Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Performance Analysis of a Convolutional Neural Network across 2D Medical Imaging Modalities

Lauren Choi[1], Sydney Kim[2], Jimin Park[3], Justin Park[4], Youngjoon Ryu[5]

[1]Lambert High School, 805 Nichols Rd, Suwanee, GA, 30024, USA
[2]The Bronx High School of Science, 75 W 205th St, Bronx, NY, 10468, USA
[3]US International School, 32, Seocho-daero 70-gil, Seocho-gu, Seoul, Republic of Korea
[4]Seckinger High School, 3655 Sardis Church Rd. Buford, GA, 30519, USA
[5]Taejon Christian International School, 77, Yongsan 2-ro, Yuseong-gu, Daejeon, Republic of Korea

Abstract

Medical imaging is an essential component of modern healthcare. Medical imaging techniques such as X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) provide complementary views of anatomical structures and pathological changes, often serving as the first line of evidence in clinical decision-making. The growing demand for faster and more accurate interpretation of medical images has increased interest in artificial intelligence, in particular, convolutional neural networks (CNNs). CNNs have achieved high performances in many computer vision tasks, but their effectiveness can vary depending on the imaging modality, data quality, and the disease context. The images used in this experiment include dental X-rays, bone fracture X-rays, brain stroke CT scans, and Alzheimer's MRI images. The goal of this study is to conduct a comparative evaluation of CNN architectures across multiple two-dimensional medical imaging modalities. The results showed a strong overall performance, with high accuracy and balanced precision-recall tradeoffs in most datasets, and particularly strong outcomes from the brain stroke and dental datasets. The model consistently achieved competitive AUC values, underscoring its robustness and adaptability across diverse imaging modalities.

Keywords : Machine learning; convolutional neural network; multiclass classification; medical imaging; magnetic resonance imaging; computed tomography; X-ray; radiography

1. Introduction

Medical imaging plays a crucial role in modern healthcare, supporting disease diagnosis, treatment planning, and monitoring. Modalities such as X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) enable clinicians to visualize internal structures non-invasively and often serve as the first step in identifying a disease or injury. However, diagnostic errors remain a concern, particularly in high-volume clinical settings. A study has shown that radiologist fatigue from long workdays can reduce focus and diagnostic accuracy, increasing the likelihood of missed fractures.[1] In such stressful environments, where radiologists may interpret hundreds of images daily, subtle findings, overlapping anatomical structures, or physician fatigue can contribute to diagnostic oversights. These errors carry significant

clinical consequences: among 1054 patients studied, 199 experienced adverse outcomes, of whom 34.7% died within 30 days, 30.7% required intensive care, and 51.8% experienced delays in necessary surgery.[2]

To address these challenges, researchers have increasingly explored artificial intelligence (AI), particularly convolutional neural networks (CNNs), as decision-support tools for medical image interpretation. CNNs are deep learning models specialized for image recognition tasks, capable of automatically extracting hierarchical features from complex medical images with high accuracy. For example, AI assistance in chest radiography has been shown to improve radiologists' sensitivity in detecting pneumonia, pneumothorax, and lung nodules while reducing interpretation time.[3,4] Similarly, CNN-based models have achieved performance comparable to that of experienced radiologists in detecting thyroid nodules, hepatocellular carcinoma, and musculoskeletal fractures.[5-7] Systematic reviews and meta-analyses further confirm that AI systems often perform at levels similar to human experts, with optimal outcomes achieved when AI complements radiologist expertise.[8,9]

Despite these advances, challenges persist. CNNs often entail high computational cost and significant training time, require large annotated datasets, and are prone to overfitting, especially with limited data. Regularization strategies such as dropout, batch normalization, and data augmentation are essential to mitigate these issues. Moreover, the predominance of supervised learning in CNN training limits applicability in data-scarce settings, prompting research into unsupervised, semi-supervised, and transfer learning approaches. Ongoing efforts in model compression, pruning, and quantization also aim to make CNNs more lightweight for mobile and embedded devices.[7] Future directions include enhancing interpretability through biologically inspired modeling, developing data-efficient architectures to reduce dependency on annotated datasets, and integrating CNNs with complementary computational paradigms.[10]

In this study, we evaluated the performance of a CNN across four publicly available datasets representing different medical imaging modalities. Our goal was to assess how imaging modality and image characteristics influence diagnostic accuracy. The model achieved the highest performance on high-contrast images with well-defined structural features, while performance declined on datasets with visually similar or low-contrast features. These findings suggest that CNN performance is strongly influenced by the visual characteristics and feature separability of the dataset rather than the imaging modality alone, emphasizing the importance of image clarity and quality in medical AI applications. Under suitable imaging conditions, CNNs may have the potential to support radiologists in achieving more accurate and consistent interpretations, provided they are validated in clinical settings.

This paper is organized as follows: In the Methods and Materials section, we describe the datasets and methods used, including data preparation and model development. The Results section reports the experimental results and evaluation metrics. The Discussion section discusses the limitations of this work and suggests future directions. Finally, the Conclusion section concludes the study.

Methods and Materials

a. Dataset Description

For this research, four publicly available medical imaging datasets were acquired from Kaggle®, a platform that provides reliable datasets for research in various domains. The datasets include the *Dental Radiography*, the *Bone Fracture Detection: Computer Vision Project*, the *Brain Stroke CT Dataset*, and the *Alzheimer's Disease Multiclass Images Dataset.* These datasets represent different medical imaging modalities, such as X-ray, CT, and MRI, and were selected to evaluate the generalizability of CNNs across diverse classification tasks. A summary of the datasets is provided in Table 1, and detailed descriptions are presented in the following subsections. The datasets are ordered by imaging modality and, within each modality, by increasing number of classes. This order is maintained throughout the paper.

Table 1. Overview of the datasets.

Dataset	Modality	# Images Used	# Classes	Class Indices/Names
Dental	X-ray	4652	4	[0: Cavity, 1: Fillings, 2: Impacted Tooth, 3: Implant]
Bone Fracture	X-ray	2060	6	[0: Elbow positive, 1: Fingers positive, 2: Forearm fracture, 3: Humerus, 4: Shoulder fracture, 5: Wrist positive]
Brain Stroke	CT	6650	3	[0: Bleeding, 1: Ischemia, 2: Normal]
Alzheimer's	MRI	12000	4	[0: MildDemented, 1: ModerateDemented, 2: NonDemented, 3: VeryMildDemented]

i. Dental Dataset

The *Dental Radiography* dataset contains 1272 X-ray images, divided into training (1076), validation (122), and test (74) sets, each accompanied by annotation files specifying bounding box coordinates and class labels. Each bounding box defined a region of interest (ROI) corresponding to dental conditions such as fillings, implants, impacted teeth, or cavities. To exclude unusually small or large regions, only ROIs with width, height, and area within the interquartile range (25th–75th percentile) were retained. Each X-ray was then converted to grayscale, and the retained ROIs were individually cropped and resized to 224 × 224 pixels. This procedure yielded a total of 4652 cropped images, with 4023 for training, 392 for validation, and 237 for testing. Dental radiographs enable dentists to observe changes in hard and soft tissues, assess dental and jawbone development in children, and evaluate facial or oral injuries. The dataset is suitable for training and evaluating machine learning models for dental condition classification.

ii. Bone Fracture Dataset

The *Bone Fracture Detection: Computer Vision Project* dataset contains 4148 X-ray images divided into training (3631), validation (348), and test (169) sets, each organized into separate folders for images and

labels. The images are labeled across six classes based on anatomical location: Elbow Positive, Fingers Positive, Forearm Fracture, Humerus, Shoulder Fracture, and Wrist Positive. Each image is annotated with bounding boxes or pixel-level segmentation masks indicating the location and extent of the fracture. Images with empty or missing annotations were excluded. This procedure yielded a total of 2060 images, with 1804 for training, 173 for validation, and 83 for testing. To address the limited size of the initial test set, the 1804 images from the initial training set were further redistributed into training (1443; 80%), validation (180; 10%), and test (181; 10%) subsets, using an initial 80-20 split for training and validation sets, followed by a 50-50 split of the validation set to create the test set. All random operations were performed with a fixed seeding protocol as described in the Validation and Reliability section to ensure reproducibility and performance stability. The dataset has a diversity of anatomical regions and fracture types, making it suitable for training and evaluating machine learning models for automated fracture detection and classification.

iii. Brain Stroke Dataset

The *Brain Stroke CT Dataset* contains 6650 labeled brain CT images categorized into three classes: Bleeding (1093), Ischemia (1130), and Normal (4427). An additional External Test folder with 200 CT scans was excluded from this study. The images were randomly split into training (4256; 64%), validation (1064; 16%), and test (1330; 20%) subsets, using an initial 80–20 split for training and test sets, followed by an 80–20 split of the training set to create the validation set. All random operations were performed with a fixed seeding protocol as described in the Validation and Reliability section to ensure reproducibility and performance stability. The dataset contains images with varying resolutions, reflecting real-world variability in medical imaging, and includes both ischemic and hemorrhagic stroke types, making it suitable for training and evaluating machine learning models for stroke detection and classification.

iv. Alzheimer's Dataset

The *Alzheimer's Disease Multiclass Images Dataset* contains 44000 brain MRI images categorized into four classes based on disease severity: NonDemented (12800), VeryMildDemented (11200), MildDemented (10000), and ModerateDemented (10000). For this study, 1000 images per class were randomly sampled to create independent training, validation, and test splits. This procedure yielded 4000 images per split and a total of 12000 images. All images are skull-stripped, and the dataset was augmented and upsampled by its curators to address class imbalance, making it suitable for training and evaluating machine learning models for Alzheimer's stage classification.

b. Dataset Preparation

To ensure consistency and facilitate effective CNN training across all datasets, several preprocessing steps were applied. Class distributions were balanced through downsampling of majority classes or selection of a subset, reducing computational cost and mitigating the risk of bias, underfitting, or overfitting. All images were resized to fixed dimensions: 256 × 256 pixels for the Bone Fracture dataset and 224 × 224 for the Dental, Brain Stroke, and Alzheimer's datasets. Pixel intensities, initially ranging from 0-255, were normalized to the range [0, 1] to standardize input ranges, improving optimization stability and training efficiency. Class labels were converted into one-hot encoded vectors, with the encoding procedure

adjusted according to each dataset's number of classes, guaranteeing consistency across training, validation, and test sets. No data augmentation techniques such as rotation, flipping, or contrast adjustment were applied to any of the datasets, and no additional preprocessing steps beyond those described above were performed.

c. Model

All experiments were conducted in Google Colab using an NVIDIA Tesla T4 GPU with 12.6 GB RAM and 15 GB storage. The environment was configured with Python 3.10, TensorFlow 2.15, and other supporting libraries, including NumPy, Matplotlib, Pandas, Scikit-learn, Keras, Pathlib, and Tqdm.

i. Architecture

The CNN model architecture is shown in Figure 1.

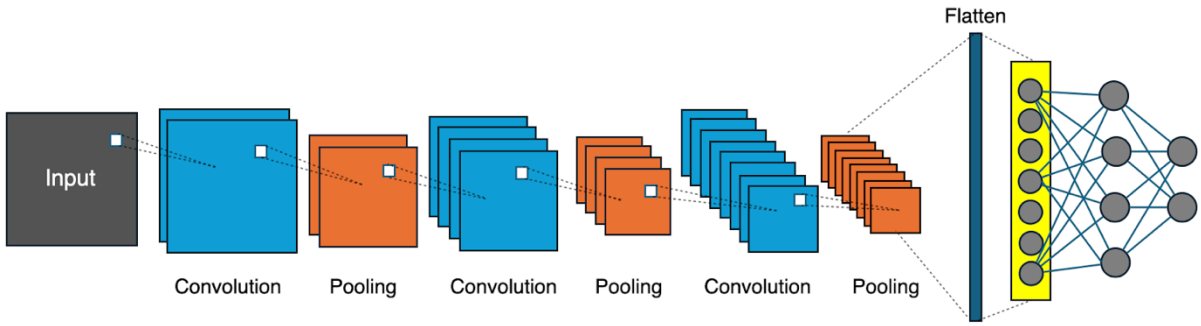

Fig 1. CNN model architecture. The model consists of three convolutional blocks (Conv2D with ReLU activation followed by MaxPooling), a flattening layer, and two dense layers with dropout regularization. The final output layer utilizes softmax activation for multi-class classification. Each square in the feature map represents 16 filters.

The model begins with a two-dimensional convolutional layer composed of 32 filters of size 3 × 3. This layer applies the rectified linear unit (ReLU) activation function to introduce non-linearity and detect basic features such as edges and textures. A subsequent max pooling layer with a 2 × 2 pool size reduces the spatial dimensions of the feature maps, retaining salient information while minimizing computational load.

The second convolutional block expands to 64 filters (3 × 3 kernel) with ReLU activation, followed by max pooling. This block allows the model to capture more complex patterns, such as curves and localized shapes. The third convolutional block expands to 128 filters, continuing hierarchical feature extraction and enabling learning of high-level abstractions such as object parts or structural patterns. Another max pooling layer follows, after which the feature maps are flattened into a one-dimensional vector of activations, preparing the data for fully connected processing.

The fully connected portion begins with a dense layer of 256 neurons with ReLU activation, integrating the extracted features into complex and discriminative representations. To mitigate overfitting, a dropout layer with a rate of 50% is applied, randomly deactivating half of the neurons during training. A second dense layer of 128 neurons with ReLU activation follows. Another dropout layer, this time with a dropout rate of 30%, provides additional regularization. The final output layer contains a number of neurons equal

to the number of target classes, with softmax activation producing a probability distribution across classes and enabling clear classification decisions.[11] For example, a dataset with four classes (e.g., dental dataset; fillings, implant, impacted tooth, cavity) corresponds to four output nodes, each representing one class.

ii. Training and Hyperparameters

The model was trained using the Adam optimizer with an initial learning rate of 110-4, providing a balance between convergence speed and stability.[12] No learning rate scheduling was applied. The categorical cross-entropy loss function was employed, appropriate for multiclass classification tasks with softmax outputs. Training was performed for a maximum of 100 epochs. Early stopping was employed with a patience of 5 epochs, monitoring validation loss to prevent overfitting and unnecessary computation. The model achieving the lowest validation loss during training was retained using model checkpointing. A batch size of 64 was used for all datasets, except for the Bone Fracture dataset, which required a reduced batch size of 32 due to GPU memory constraints associated with higher-resolution images.

Training hyperparameters and configurations used across all experiments are summarized in Table 2.

Table 2. Training hyperparameters and configurations used across all experiments

Parameter	Value
Batch size	64 (32 for Bone Fracture dataset)
Early stopping	Yes (patience = 5, monitor = ‘val_loss’)
Learning rate	110-4
Learning rate scheduling	Not used
Loss function	Categorical cross-entropy
Maximum epochs	100
Optimizer	Adam

iii. Validation and Reliability

To ensure the reliability of our findings, we conducted two independent training realizations for each dataset. For the Bone Fracture, Brain Stroke, and Alzheimer’s datasets, the two runs utilized different random seeds (42 and 123) to vary the data partitioning. For the Dental dataset, the experiment was repeated to reduce the influence of training stochasticity. The original fixed partitions provided by the dataset curators were used, as described in the Dental Dataset section, to preserve reproducibility, direct comparison, and avoid potential data leakage arising from arbitrary re-splitting.

The results reported in the Results section represent the mean performance metrics across these independent runs. This protocol was adopted to observe the model's sensitivity to data shuffling and to provide an initial measure of performance stability given the computational constraints of the training environment.

Results

The performance of the model was evaluated on all four datasets. To confirm the reliability of the findings, all metrics were averaged across two independent experimental runs. Quantitative metrics, including accuracy, precision, recall, F1-score, specificity, and AUC-ROC, were used to evaluate model performance. Precision, recall, F1-score, and specificity were calculated using weighted averaging to account for class imbalance. The evaluation results are summarized in Table 3 as Mean ± Range. While performance metrics represent aggregated results, the accuracy and loss curves, confusion matrices, and ROC curves presented in Figures 2, 3, and 4 correspond to the primary experimental realization to provide a granular view of model behavior. For the confusion matrices and ROC curves, class indices correspond to those reported in Table 1.

Table 3. Evaluation results. The best performance for each metric is highlighted in bold.

Dataset	**Evaluation metrics**					
	Accuracy	Precision	Recall	F1-score	Specificity	AUC
Dental	0.8439	0.8325	0.8439	0.8307	0.8536	0.97
Dental_2 (temporary)	0.8439	0.8325	0.8439	0.8307	0.8536	0.97
Bone Fracture	0.7182	0.7513	0.7182	0.7240	**0.9421**	0.93
Bone_2 (temporary)	0.8111	0.9166	0.8000	0.8543	0.9359	0.94
Brain Stroke	**0.9436**	**0.9440**	**0.9436**	**0.9427**	0.9056	**0.99**
Brain_2 (temporary)	0.9233	0.9230	0.9233	0.9223	0.9398	0.99
Alzheimer's	0.7505	0.7581	0.7505	0.7500	0.9168	0.93
Alz_2 (temporary)	0.7792	0.7783	0.7792	0.7746	0.9264	0.95

Table 3. Evaluation results reported as the Mean ± Range calculated from two independent realizations. The best performance for each metric is highlighted in bold.

Dataset	**Evaluation metrics**					
	Accuracy	Precision	Recall	F1-score	Specificity	AUC
Dental	0.8439 ± 0.0000	0.8325 ± 0.0000	0.8439 ± 0.0000	0.8307 ± 0.0000	0.8536 ± 0.0000	0.9700 ± 0.0000
Bone Fracture	0.7647 ± 0.0465	0.8339 ± 0.0827	0.7591 ± 0.0409	0.7892 ± 0.0651	0.9390 ± 0.0031	0.9350 ± 0.0050
Brain Stroke	0.9335 ± 0.0101	0.9335 ± 0.0105	0.9335 ± 0.0101	0.9325 ± 0.0102	0.9227 ± 0.0171	0.9900 ± 0.0000
Alzheimer's	0.7649 ± 0.0144	0.7682 ± 0.0101	0.7649 ± 0.0144	0.7623 ± 0.0123	0.9216 ± 0.0048	0.9400 ± 0.0100

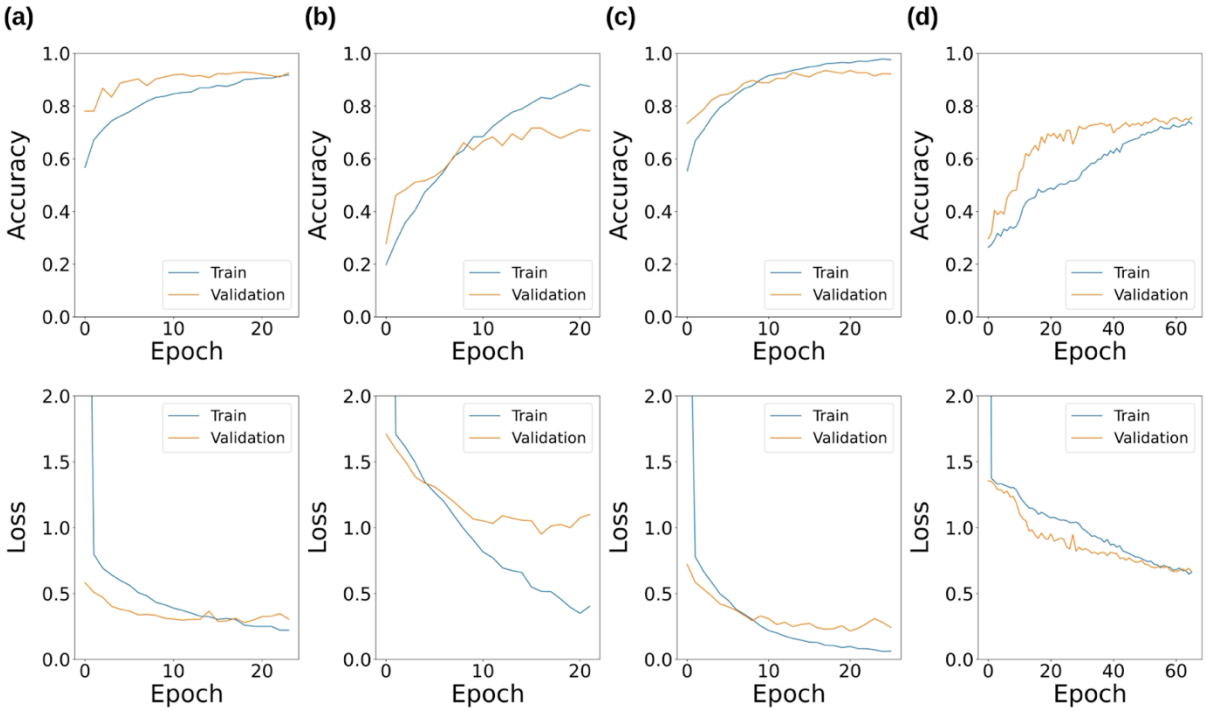

Fig 2. Training and validation accuracy (top) and loss (bottom) over epochs for the datasets: (a) Dental, (b) Bone Fracture, (c) Brain Stroke, and (d) Alzheimer's.

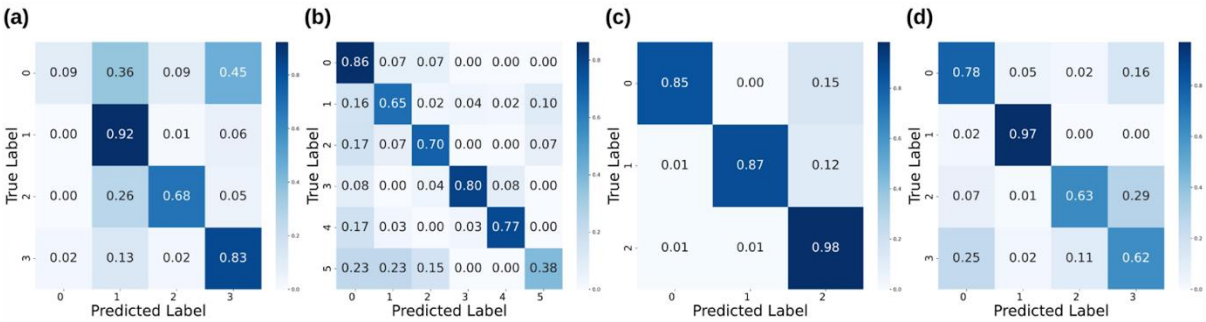

Fig 3. Confusion matrices for the datasets: (a) Dental, (b) Bone Fracture, (c) Brain Stroke, and (d) Alzheimer's.

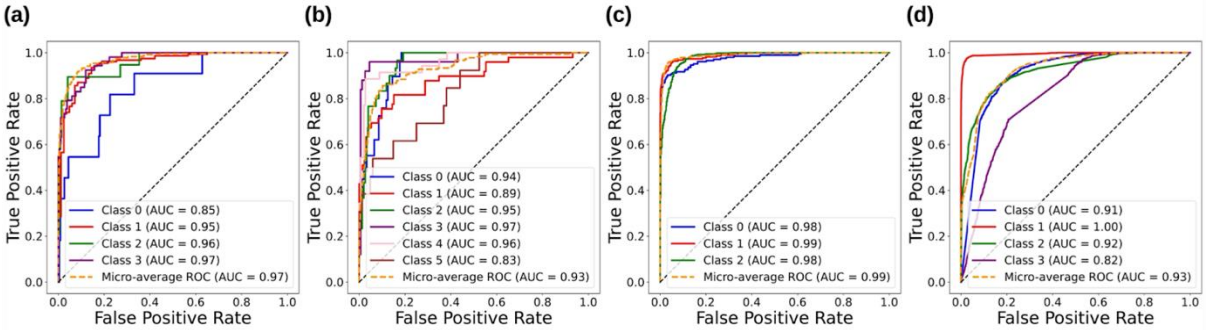

Fig 4. ROC curves for the datasets: (a) Dental, (b) Bone Fracture, (c) Brain Stroke, and (d) Alzheimer's.

a. Dental

The Dental dataset demonstrated stable and balanced performance, with a mean accuracy of 0.8439, precision of 0.8325, recall of 0.8439, F1-score of 0.8307, and specificity of 0.8536. The results showed zero variance across runs due to the fixed-split protocol. High specificity indicates that the model effectively identified non-target dental conditions, supporting its suitability for automated dental radiograph classification. The AUC of 0.97 confirms strong class separability. Figure 2a shows smooth convergence with low variance, and the confusion matrix in Figure 3a indicates low misclassification rates overall. However, the Cavity class exhibited noticeably poorer performance, with precision of 0.50, recall of 0.09, and F1-score of 0.15, alongside higher confusion with the Implant and Fillings classes. This suggests that the model struggles to distinguish cavities from visually similar dental features, likely due to class imbalance, as only 11 of the 237 test images belonged to the Cavity class. These findings highlight the potential need for additional training examples or targeted augmentation.

b. Bone Fracture

The Bone Fracture dataset posed the most significant challenges, but showed notable improvement across realizations, achieving a mean accuracy of 0.076470.05, precision of 0.7513, recall of 0.7182, F1-score of 0.7240, and specificity of 0.9421. Higher precision than recall indicates the model was more successful at correctly identifying fractures when it predicted positive cases, although it missed some true fractures. High specificity (0.9390) and an AUC of 0.93 suggest the model captured class separability at the probability level despite lower overall classification accuracy. Figure 2b shows slower and noisier convergence, while the confusion matrix in Figure 3b reveals frequent misclassifications among certain fracture types; images of Forearm Fracture and Fingers Positive were often misclassified as Elbow Positive, reflecting challenges in distinguishing smaller or visually similar fracture types. Limited representation of specific classes and subtle fracture features appear to be primary bottlenecks.

c. Brain Stroke

The brain stroke dataset achieved the highest and most consistent performance, with a mean accuracy of 0.9436, precision of 0.9440, recall of 0.9436, F1-score of 0.9427, and specificity of 0.9497. This indicates the model reliably captured both positive and negative cases across different data partitions. Figure 2c shows smooth convergence, indicating stable optimization during the training process, and Figure 3c exhibits minimal misclassification, confirming that high contrast CT features facilitated accurate discrimination. With Figure 4c displaying a sharply rising ROC curve nearing the maximum, the model also demonstrated an AUC of 0.99, showcasing a near-perfect performance.

d. Alzheimer's

The Alzheimer's dataset showed moderate but stable performance across runs, with a mean accuracy of 0.7505, precision of 0.7581, recall of 0.7505, F1-score of 0.7500, and specificity of 0.9168. Precision slightly exceeding recall suggests the model was better at avoiding false positives than capturing all true cases. The consistency of these results across different seeds (Range = 0.01) reinforces the model's reliability. Figure 2d indicates slower convergence and various fluctuations, and Figure 3d shows frequent

misclassifications between adjacent classes, often confusing NonDemented with VeryMildDemented and VeryMildDemented with MildDemented. This is likely due to the subtle differences across Alzheimer's stages and the variability of MRI scans. With an AUC of 0.93, Figure 4d confirms reasonable separability between classes despite these challenges.

Discussion

Across datasets, the Brain Stroke dataset achieved the highest and most consistent performance, confirming that high-contrast CT features and clear structural abnormalities facilitate highly accurate discrimination. In contrast, the Bone Fracture and Alzheimer's datasets exhibited moderate performance. While the Bone Fracture dataset showed improved reliability across runs, it continues to reflect the challenges of distinguishing subtle, small-scale fracture lines across multiple anatomical classes. Similarly, the Alzheimer's dataset highlighted the difficulties posed by subtle and variable MRI features. Notably, the Dental dataset maintained strong results, benefiting from the clear structural features of radiographs, despite the specific challenges identified in the Cavity class.

Despite these differences in mean classification accuracy, the consistently high AUCs and the low variance across realizations suggest that the CNNs learned meaningful and stable separability across all datasets. Future improvements in dataset balancing, targeted data augmentation, or the implementation of transfer learning could further enhance performance, particularly for the more underrepresented or visually ambiguous cases found in the Bone Fracture and Alzheimer's datasets.

We acknowledge the limitations regarding the statistical depth of our performance variance analysis. While we conducted dual-run experiments across all datasets to observe model stability, the high computational requirements and GPU time constraints associated with training CNN architectures on large-scale medical datasets precluded the use of intensive k-fold cross-validation or statistical testing (e.g., t-tests or ANOVA). However, the minimal variance observed between runs (<5%) suggests that the model is robust to different data partitions.

This study relied on publicly available datasets, which may not represent the full diversity of real-world medical imaging. Additional validation using diverse, real-world imaging data is necessary before extending these comparative modeling results toward clinical applications. Variability in image quality, scanner settings, and patient demographics could also affect generalizability[13]. Future work should focus on expanding datasets, particularly for underrepresented classes, and leveraging techniques such as attention mechanisms, multimodal learning, and transfer learning to improve performance on subtle or complex imaging features. These enhancements could improve the generalizability of CNNs, forming a starting place for future studies that evaluate clinical reliability.

Additionally, this study did not compare our model to pretrained architectures such as those using ImageNet. Although transfer learning is a common strategy for improving performance on small or subtle medical imaging datasets, using pretrained models was not possible within the available hardware resources. The significant GPU memory requirements and extended training times associated with state-of-the-art pretrained CNNs exceeded the constraints of this project. As a result, the comparative

evaluations presented here reflect only models trained from scratch, which may underestimate the performance achievable with more compute-intensive approaches.

Across the four imaging modalities, the performance patterns observed are consistent with modality-dependent limitations in previous deep learning research. The Dental and Bone Fracture datasets, which both used X-ray imaging, required the model to detect fine, low-contrast structural differences which CNNs trained from scratch often struggle to learn reliably[14]. This aligns with known CNN sensitivity to texture and contrast, particularly when class boundaries are small or visually ambiguous[15]. In contrast, the Brain Stroke CT dataset exhibited stronger class separability due to larger, more distinct density differences between ischemic, hemorrhagic, and normal tissue, allowing the model to achieve higher AUC values despite class imbalance[16]. The Alzheimer's MRI dataset displayed the opposite pattern. Although AUC values were high, accuracy and F1-scores were lower. This reflects the challenge CNNs face when distinguishing disease severity stages that differ by subtle volumetric changes[17]. These modality-specific outcomes are similar to established limitations in CNNs with their dependence on high-contrast features, vulnerability to class imbalance, and difficulty modeling nuanced anatomical variation in particular. This supports that the observed errors are not dataset artifacts but reflect broader constraints of CNN-based medical image classification systems.

Some of the performances observed across the four datasets can be related to this study's technical constraints. The imbalance in training and validation curves, particularly in the Bone Fracture and Alzheimer's datasets, is reflected in their lower recall. This suggests the inherent visual complexity and subtle class boundaries of these modalities made it difficult for a model trained from scratch to generalize effectively. The confusion matrices reveal that misclassifications were concentrated in visually similar classes. This indicates that the model's capacity for fine-grained feature extraction was limited compared to deeper, pretrained architectures. This is further evidenced by AUC values that are noticeably higher than accuracy or F1-scores, implying that while the model has high discriminative potential, it struggled to achieve consistent precision and recall across categories where visual features are highly overlapped.

Despite the robust numerical evaluation provided, this study is not without limitations. Specifically, visual interpretability tools, such as Grad-CAM heatmaps or feature visualizations, were not implemented in the current pipeline. While the confusion matrices and metric gaps offer a proxy for understanding model behavior, future work will integrate these visualization techniques to more precisely localize the diagnostic features the CNN prioritizes across different medical modalities.

Conclusion

This study conducted a comparative evaluation of CNN performance across multiple medical imaging modalities, including X-rays, CT, and MRI. The results indicate that CNNs perform best on high-contrast images with clearly distinguishable features, whereas high variability and subtle inter-class differences limit performance. Dataset characteristics appear to influence outcomes more than imaging modality alone. These findings underscore both the potential and the limitations of CNNs, emphasizing that performance depends heavily on image quality and feature representation. Further research on advanced architectures, transfer learning, and dataset expansion will be important for improving model performance and future studies on potential clinical applications.

Acknowledgements

All student authors contributed equally under the supervision of Dr. Soo Min Oh. The authors are listed in alphabetical order by their last names. The authors would like to thank Dr. Oh for valuable guidance and feedback during the preparation of this work.

Data Availability

The data supporting this study are publicly available in the Kaggle dataset, accessible at the addresses https://www.kaggle.com/datasets/imtkaggleteam/dental-radiography (Dental), https://www.kaggle.com/datasets/pkdarabi/bone-fracture-detection-computer-vision-project (Bone Fracture), https://www.kaggle.com/datasets/ozguraslank/brain-stroke-ct-dataset/data (Brain Stroke), https://www.kaggle.com/datasets/aryansinghal10/alzheimers-multiclass-dataset-equal-and-augmented/data (Alzheimer's).

No new data were generated for this study.

Code Availability Statement

The codes developed in this study are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

This study did not involve any human or animal subjects.

Declaration of conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

CRediT authorship contribution statement

Lauren Choi: Investigation, Software, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Sydney Kim:** Investigation, Writing - Original Draft, Writing - Review & Editing. **Jimin Park:** Investigation, Writing - Original Draft, Visualization. **Justin Park:** Investigation, Software, Writing - Review & Editing, Visualization. **Youngjoon Ryu:** Investigation, Writing - Original Draft.

References

1. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long Radiology Workdays Reduce Detection and Accommodation Accuracy. Journal of the American College of Radiology. 2010 Sep;7(9):698–704.

2. Ahn Y, Hong GS, Park KJ, Lee CW, Lee JH, Kim SO. Impact of diagnostic errors on adverse outcomes: learning from emergency department revisits with repeat CT or MRI. Insights into Imaging. 2021 Nov 3;12(1).
3. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* [Internet]. 2021 Mar 31;8(1):1–74. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8
4. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Comput Vis ECCV*. 2014;346–61.
5. Pimrada P, Natamon C, Prakobkiat H. A comparison of artificial intelligence versus radiologists in the diagnosis of thyroid nodules using ultrasonography: a systematic review and meta-analysis. *Sci Rep*. 2022 Jun 29;279(11):5363–73.
6. Chatzipanagiotou OP, Loukas C, Vailas M, Machairas N, Kykalos S, Charalampopoulos G, et al. Artificial intelligence in hepatocellular carcinoma diagnosis: a comprehensive review of current literature. *J Gastroenterol Hepatol*. 2024 Jun 23.
7. Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, et al. CNN variants for computer vision: history, architecture, application, challenges and future scope. *Electronics*. 2021 Oct 11;10(20):2470.
8. Abadia AF, Yacoub B, Stringer N, Snoddy M, Kocher M, Schoepf UJ, et al. Diagnostic accuracy and performance of artificial intelligence in detecting lung nodules in patients with complex lung disease: a noninferiority study. *J Thorac Imaging* [Internet]. 2022 May 1;37(3):154–61. Available from: https://pubmed.ncbi.nlm.nih.gov/34387227/
9. Roest C, Fransen SJ, Kwee TC, Yakar D. Comparative performance of deep learning and radiologists for the diagnosis and localization of clinically significant prostate cancer at MRI: a systematic review. *Life (Basel)*. 2022 Sep 26;12(10):1490.
10. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev*. 2020 Apr 21;53.
11. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net. arXiv:14126806 [Internet]. 2015 Apr 13; Available from: https://arxiv.org/abs/1412.6806
12. Bock S, Goppold J, Weiß M. An improvement of the convergence proof of the ADAM-Optimizer. arXiv:180410587 [Internet]. 2018 Apr 27; Available from: https://arxiv.org/abs/1804.10587
13. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. Sheikh A, editor. PLOS Medicine [Internet]. 2018 Nov 6;15(11):e1002683. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683
14. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis. 2017 Dec;42(1):60–88.
15. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness [Internet]. Openreview.net. 2019 [cited 2025 Nov 24]. Available from: https://openreview.net/forum?id=Bygh9j09KX&trk=public_post_comment-text

16. Zhu G, Chen H, Jiang B, Chen F, Xie Y, Wintermark M. Application of Deep Learning to Ischemic and Hemorrhagic Stroke Computed Tomography and Magnetic Resonance Imaging. Seminars in Ultrasound, CT and MRI [Internet]. 2022 Apr 1 [cited 2022 Nov 14];43(2):147–52. Available from: https://www.sciencedirect.com/science/article/pii/S0887217122000166
17. Jain R, Jain N, Aggarwal A, Hemanth DJ. Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images. Cognitive Systems Research [Internet]. 2019 Jan [cited 2019 Apr 17]; Available from: https://www.sciencedirect.com/science/article/pii/S1389041718309562

Tables and Figures

Table 1. Overview of the datasets.

Dataset	Modality	# Images Used	# Classes	Class Indices/Names
Dental	X-ray	4652	4	[0: Cavity, 1: Fillings, 2: Impacted Tooth, 3: Implant]
Bone Fracture	X-ray	2060	6	[0: Elbow positive, 1: Fingers positive, 2: Forearm fracture, 3: Humerus, 4: Shoulder fracture, 5: Wrist positive]
Brain Stroke	CT	6650	3	[0: Bleeding, 1: Ischemia, 2: Normal]
Alzheimer's	MRI	12000	4	[0: MildDemented, 1: ModerateDemented, 2: NonDemented, 3: VeryMildDemented]

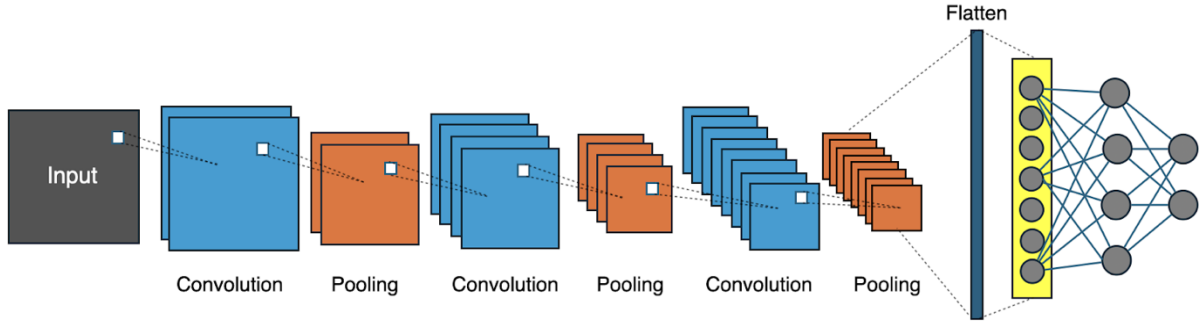

Fig 1. CNN model architecture. The model consists of three convolutional blocks (Conv2D with ReLU activation followed by MaxPooling), a flattening layer, and two dense layers with dropout regularization. The final output layer utilizes softmax activation for multi-class classification. Each square in the feature map represents 16 filters.

Table 2. Training hyperparameters and configurations used across all experiments

Parameter	Value
Batch size	64 (32 for Bone Fracture dataset)
Early stopping	Yes (patience = 5, monitor = 'val_loss')
Learning rate	110-4
Learning rate scheduling	Not used
Loss function	Categorical cross-entropy
Maximum epochs	100
Optimizer	Adam

Table 3. Evaluation results reported as the Mean ± Range calculated from two independent realizations. The best performance for each metric is highlighted in bold.

Dataset	Evaluation metrics						
	Accuracy	Precision	Recall	F1-score	Specificity	AUC	
Dental	0.8439 ± 0.0000	0.8325 ± 0.0000	0.8439 ± 0.0000	0.8307 ± 0.0000	0.8536 ± 0.0000	0.9700 ± 0.0000	
Bone Fracture	0.7647 ± 0.0929	0.8340 ± 0.1653	0.7591 ± 0.0818	0.7892 ± 0.1303	**0.9390 ± 0.0062**	0.9350 ± 0.0100	
Brain Stroke	**0.9335 ± 0.0203**	**0.9335 ± 0.0210**	**0.9335 ± 0.0203**	**0.9325 ± 0.0204**	0.9227 ± 0.0342	**0.9900 ± 0.0000**	
Alzheimer's	0.7649 ± 0.0287	0.7682 ± 0.0202	0.7649 ± 0.0287	0.7623 ± 0.0246	0.9216 ± 0.0096	0.9400 ± 0.0200	

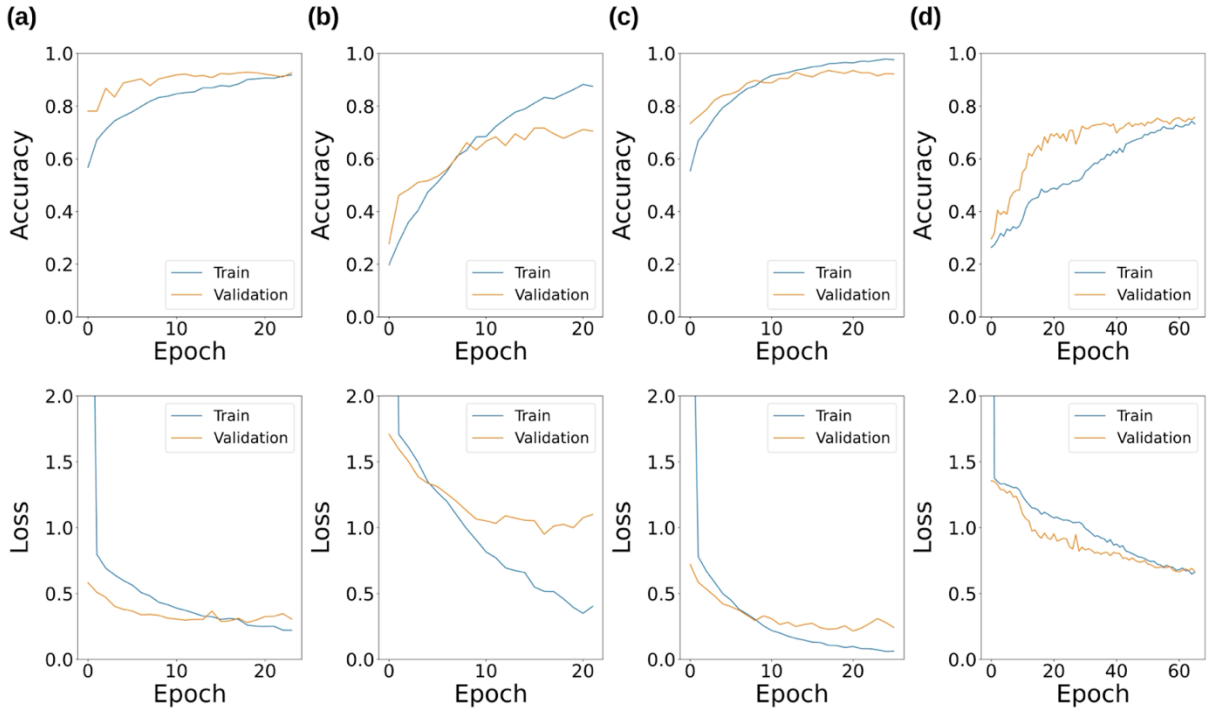

Fig 2. Training and validation accuracy (top) and loss (bottom) over epochs for the datasets: (a) Dental, (b) Bone Fracture, (c) Brain Stroke, and (d) Alzheimer's.

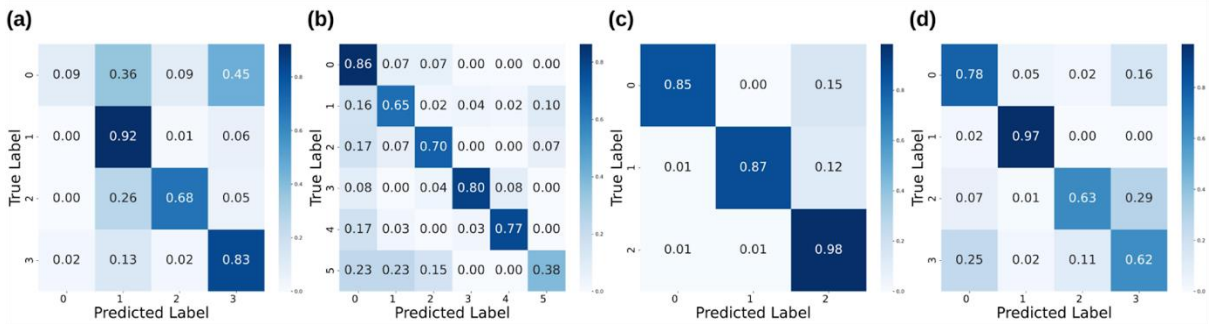

Fig 3. Confusion matrices for the datasets: (a) Dental, (b) Bone Fracture, (c) Brain Stroke, and (d) Alzheimer's.

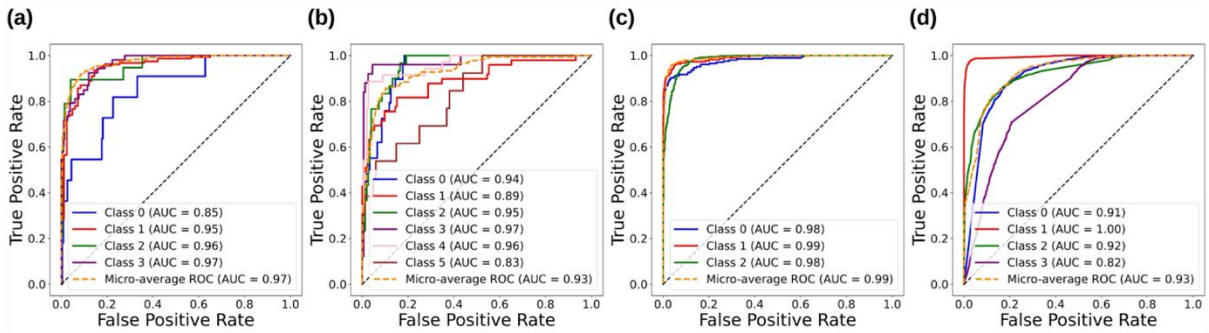

Fig 4. ROC curves for the datasets: (a) Dental, (b) Bone Fracture, (c) Brain Stroke, and (d) Alzheimer's.