Facilitating the Development of Personalized Healthcare AI Models: Optimizing Spinal X-Ray Imaging Protocols for Data-Constrained Environments

By Dev Gopal and Jack Whitaker

Abstract

Back pain is among the most common causes of disability in the United States, and up to 80% of individuals in the US will experience lower back pain at some point in their lives. Back pain in fact accounts for 2-3 percent of the total physician visits in the US, and is the second most common reason for hospitalization [1].

Although much work in AI imaging diagnostics has focused on maximizing accuracy through new techniques, an overlooked but crucial area of exploration is analysis in data-constrained environments. Due to the momentum of AI towards democratizing healthcare, we believe that one near-future is personalized AI models for diagnostics, driven from the most relevant data to you. For this paper, we evaluate the best imaging to take in a data-constrained environment by experimenting on different neural architectures performance on a small dataset.

Specifically, we developed and tested various lightweight deep-learning models to classify spine X-rays from the BUU-LSPINE dataset (with 800 classified images) as either normal or disordered. Facing constraints of limited data and significant class imbalance, we compared single-view (either AP or lateral) and dual-view approaches across four different CNN architectures: ResNet-18, MobileNetV2, DenseNet-121, and a simpler custom CNN. Our simple CNN achieved the highest accuracy at 85%. We also trained a dual-CNN, and experimented with different sampling methodologies. Although the results from these models were not able to optimally fit to the dataset, we were able to see different results that help us to better understand the ways that models fit to small datasets.

We present both a comparison between these architectures, and an analysis of the underlying data, which helps to contextualize the issues our models faced.

Introduction

Spine injuries are the second most common reason for hospitalization in the U.S. These hospitalizations come from spinal conditions, which can be serious if not caught early. Common causes range from anterolisthesis to retrolisthesis and more, all of which can be detected through X-Ray scans.

We envision that breakthroughs in AI will democratize healthcare access beyond what people can currently imagine. One likely advance of this is the development of personalized healthcare models [2]. However, helping develop techniques and models that can withstand low-data and biased data would be crucial to this breakthrough. We hope to help explore these low- and biased- data situations through our project, heralding a new age of AI.

While our initial work focused on a smaller dataset, specifically the open Mendeley Spine dataset, we've broadened our scope to focus instead on the larger BUU spine dataset. This had both advantages and disadvantages. While the larger dataset featured more realistic data and allowed us to train more general models, it also had less well segmented classes and focused on more real image scans, ideally allowing us to train more broadly applicable classification models.

While research on spinal diagnostics using machine learning is extensive, it features little coverage of low-data environments. In this study, we trained many different types of training regimes to see if we could get accurate data-fitting in low-data environments. Radiographic screening is often the only spine-imaging option in small clinics and rural hospitals, yet interpreting a single X-ray still demands scarce specialist time. Automated triage models could close that gap, but most computer-vision papers assume tens of thousands of labeled studies and balanced classes, conditions that do not exist where the clinical burden is sometimes highest. Building tools that work when data are scarce is therefore a prerequisite for equitable spine care.

Compact CNNs and pragmatic sampling can, in principle, squeeze strong performance out of limited material, but the field lacks a head-to-head analysis of how these techniques behave when both image count and class distribution are sharply constrained. Our study attempts to fill that gap. Using the BUU-LSPINE collection, 400 patient pairs of anteroposterior and lateral films, with four slip disorders represented alongside normals, we benchmark four lightweight backbones plus a custom CNN, test single versus dual-view pipelines, and probe three simple re-sampling strategies. The input to our CNN was a spinal X-Ray image, and the output is a binary prediction (either 'normal' or 'non-normal').

First, we will explore related work, then explain the details of the dataset we're working on, then explain our methodology, then show our results, and finally analyse what we've learned from this process.

Related Work

Spinal diagnostics is a rapidly growing field in the world of medical AI. One major breakthrough came in 2018, which was the introduction of multi-view convolutional neural networks in spinal diagnostics [3, 6, 7, 8]**.** These analyzed both the anterior-posterior (AP) and LA (lateral) angles to allow more rich analysis. After all, if surgeons often used multi-angle imaging, why shouldn't the machine learning models inspired by those surgeons do the same?

In prior literature, we found that CNNs were incredibly effective, and perhaps the most effective, in medical imaging diagnostics like spinal radiograph diagnostics. One literature review found pooled AUCs of 0.933–1.000 in ophthalmology, 0.864–0.937 in respiratory imaging, and 0.868–0.909 in breast imaging for DL algorithms, showing the broad applicability for CNNs in medical diagnostics [4, 9, 10].

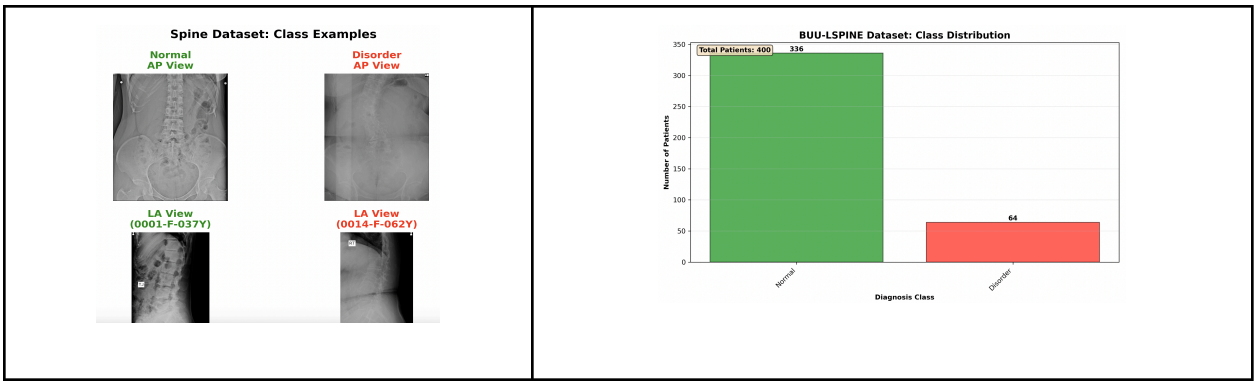
However, we also recognized that other architectures have been used more commonly in vision in recent years, and medical AI can sometimes lag behind technologically. In the past few years, some researchers have begun using transformer-based architectures very successfully. Some of these research studies use transformers in medical imaging, including analyzing spinal deformity[5, 11, 12]**.**

Generally most large scale studies on spinal imaging favors deep CNNs trained on large, private datasets. We experiment instead with a public 800-image set, a single cloud GPU, and models trainable in minutes. By building on compact CNNs and a handful of simple resampling tricks, we probe whether careful budgeting of parameters can stand in for vast data reserves. Additionally, we experimented with training dual-CNNs to evaluate this data, which was inspired by Wu et. al's 2018 work. We also use major advances in transformer techniques to inform advanced DualCNN and HybridCNN approaches, such as the attention mechanism. Although not using transformers due to their low accuracy in low-data low-compute environments (and in our preliminary trials), we were excited to apply the most exciting insights in ML to healthcare diagnostics.

Dataset

We used data from the BUU (Burapha University) Spine Dataset. This dataset comprises 400 pairs of spine X-ray images, with each pair containing an anteroposterior (AP) and a lateral (LA) view, collected from 400 unique patients. The total dataset therefore contains 800 images. These images consist of plain film radiographs suitable for developing and evaluating spine-related diagnostic models using multimodal view inputs. Resolution and image size varied, which was another challenge that we had to overcome in this data situation.

The dataset is labeled with subject ID, gender, age, and spinal diagnostic. In terms of patient demographics, this dataset includes 127 male and 273 female subjects, with ages ranging from 6 to 89 years (with a mean age 50.21 years). he dataset contains JPG files for each image, a CSV associated with each image with additional information, and a .xlxv master file containing additional information on each subject. Within the 400 images, 76 subjects have some sort of spinal disorder. Those images flagged as disordered are further categorized as Anterolisthesis, Left Laterolisthesis, Retrolisthesis, or Right Laterolisthesis. The dataset features 57 samples with Anterolisthesis, 3 samples Left Laterolisthesis, 9 samples Retrolisthesis and 7 samples with Right Laterolisthesis.

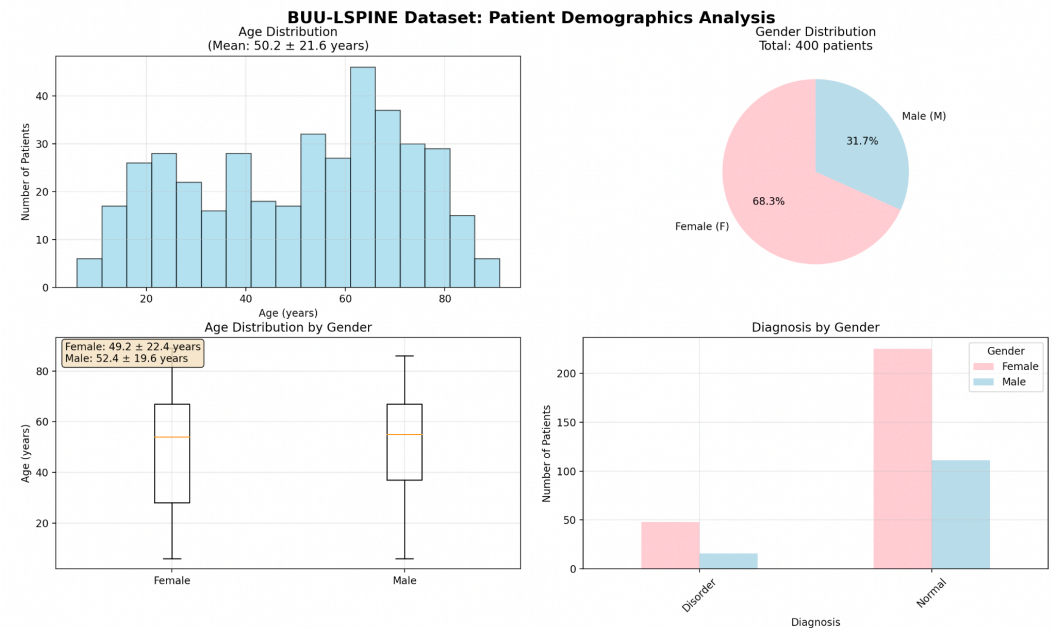


Dataset Size and Bias

Due to constraints on timing and issues with approval (a Stanford administrator no-showed a data approval call), we had to use this smaller dataset. In particular, our dataset lacked sufficient disordered samples. We experimented with undersampling, combined sampling, and other techniques. This was described further in the *Methods* and *Experiments* section. We used an 80:10:10 train:test:val split with 800 total images.

Our dataset is biased towards women and older individuals, reflecting the demographics of the patient population from which the data were collected.The age bias is relatively expected, as spinal disorders are more prevalent among older individuals, who typically constitute a greater proportion of patients seeking clinical care for spine-related issues. Consequently, our models might inherently capture age-related spinal characteristics, potentially limiting their accuracy when applied to younger patient populations. This may limit trained model's applicability to younger patients with abnormal spinal deformities.

The gender imbalance (the dataset features approximately 68% female subjects in the dataset) further introduces potential limitations. This skew could result in reduced generalizability of diagnostic models, possibly leading to diminished predictive performance or increased diagnostic uncertainty in male populations. Larger training runs of spinal diagnostic models should mitigate these biases by collecting data from more gender-balanced and age-diverse cohorts, which would improve robustness and ensure broader applicability.



Methods

Our methodology focused on experimentation and exploration for optimizing a model in situations with constrained and biased data. The most important experiments, across (1) Different pre-trained (or non-pre-trained) CNNs, (2) Single, Multi-View, and Hybrid CNN Architectures, and (3) Sampling to adjust for the data-constrained environment.

We decided to narrow to those three experiments from preliminary results. We ran many dozens of mini-experiments (from comparison of Transformers and CNNs to data-based segmentation based on lateral and anterial-posterior viewpoints in spine data). Interesting preliminary results prompted us to delve further, which culminated in the following three architectural designs:

Section I: Architectures.

We analyzed comparing different CNN architectures. We compared four main architectures: ResNet18, MobileNetV2, DenseNet121, and a non-pretrained CNN (which we called SimpleCNN). Based on preliminary trials, we knew that overfitting was a major issue, so we used models that we believed would best be able to reduce that risk. We used pre-trained backbones for transfer learning for every model except for the non-pretrained CNN.

The major consistency across the different architectures was replacement of classification heads to the binary prediction relevant to our models via final layer modification. This included dropout optimized to 0.4 after hyperparameter tuning for regularization to reduce overfitting and a final linear layer for binary prediction.

Section II: Multi-View CNNs

We analyzed a multi-view CNN (“DualCNN”) architecture in comparison to a standard CNN and a hybrid model (“HybridCNN”). Specifically, this was drawn from how surgeons and radiologists often use both the frontal and lateral spinal images to diagnose spinal deformity. If we wouldn’t expect professionals to diagnose with just one data point, how could we expect an AI model to do so? After all, some images may not show the curvature that another image might.

The Standard CNN was a traditional single-image and single-view approach using individual pre-trained ResNet18 models that separately evaluated front and side views in an ensemble approach.

The DualCNN was a multi-view approach, inspired by and extending beyond Wu et al. 's 2018 work to include slightly modern techniques. This included multiple ResNet18 feature extractors for each of front and side views. Afterwards, they were followed by attention-based and concatenation fusion.

Finally, the HybridCNN approach was a more adaptive architecture, meant to serve almost as a “middle ground” between the above two. This used availability masks and fusion strategies to attempt to also add extra redundancy to protect against unpaired images (which was developed in response to a data cleaning/wrangling glitch, which was later resolved).

Fusion approaches used included concatenation, which was direct feature vector combinations, attention-based fusion which were learned weighting to emphasize more informative features, and ensemble fusion that combined separate predictions through an average.

Section III: Sampling and Other Data Techniques,

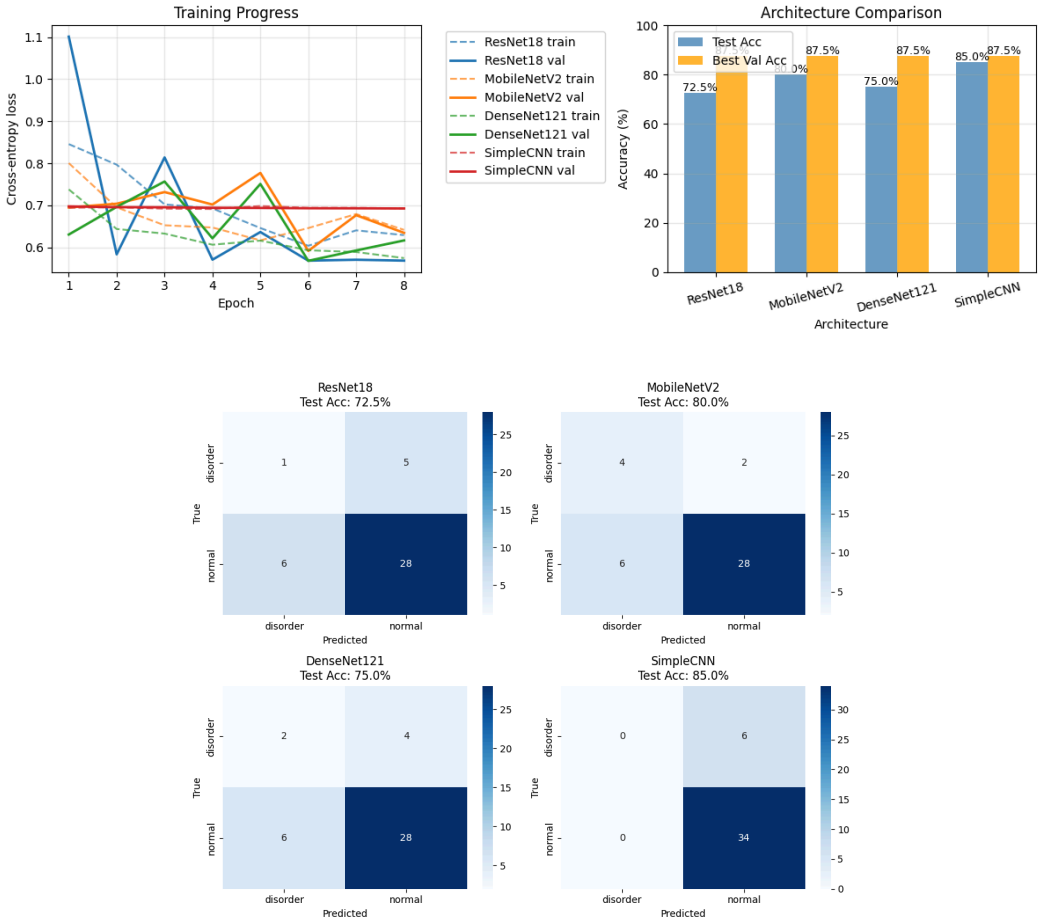
In response to the class imbalance, we also tested many different sampling techniques. These are often used in both medical and non-medical imaging tasks to account for low-data environments. Often, some of these techniques can help prevent overfitting, which was a major problem from our low-data imaging problem.

Using the best architectures from *Section I* and *Section II* (ResNet18 in a Standard CNN approach), we evaluated undersampling and combined sampling to try to overcome our data issues.

Experiments

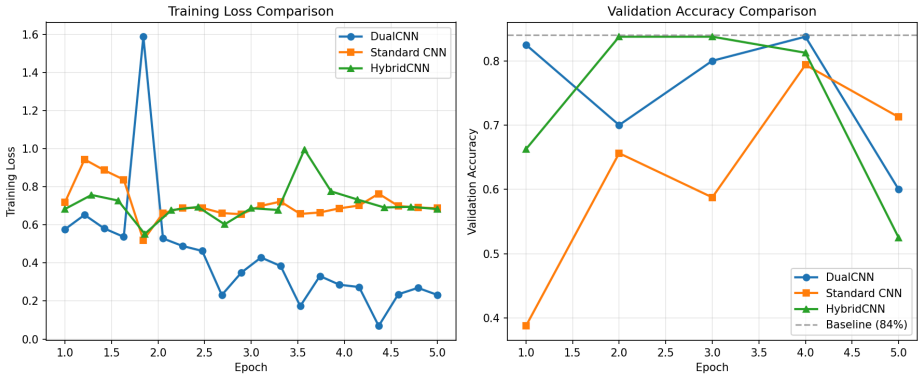
After countless rounds of testing, we landed on a 1e-4 learning rate and the Adam optimizer as ideal on this dataset. Our primary metrics were accuracy, loss, and also manually ensuring overfitting didn't occur. One issue was that the model often simply guessed "normal" due to the data bias of having most of the data being "normal".

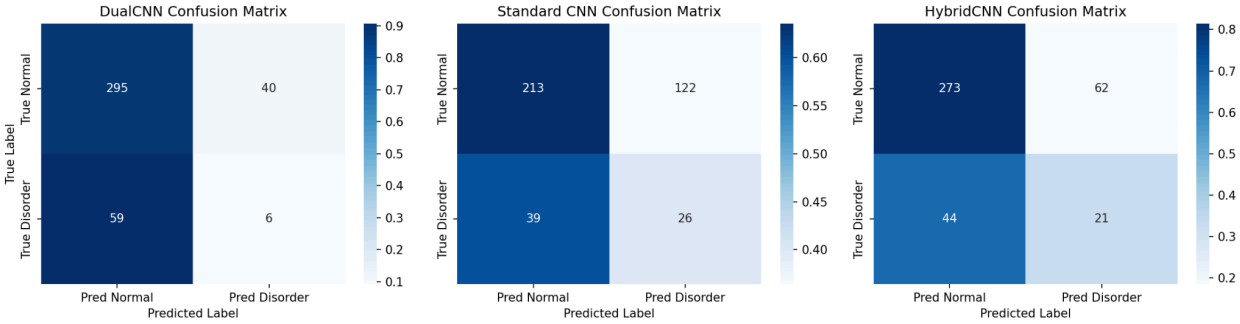
Section I: Architecture.

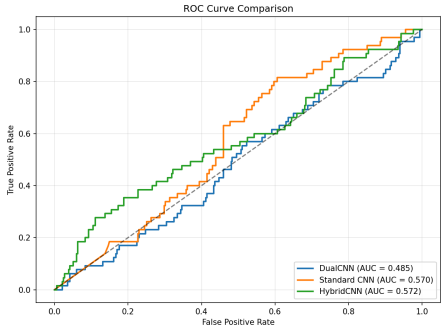


These results show the comparison across model architectures. As you can see, all models ultimately outperformed our simple CNN, which categorized all results as normal, on loss, but ended up overfitting and failing at test time. While Simple-CNNs policy did outperform, simply categorizing all models as normal is a very naive strategy, so we attempted more sophisticated sampling processes later to mitigate this. The best result was by ResNet18, which was surprising as some of the other architectures had less or more complexity. It was interesting that there was an almost "goldilocks" phenomenon without a clear connection between complexity and accuracy.

Section II: Multi-View CNNs.

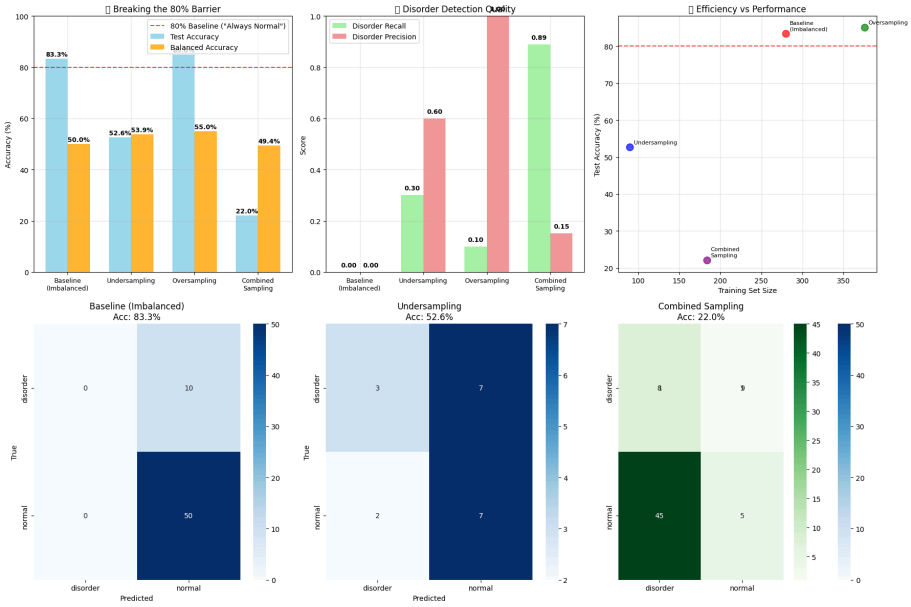




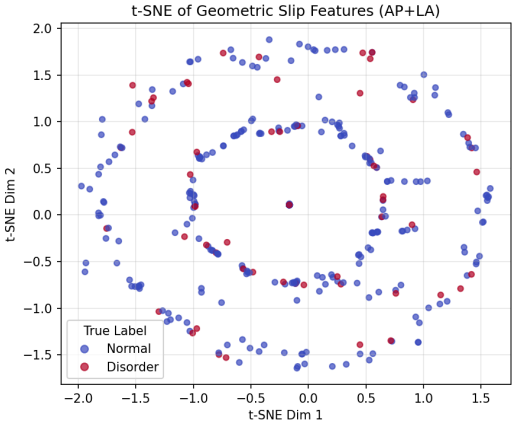


Now, we evaluate the Multi-View CNNs. Although we see that the DualCNN has the lowest loss value, it doesn't connect to a higher accuracy value. This seems to imply an overfit. We can similarly see overfitting across the StandardCNN and the HybridCNN, with none surpassing baseline.

Section III. Sampling and Other Data Techniques.



Here are our results testing different sampling methodologies. Interestingly enough, undersampling and combined sampling ultimately resulted in lower accuracy. Oversampling on the other hand resulted in a slight accuracy increase, suggesting that it may be a good fit for similar datasets.



Lastly, here are our results for data analysis. This t-SNE plot projects our geometric slip feature vectors from paired AP and lateral radiographs into two dimensions. As the points arrange into two loose concentric rings, revealing that slip metrics live on a curved low-dimensional manifold and samples overlap throughout, networks may struggle to cleanly separate healthy from pathological cases. The two t-SNE dimensions were based on the labels provided for different spinal curvature and regions, seemingly implying that the complexity of spinal diagnosis in many of these cases was beyond simply looking at spinal curvature. This would likely explain why off-the-shelf techniques fell short of truly diagnosing the spinal deformities.

Conclusion

Personalized healthcare AI is the future, and better understanding low- and biased- data environments is a crucial frontier to better advise physicians and patients. In our data, we analyzed a multitude of techniques across some of the top literature studies. We found that none of those advanced techniques provided exceptionally impressive results, as overfit on our limited dataset. Across four different CNN architectures: ResNet-18, MobileNetV2, DenseNet-121, and a simpler custom CNN, our simpler CNN achieved the highest accuracy at 87.5%. ResNet-18, MobileNetV2, DenseNet-121, all seemed to overfit to the data and achieved accuracies between 72% and 80%. This model overfitting was likely due to how more complex models are better able to memorize training data, which the simple CNN was less likely to do. In the multi-view approach, a similar conclusion was reached, despite literature supporting its prowess. Both the DualCNN and HybridCNN we tested seemed to perform slightly worse than the standard CNN. Finally, perhaps most interesting, the sampling techniques actually made our model perform worse than the baseline.

The future of low-data ML is just beginning, and personalized healthcare AI will rely on similar explorations of advanced techniques like we have done in this paper. Further studies could include analyzing other imaging types (MRI, CT scans, etc), using more recent techniques, or potentially even using advanced transformer systems trained in imaging data followed by personalization in the data-constrained environment. Further data wrangling would also be quite interesting, potentially focused on specific regions in images. For example, which parts of the spine should we focus on?

All in all, our results are just the beginning of an exciting journey in developing AI with constrained data. Beyond healthcare, we believe this is one of the most important frontiers to advance technology and democratize so many quality of life improvements for so many people. Data is a bottleneck, and we're fighting to change that!

REFERENCES:

[1] Martin, B. I., Tosteson, A. N. A., Lurie, J. D., Goodman, D. C., Schoenfeld, A. J., Brophy, R., Weinstein, J. N., & Wennberg, J. E. (2014, October 28). Back pain in the United States. In Variation in the care of surgical conditions: Spinal stenosis: A Dartmouth Atlas of Health Care Series. The Dartmouth Institute for Health Policy and Clinical Practice. https://www.ncbi.nlm.nih.gov/books/NBK586768/

[2] ICA.ai. (2025, February 5). Personalized medicine meets AI: Unlocking the power of big data. ICA Insights. https://www.ica.ai/insights/personalized-medicine-meets-ai-unlocking-the-power-of-big-data/

[3] Wu, H., Bailey, C., Rasoulinejad, P., & Li, S. (2018). Automated comprehensive Adolescent Idiopathic Scoliosis assessment using MVC-Net. Medical Image Analysis, 48, 1-11. DOI: 10.1016/j.media.2018.05.005

[4] R. Aggarwal et al., "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," npj Digit. Med., vol. 4, Art. 65, Apr. 2021

[5] Zhao, M., Meng, N., Cheung, J.P.Y., et al. (2023). SpineHRformer: A Transformer-Based Deep Learning Model for Automatic Spine Deformity Assessment with Prospective Validation. Bioengineering, 10(11), 1333. DOI: 10.3390/bioengineering10111333

[6] Kim, H. J., Kim, S. S., Wu, H. G., Lee, H. M., Kim, H. S., Khang, G., & Moon, S. H. (2023). Multi-pose-based convolutional neural network model for diagnosis of patients with central lumbar spinal stenosis. Scientific Reports, 13, 24658. https://doi.org/10.1038/s41598-023-50885-9

[7] Lee, G. W., Shin, H., & Cho, M. C. (2022). Deep learning algorithm to evaluate cervical spondylotic myelopathy using lateral cervical spine radiograph. BMC Neurology, 22, 195. https://doi.org/10.1186/s12883-022-02670-w

[8] Zheng, H., Wang, R., Yu, L., Luo, S., Li, J., Xu, S., & Guo, C. (2022). Image quality control in lumbar spine radiography using enhanced U-Net neural networks. Frontiers in Public Health, 10, 891766. https://doi.org/10.3389/fpubh.2022.891766

[9] Dhillon, A., & Verma, G. K. (2021). Convolutional neural networks in medical image understanding: a survey. Evolutionary Intelligence, 14(2), 1-22. https://doi.org/10.1007/s12065-020-00540-3

[10] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into Imaging, 9(4), 611-629. https://doi.org/10.1007/s13244-018-0639-9

[11] Li, H., Zhang, L., Qin, C., Yap, M. H., Rajpoot, N., & Li, C. (2024). VerFormer: Vertebrae-aware transformer for automatic spine segmentation from CT images. Scientific Reports, 14, 20599. https://doi.org/10.1038/s41598-024-71050-3

[12] Lyu, J., Bi, S., Ling, S. H., Banerjee, S., & Su, S. (2025). SSAT-Swin: Deep learning-based spinal ultrasound feature segmentation for scoliosis using self-supervised Swin transformer. Ultrasound in Medicine and Biology, 51(3), 102748. https://doi.org/10.1016/j.ultrasmedbio.2025.01.002

LIBRARIES:

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2 (nature.com)

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 56-61). https://doi.org/10.25080/Majora-92bf1922-00a (scirp.org)

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55 (scirp.org)

Clark, A., & Pillow Contributors. (2025). Pillow (Version 11.1.0) [Computer software]. https://pillow.readthedocs.io/ (pypi.org)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html (jmlr.org)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32 (pp. 8024-8035). https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (papers.nips.cc)

TorchVision Maintainers & Contributors. (2016). TorchVision: PyTorch's computer vision library (Version 0.22.1) [Computer software]. https://github.com/pytorch/vision (pypi.org)

Waskom, M. L. (2021). seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021 (joss.theoj.org)

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1-5. https://jmlr.org/papers/v18/16-365.html (jmlr.org)