its inference speed represents a 17.90% improvement over the widely-used ResNet baseline, enabling faster processing of large MRI series. Furthermore, its rapid training time facilitates quicker development cycles, model retraining, and iterative improvements. This computational efficiency, combined with its high accuracy, makes it a powerful and practical tool. In contrast, architectures like MobileNet V4 and RepViT were not only less accurate but also demanded significantly more computational resources for training without providing a corresponding benefit in inference speed.

A detailed error analysis reveals that the primary sources of performance degradation originated from two specific stages: the initial Anatomy Classification and the final Spinal Cord Classification, both of which had overall F1-scores below 0.95.

TABLE 5
CONFUSION MATRIX ANATOMY TASK EFFICIENT VIT MODEL

Model	MR brain	MR knee	MR mriabd	MR spine
Other CT	1	1	20	90
MR_af	2	19	8	229
MR_brain	**1986**	3	12	282
MR_hip	0	23	176	143
MR_knee	1	**923**	0	184
MR_mriabd	7	27	**1480**	302
MR_shoulder	0	3	5	18
MR_spine	3	2	14	**2685**
Other US	0	0	2	67

This analysis shows that the lower recall for the anatomy task was primarily driven by a significant number of MR_spine images being misclassified as other anatomical regions. The most frequent confusions were with MR_mniabd (302 instances) and MR_brain (282 instances), and the visual basis for these errors is illustrated in Figure 4. The confusion with abdominal scans arises because they often include the lumbar spine, presenting vertebral features (Figure 4a) that mimic dedicated spine images. Similarly, the anatomical boundary between the brain and spine is often indistinct; images capturing the craniocervical junction (Figure 4b) or coronal views showing both regions (Figure 4c) contain overlapping features. This inherent ambiguity underscores the difficulty of classifying scans based on gross anatomical features alone, leading the model to produce these specific false negatives.

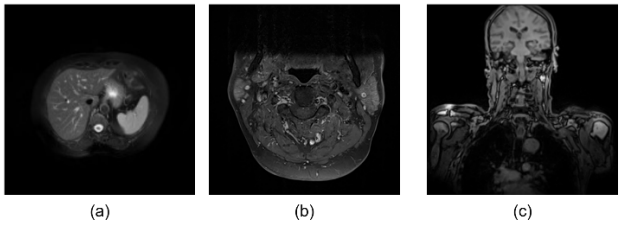

Figure 4. Representative examples of visually ambiguous MRI slices that challenge the anatomy classification task

For the final Spinal Cord Classification task, the slightly lower precision indicates the model generated a number of false positives. This behavior stems not from the model misidentifying other anatomical features, but from its difficulty with transitional slices, as illustrated in the figure 5. The example shows how a slice immediately adjacent to a clear view of the spinal cord may contain only a small, partial segment of the spinal canal. The presence of some cord tissue in these boundary slices creates an inherent ambiguity, making it challenging for the model to apply a consistent classification threshold. Consequently, the model tends to be overly inclusive by flagging these ambiguous cases. While high recall is critical, these false positives introduce noise that would require a subsequent manual review to filter out, slightly reducing the overall efficiency of the automated pipeline.

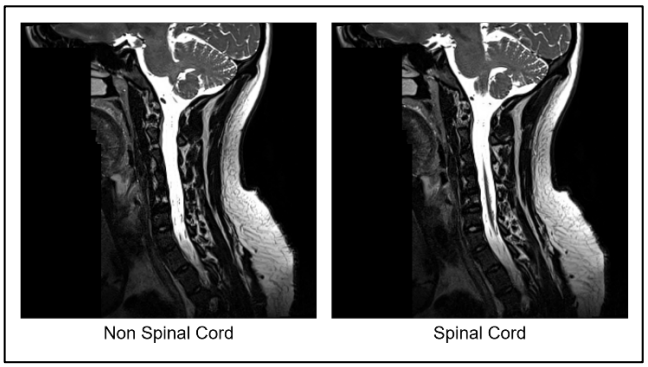

Figure 5. Example of a Transitional MRI Slice Causing Spinal Cord Classification Ambiguity

While the results presented in this study are promising, several limitations should be acknowledged. First, the models were trained and evaluated on a dataset from a single source. Consequently, the system's generalizability has not yet been established. Future work must involve validation on diverse, multi-institutional datasets to test the model's robustness against variations in image acquisition protocols, scanner manufacturers, and different patient populations, as these factors can significantly impact performance in real-world clinical settings.

Second, the current hierarchical system relies on a simple majority voting mechanism for series-level decisions. While effective, this approach could be enhanced. Future research

should explore more sophisticated aggregation techniques to potentially improve end-to-end accuracy. For instance, implementing weighted voting based on model confidence scores or employing other ensemble methods could provide a more nuanced decision-making process. Addressing these limitations will be a crucial step toward developing a clinically viable and reliable automated tool.

VII. CONCLUSION

This research directly confronted the challenge of inconsistent and time-consuming manual selection of spinal MRI series, a common impediment in radiological workflows. We successfully demonstrated that a hierarchical deep learning framework can effectively automate the identification of clinically relevant T2-weighted, sagittal, cervical MRI slices containing the spinal cord. By systematically breaking down the classification problem into four logical stages, our system efficiently filters large, raw DICOM series to isolate the target images with high precision.

Our comprehensive evaluation of modern deep learning architectures revealed that EfficientViT is the superior model for this application, establishing the best synergy between predictive accuracy and operational performance. The model's robustness was evidenced by its top-tier end-to-end F1-score of 0.8357 and recall of 0.7905. Critically, its computational efficiency—marked by the fastest training time and an inference speed 21% faster than the ResNet baseline—positions it as a practical tool for real-world integration.

The broader impact of this automated system is its potential to streamline the pre-analytical phase of medical image analysis, leading to tangible improvements in clinical efficiency. By ensuring that subsequent diagnostic or analytical tasks are performed on accurately selected data, our framework can reduce processing time and support radiologists' decision-making. Although the current system relies on single-source data and a majority voting mechanism, future work will be directed at large-scale, multi-institutional validation and the implementation of more sophisticated ensemble techniques. In conclusion, this work provides a powerful and efficient solution that standardizes a critical step in neuroimaging, paving the way for faster diagnostic workflows and enhanced patient care.

AUTHORS CONTRIBUTION

David Cahyadi: Formal Analysis, Investigation, Project Administration, Resources, Software, Validation, Visualization, Original Drafting Writing, Review & Editing Writing;
Edwin Pramana: Project Administration, Supervision, Validation, Review & Editing Writing;
Rudi Limantara: Project Administration, Supervision, Formal Analysis, Review & Editing Writing;
I Gusti Lanang Ngurah Agung Artha Wiguna: Supervision, Validation, Review;
Maria Florencia Deslivia: Supervision, Validation, Acquisition of Data;
Ivan Alexander Liando: Conceptualization, Acquisition of Data, Investigation;

COPYRIGHT

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

REFERENCES

[1] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention," May 11, 2023, *arXiv*: arXiv:2305.07027. doi: 10.48550/arXiv.2305.07027.

[2] D. Qin *et al.*, "MobileNetV4 -- Universal Models for the Mobile Ecosystem," Apr. 16, 2024, *arXiv*: arXiv:2404.10518. doi: 10.48550/arXiv.2404.10518.

[3] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "RepViT: Revisiting Mobile CNN From ViT Perspective," Mar. 14, 2024, *arXiv*: arXiv:2307.09283. doi: 10.48550/arXiv.2307.09283.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[5] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 04, 2019, *arXiv*: arXiv:1711.05101. doi: 10.48550/arXiv.1711.05101.

[6] E. Giunchiglia and T. Lukasiewicz, "Coherent Hierarchical Multi-Label Classification Networks," arXiv.org. Accessed: Oct. 05, 2024. [Online]. Available: https://arxiv.org/abs/2010.10151v1

[7] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection," in *Computer Vision – ECCV 2022*, vol. 13672, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13672. , Cham: Springer Nature Switzerland, 2022, pp. 596–613. doi: 10.1007/978-3-031-19775-8_35.

[8] K. Kowsari *et al.*, "HMIC: Hierarchical Medical Image Classification, A Deep Learning Approach," *Information*, vol. 11, no. 6, p. 318, Jun. 2020, doi: 10.3390/info11060318.

[9] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Min Knowl Disc*, vol. 22, no. 1, pp. 31–72, Jan. 2011, doi: 10.1007/s10618-010-0175-9.

[10] Z. Yan *et al.*, "HD-CNN: Hierarchical Deep Convolutional Neural Network for Large Scale Visual Recognition," May 15, 2015, *arXiv*: arXiv:1410.0736. doi: 10.48550/arXiv.1410.0736.

[11] I. G. L. N. A. Artha Wiguna *et al.*, "A deep learning approach for cervical cord injury severity determination through axial and sagittal magnetic resonance imaging segmentation and classification," *Eur Spine J*, Aug. 2024, doi: 10.1007/s00586-024-08464-7.

[12] R. Limantara, Y. Kristian, E. I. Setiawan, D. Cahyadi, I. G. L. N. A. A. Wiguna, and M. F. Deslivia, "SpinalAI: A Deep Learning Approach to Predict Vertebrae-Column Level, Structure, and Foraminal on Cervical Spine Axial MRI Images," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, Jul. 2024, pp. 42–47. doi: 10.1109/ICICoS62600.2024.10636887.

[13] T. Shimizu *et al.*, "A multimodal machine learning model integrating clinical and MRI data for predicting neurological outcomes following surgical treatment for cervical spinal cord injury," *Eur Spine J*, Apr. 2025, doi: 10.1007/s00586-025-08873-2.

[14] L.-D. Azoulay *et al.*, "Deep learning approaches to predict late gadolinium enhancement and clinical outcomes in suspected cardiac sarcoidosis," *Sarcoidosis Vasc Diffuse Lung Dis*, vol. 42, no. 1, p. 15378, Mar. 2025, doi: 10.36141/svdld.v42i1.15378.

[15] C. Su, K. Miao, L. Zhang, and X. Dong, "Deep learning based on ultrasound images to predict platinum resistance in patients with epithelial ovarian cancer," *Biomed Eng Online*, vol. 24, no. 1, p. 58, May 2025, doi: 10.1186/s12938-025-01391-8.

[16] X. Mei *et al.*, "RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e210315, Sep. 2022, doi: 10.1148/ryai.210315.
[17] M. Lyu *et al.*, "M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research," *Sci Data*, vol. 10, p. 264, May 2023, doi: 10.1038/s41597-023-02181-4.
[18] A. E. Kavur *et al.*, "CHAOS Challenge -- Combined (CT-MR) Healthy Abdominal Organ Segmentation," *Medical Image Analysis*, vol. 69, p. 101950, Apr. 2021, doi: 10.1016/j.media.2020.101950.
[19] J. Cohen-Adad *et al.*, "Open-access quantitative MRI data of the spinal cord and reproducibility across participants, sites and manufacturers," *Sci Data*, vol. 8, p. 219, Aug. 2021, doi: 10.1038/s41597-021-00941-8.
[20] R. Tibrewala *et al.*, "FastMRI Prostate: A Publicly Available, Biparametric MRI Dataset to Advance Machine Learning for Prostate Cancer Imaging," Apr. 18, 2023, *arXiv*: arXiv:2304.09254. doi: 10.48550/arXiv.2304.09254.
[21] J. W. van der Graaf *et al.*, "Lumbar spine segmentation in MR images: a dataset and a public benchmark," *Sci Data*, vol. 11, no. 1, p. 264, Mar. 2024, doi: 10.1038/s41597-024-03090-w.
[22] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.
[23] A. Howard *et al.*, "Searching for MobileNetV3," Nov. 20, 2019, *arXiv*: arXiv:1905.02244. doi: 10.48550/arXiv.1905.02244.