Towards Data-Centric Approaches to Lung Cancer Classification

Mark Movh and Isah A. Lawal(✉)

Noroff University College, Kristiansand, Norway
isah.lawal@noroff.no

Abstract. There is an ever-growing need to review Artificial Intelligence and its corresponding implementation methodology in medical image analysis. The discussion of optimizing code versus improving data is of prime importance when maximizing model performance in medical image classification. Recently, a majority of studies have been model-centric. It is crucial to investigate data-centric methodologies and how medical image quality impacts a model's learning capabilities. This study opts toward data-related modifications for model improvement in lung cancer classification, acting as a proof of concept for developing data-centric AI. The proposed data-centric approach (DCA) modifies CT-scan images of the lung through 3 stages; image preprocessing, image segmentation, and feature extraction. The modified images were used to train a simple Convolutional Neural Network (CNN) for the classification task. We evaluate the performance of the proposed method using a publicly available real-world dataset of lung CT scans. Our method achieves a classification score (F1 score) of up to 0.889. This performance is superior to that reported using a model-centric approach on the same dataset, which conducted automatic hyperparameter optimization using the random search algorithm.

Keywords: Lung Cancer Classification · Data-Centric AI · Medical Image Analysis

1 Introduction

Lung cancer is the most prominent and deadly variant of cancer, indicating the need for accurate identification and diagnosis. Computed Tomography (CT) scans are a method for capturing images of the lungs, to identify clumps of abnormal cells. CT is considered the most common, due to its accuracy and benefit of perspective without having features overlap one another [3]. However, even with the technological development of these machines, human limitations are often the cause of inefficient information evaluation. Manual reading, understanding, and overall analysis of the scans could become unreliable without an experienced radiologist [4]. Therefore, to combat these issues, the process of automating detection and diagnosis in healthcare has been researched and developed through Machine Learning.

S. Kadry and R. Prasath (Eds.): MIKE 2023, LNAI 13924, pp. 54–66, 2023.
https://doi.org/10.1007/978-3-031-44084-7_6

Machine Learning (ML) is an area within Artificial Intelligence that revolves around computers and how they can learn in order to process data, find more complex patterns, and present information collected from these patterns. The process mainly consists of training a model on real-world examples, such that it is capable of learning their distinctive characteristics. Three major ML types are supervised, semi-supervised, and unsupervised learning [9]. Each category has various problem-solving capabilities, however this study will primarily focus on supervised ML as it focuses on lung cancer classification. In supervised ML there are features (inputs) and labels (outputs); the goal of the algorithm is to learn how it can map these features to their respective labels [2]. Once an ML model has been created and trained, it can be further improved through specific optimization. However, these optimizations may come in many forms based on the two important components that build up the ML model; the data, which would be some type of input (images etc.), and the code, that is, the algorithm that undergoes the learning described above [6].

Recently, there has been a large discussion as to which part should be optimized for a more accurate model; the data, or the code. Andrew Ng [7] popularized this discussion through his presentation of model-centric and data-centric AI concepts, bringing to light an important consideration as ML applications continue growing. A model-centric approach refrains from adjusting the data while continuously optimizing a model's structure by tuning the model parameters and hyperparameters to maximize performance (see Fig. 1) [6]. A data-centric approach, on the other hand, is concerned with keeping the model structure fixed while improving the quality of data through pre-processing (see Fig. 2) [8].

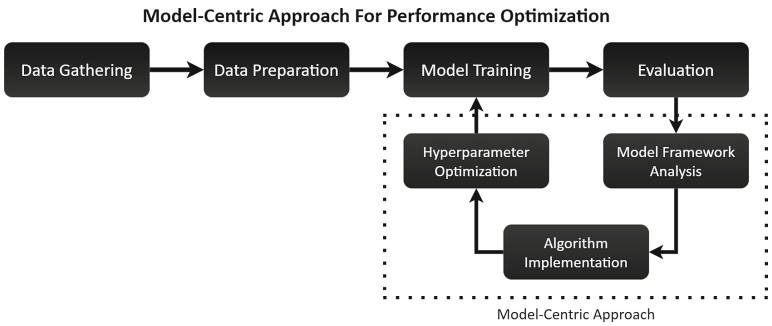

Fig. 1. Model-centric approach focused on model optimization. The dotted region encapsulates the overarching stages of this optimization for performance improvement.

The problem with the model-centric approach for medical image classification is that the model sometimes becomes too complex due to excessive parameter tuning to fit the raw medical images. Thus potentially leading to a model with poor classification accuracy that is undesirable for medical applications. Additionally, a lack of available data in large amounts and a lack of refined solutions

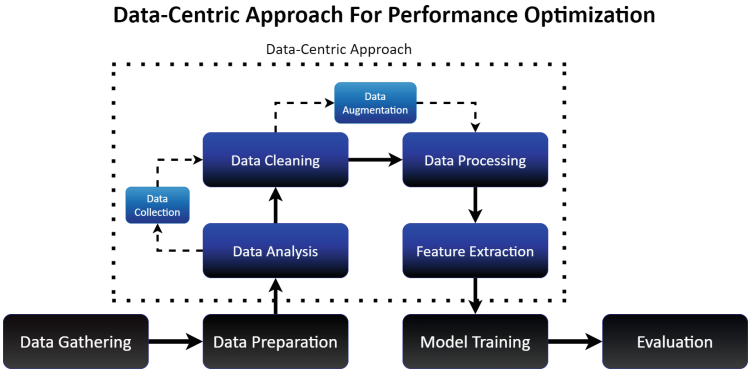

Fig. 2. Data-centric approach focused on data enhancement. The dotted region encapsulates the primary stages to achieve higher-quality data for performance improvement.

within the medical sector should pave the way for an increase in data-centric methodologies [6,12].

The main idea of this study is to improve the quality of medical images and data preparation stages through a data-centric approach, thus achieving a more accurate lung cancer classification model. We improve the performance of a simple CNN model implementation through modification of image pre-processing, image segmentation, and feature extraction stages. Firstly, we extract prominent information through data analysis. Secondly, we undergo pre-processing stages to transform the data into more suitable input. Finally, determining the most informative features for final input such that even a simple machine learning model can easily and accurately classify them without the rigor of parameter tuning by an expert. It is acknowledged that this paper presents simple but effective image modification processes for improving lung cancer classification. Moreover, this paper serves as a proof of concept for developing data-centric AI in lung cancer classification and the importance of improving data quality. It shows the efficiency of employing simple data processing implementations for performance gain when measured against a base model with a simple architecture. The goal is to demonstrate the difference in results when conducting a data-centric approach.

2 Related Work

There has been a growing interest in research on the concept of data-centric AI since the term became popularized. Still there exists a large knowledge gap within the field due to a lack of available studies, especially in the medical sector.

Research by Angela Zhang et al. [13] in 2022 expressed the importance and limitations of transforming model practices into data practices in healthcare. The paper outlines several concepts that could prove useful in the future, placing heavy emphasis on dataset creation and improvement, along with model

optimizations. While they do praise data-centric AI, they rightly point out the limitations. Large, quality and standard datasets are hard to gain access to, much less create. In the medical sector, this may be especially difficult, as data is not so easily available, either due to the privacy of individuals or the differences in data structures within institutions.

Overall, there is a lack of available literature on data-centric approaches in medical image applications, mainly in lung cancer classification. Even so, despite the term being newly formed, the concept of enhancing data and images has been heavily researched. Chaturvedi et al. [3] presented the intricate nature of image modification and how it can improve the performance of models. This review of current research highlights a variety of available and refined techniques. Three primary stages are outlined for a data-centric approach in medical image applications: image preprocessing, image segmentation, and feature extraction.

Studies such as the one conducted by Vas and Dessai [10] in 2017 on small medical images, go into further detail about the effectiveness of these stages. Vas and Dessai first cropped the images to reduce unnecessary parts, then applied 3×3 median filters to remove impulse noise. Moreover, the images were segmented through morphological operations such that only the Region of Interest (ROI), the lungs in this case, are kept. For feature extraction, the Gray-Level Co-Occurrence Matrix (GLCM) was used, scanning for Haralick features. The artificial neural network was chosen for the classification algorithm for the classification task.

Vero and Srinivasan [11] more recently in 2020, also conducted such an approach by applying image pre-processing, segmentation, feature extraction, and furthermore feature selection. First, Histogram Equalization was used to make image intensity differences clearer, followed by an Adaptive Bilateral Filter to remove any noise present. For image segmentation, the nodules were segmented using the Artificial Bee Colony method. Then various techniques were tested to locate the nodules within an image. Through ROI feature extraction, the following features were considered: volumetric, texture, intensity, and geometric.

Similarly to the [11] method, our proposed method (DCA) employs basic but rigorous data-related stages allowing for a comprehensive comparison of techniques, and providing insight into the potential of data-centric approaches in lung cancer classification and other healthcare applications.

3 Methodology

3.1 Data Processing

For the lung cancer classification DCA modeling, we used the Lung Nodule Analysis (LUNA16) dataset described in Sect. 4.1. Figure 3 shows samples of some of the raw CT scan slices and the following discussion explains the data processing techniques that were used to improve the overall quality of the image and in preparing them as input to a simple convolutional neural network.

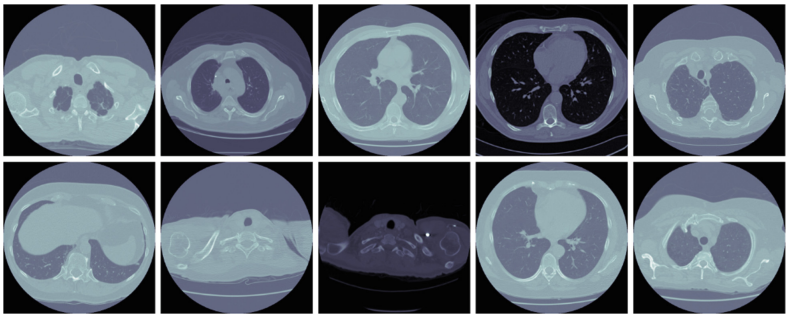

Fig. 3. 10 randomly selected samples from the LUNA16 dataset. Each sample is a single slice from a computer tomography scan.

Firstly, simple filtering techniques, such as the median filter, were used to reduce any potential noise in the slices. The prime reason for using median filters is that they are better at detail conservation, such as edges. The core implementation of this is that it selects the median value for a pixel, depending on its neighbors. The filter size determines the number of these neighbors, and a 5×5 filter was selected for our study. The transformation of one of the raw images can be seen in Fig. 4. For a larger testing environment, the Mean and Gaussian filters were later employed to measure the effectiveness of implementing other basic filters for performance changes.

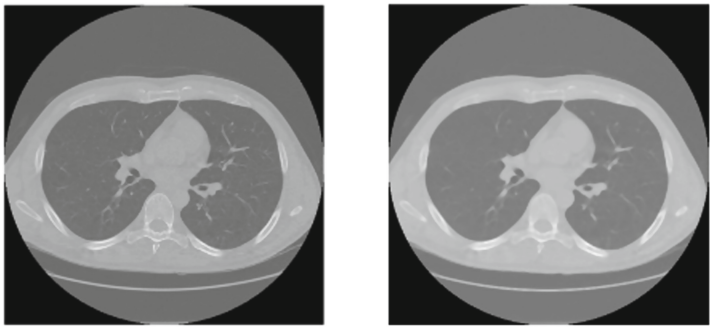

Fig. 4. Figure showing how the 5×5 median filter modifies the original raw image (left) causing smoothing and blur (right).

Secondly, we performed image segmentation. Regions of interest are first identified, then followed by morphological operations. Every slice in every CT scan was individually segmented, going through 8 steps. Firstly, the image is transformed into black and white, followed by the second step, border refinement. The third step is to label the different regions of the image and then, in step 4, remove those deemed irrelevant. This is done by only keeping the largest two

areas. The next step conducts binary erosion to create a distinction between the lungs and blood vessels. Binary closing comes after, filling in black gaps in the regions. The 7th step ensures any black holes left over are filled, and finally in step 8, the binary mask created is superimposed on the original slice. The stages and their transformations are shown in Fig. 5.

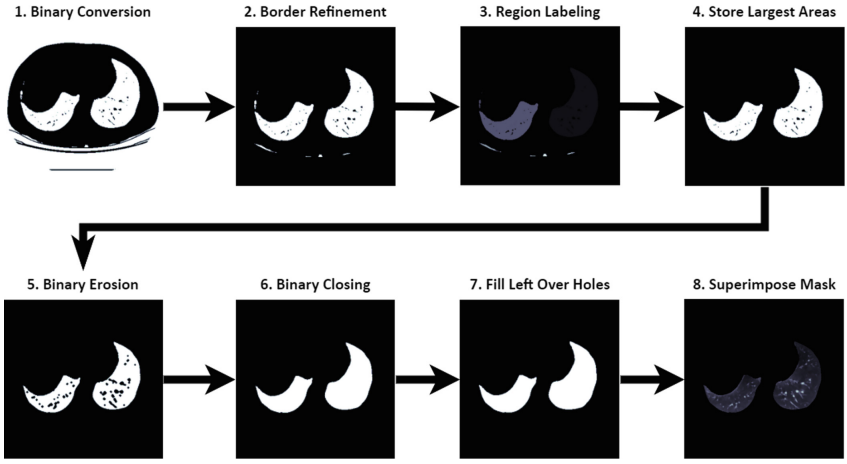

Fig. 5. Procedure of segmenting CT scan slices to prepare them for feature extraction.

Finally, the feature extraction stage. In the case of the dataset used, the images contained labeled nodule objects. The CT scans were stored as 3D arrays, with each one of the 3D arrays accompanied by real-world coordinates, allowing training and testing data generation. Thus in the feature extraction stage, the nodules within the 3D arrays would be extracted based on regions of interest. Meaning a nodule would be found given its respective coordinates, and then cubic voxels of size $36 \times 36 \times 36$ would be cut around these coordinates. These voxels would then have a corresponding label marked as 1 (nodule). Non-nodule voxels were generated by randomly picking coordinates and slicing the array at these coordinates. A sample of these is shown in Fig. 6. Not all of these cubes will be perfectly sliced. Nodules or randomly cut voxels could be taken on the border of the segmented lung, causing much of the black background to be cut along with it. While this randomness potentially results in some overlapping, it was considered reasonable to allow the model to generalize and avoid overfitting.

3.2 Classification Algorithm

The architecture of the CNN that was built for the proposed lung cancer classification is inspired by design in [1] except that for our DCA approach, relatively fewer convolutional layers are enough for distinguishing cancerous and non-cancerous lung scan images. We also modified the fully connected layer,

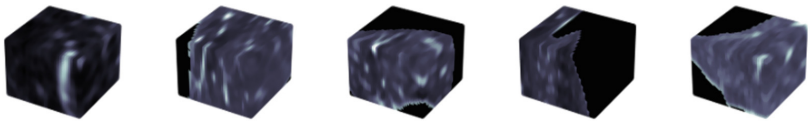

Fig. 6. 5 voxels cut from random segmented CT scan slices. The black values on the cube represent the area outside of the segmented lungs.

reducing the number of filters and adjusting the dropout layer. Lastly, the output activation was changed to sigmoid, and the loss was adjusted to binary cross-entropy. The point was to create a simple model that would rely on higher-quality data rather than trying to learn the complex features themselves. The simple CNN architecture is presented in Fig. 7.

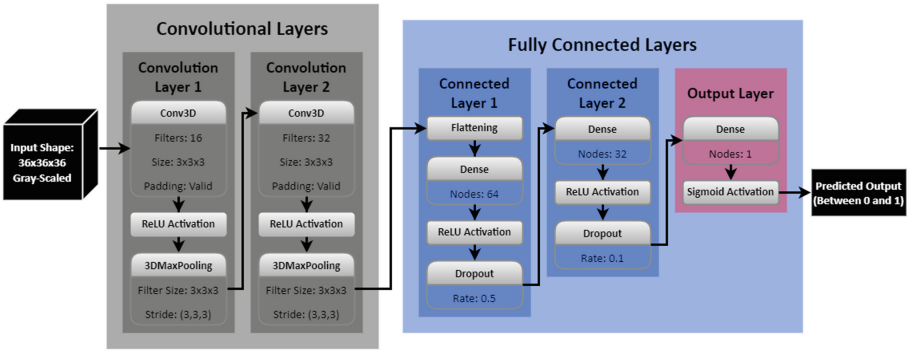

Fig. 7. Schematic of the CNN used for the classification of the processed CT scan images.

The CNN is created by putting the layers in sequential order. The left part of the diagram shows the convolutional layer, whereas the right shows the fully connected part of the CNN. The input layer accepts a 3D array as input of size $36 \times 36 \times 36$, noting that these are grayscaled. Following is a 3D convolution containing 16 filters of size $3 \times 3 \times 3$ with the ReLU activation function. The next layer contains the pooling layer, implemented through 3D max-pooling, with a kernel size of $3 \times 3 \times 3$, and the stride is the default. The second 3D convolution layer is similarly implemented, consisting of 32 $3 \times 3 \times 3$ filters and a ReLU activation function, followed by a 3D pooling layer with a $3 \times 3 \times 3$ kernel size. The results of this layer are then flattened, before going into the fully connected layer which is responsible for the classification of labels. The first Dense layer contains 64 nodes and a ReLU activation function. A dropout function is utilized, with a value of 0.5. The following Dense layer is very similar to the previous, except it contains 32 nodes and outputs into the Dropout layer set to 0.1. The final layer contains only one node and applies the sigmoid function for the binary

problem. Only one output is given, this being a predicted output between 0 and 1. Closer to 0 means higher certainty that it is not a nodule, whereas values closer to 1 indicate higher certainty of a nodule.

4 Experiments and Discussion

4.1 Dataset

The LUng Nodule Analysis (LUNA16) dataset [5] was used to evaluate our proposed method. The open-access dataset contains CT scan images of patients' lungs, published by Zenodo as a challenge for lung cancer classification through ML algorithms. 486 data points are identified within these images and then extracted for model training and testing. The training and testing data is generated as discussed in the methodology. These 486 data points are divided approximately into an 80-20% split, having 388 data points in the training set and 98 data points in the testing set.

From every data point, 96 voxels (48 positively labeled for nodule, 48 negatively labeled for non-nodule) are extracted, resulting in 37248 training entries and 9408 testing entries. The model is trained through batches, going through all 388 data points and training the model on the 96 voxels present in each batch. Predictions are generated in batches, taking one data point at a time, and are prepared accordingly such that they are comparable to the testing set. We assessed the performance of the DCA approach on the testing set using the following metrics: accuracy, precision, sensitivity, specificity, and F1 score. The experiment is repeated 5 times to gain an average for each metric.

4.2 Experimental Setup

The CNN model presented in Sect. 3.2 was developed using Tensorflow in Python version 3.9.13. The training and testing of the model were done on a PC containing the following components: Ryzen 7 3700X (3.6 GHz) CPU, ASUS ROG Strix 2060 GPU, and 32 GB (3200 MHz) RAM. Several experiments were conducted however, the model's parameters remained fixed throughout these experiments. The model is trained, and its predictions are evaluated against the processed data discussed in Sect. 3.1. The model is compiled with binary cross-entropy for the loss function and the adam optimizer with a default learning rate (0.001).

4.3 Results and Discussion

Once the model had been built and compiled, the evaluations began using the test set. A total of 6 different evaluations were conducted, and their results were collected. The first evaluation involved using only the raw, unmodified CT scan data. The second used only median filters to transform the data. The third involved using only fully segmented CT scan images. The fourth, fifth, and sixth evaluations involved using the proposed data-centric approach (DCA)

with Median, Mean, and Gaussian filters, respectively. The various evaluations were conducted to view how the lung cancer classification performance would be impacted depending on which data processing techniques were implemented.

Table 1 summarizes the result of the evaluations. As seen in the results, using only the unmodified data produces classification scores of at most 81% in almost all scoring metrics, with each additional data modification process improving the results. The proposed DCA with the Median filter achieves the highest classification score (F1 score) of 88.9%, confirming that modifying the data quality is a valid approach to improving a model's performance. Interestingly, using only correctly segmented CT scan images scored almost as high as the full data-centric approach on the test set, with sensitivity being higher at 92%. From this, it can be said that proper image segmentation can increase classification performance for nodules.

Table 1. Results of raw data, application of single processing methods, and three variants of the proposed data-centric approach, utilizing various filters. IS - Image Segmentation, DCA - Data-Centric Approach (Filter Applied).

	Raw Data	Filter Only	IS Only	DCA (Median)	DC (Mean)	DCA (Gaussian)
Accuracy	81.0%	84.0%	87.0%	88.8%	85.6%	85.6%
Precision	81.0%	85.4%	84.6%	89.4%	87.4%	85.6%
Sensitivity	80.0%	85.2%	92.0%	88.4%	87.4%	85.0%
Specificity	81.0%	83.0%	81.8%	89.4%	83.8%	87.0%
F1-Score	81.0%	84.0%	87.4%	88.9%	85.2%	86.2%

The DCA approach showed a decrease in sensitivity over applying image segmentation on the lungs. However, this decrease is resultant of the model having more stable and robust learning since the full DCA approach remained unbiased as sensitivity and specificity were similar. Additionally, with the DCA (median) approach, the results of all the scoring metrics are very close to each other, which means the model is more balanced, predicting true positives as well as true negatives equally. This is highly desirable for lung cancer classification using CT scan images as in healthcare, it is especially important to attain as high performance as possible since people's lives are considered.

Considering the results and advantages of the proposed data-centric approach, there was further investigation taken into the classification results of the best-performing approach, this being the DCA (Median) implementation. A confusion matrix was constructed, which is shown in Fig. 8. The confusion matrix of the model's predictions further confirms that the model is balanced in terms of predictions and is not biased towards one class. However, with a large number of false positives and false negatives, it would not be sufficient to employ this solution in real-life systems. Nevertheless, the improvement of quality in data allowed for an increase in performance across all measured metrics.

		Actual Values	
		Positive Nodule 1	Negative Non-Nodule 0
Predicted Values	Positive Nodule 1	88.4%	10.6%
	Negative Non-Nodule 0	11.6%	89.4%

Fig. 8. Confusion Matrix of the DCA (Median) implementation, showing the percentages of true positives (top left), false positives (top right), false negatives (bottom left), and true negatives (bottom right).

This study favored a data-centric approach in medical image analysis due to the general problem of lack of the availability of quality data. Contrarily, model-centric approaches were also tested to confirm the effectiveness of the proposed approach further. The raw data was kept fixed while the CNN model discussed in Sect. 3.2 was iteratively improved. Various testing was conducted, ranging from manual hyper-parameter optimization to automatic hyper-parameter tuning. From dozens of models generated, the best scoring (average across all metrics) was selected for comparison as shown in Fig. 9. For the model-centric implementation, the number of layers remains the same. However, the optimizer is changed to Ftrl (Follow the regularized leader) as it presented more stabilized results. Moreover, each layer was automatically hyper-tuned through the use of the random search algorithm. The first convolutional layer had a range of 16 to 64 filters, whereas the second convolutional layer had a range of 32 filters to 128. Each of these layers had a step of 8 in these ranges and tried filter sizes of 3 and 5. The first fully connected layer had a range of 128 to 256 nodes, and the second fully connected layer had a range of 64 to 128 nodes. Each fully connected layer had a step of 16 in these ranges.

As shown, the model-centric approach scored lower than the proposed data-centric approach, except in sensitivity, where it scored similarly. It should be noted that the model-centric results were obtained after an exhaustive hyperparameter search that was time-consuming and computationally complex. Whereas the data-centric approach was limited in its applied techniques, a variety of image modifications can still be applied for an improved classification rate in the scope of image enhancement applications in healthcare. This additional evaluation demonstrates the importance of improving data quality and its impact on performance that can still be built upon. Overall, the discussion shows that by minimizing the number of uninformative artifacts in an image, the model could focus better on identifying the prominent features present and improving the classification rate.

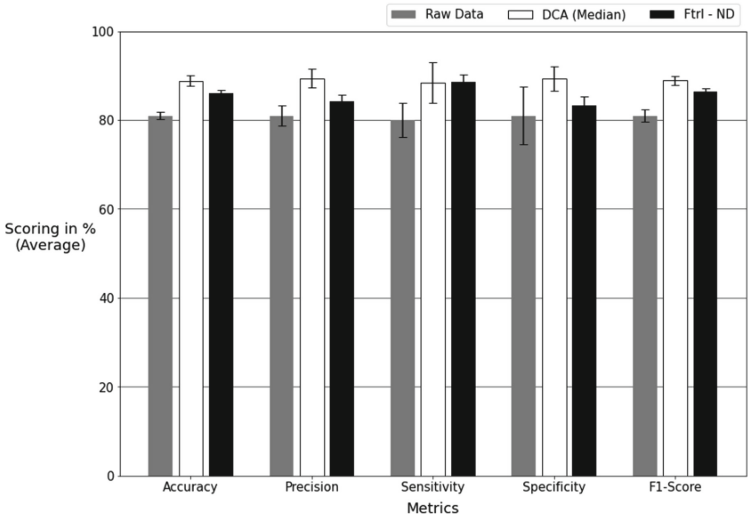

Fig. 9. Performance of model-centric approach (MCA) in black vs. proposed DCA in white. The gray bars are the results of the baseline model trained with the unmodified data only. Error bars measure standard deviation.

As a comparison of the proposed DCA to similar studies, Ge Zhang, Lin, and Wang in 2021 [14] conducted their 3D CNN implementation for lung nodule detection on CT scans in the LUNA16 dataset. Similar to our approach, the paper underwent data preprocessing where 3D nodule patches were segmented and extracted, with the additional stage of labeling the malignancy suspiciousness of said patch. Through the application of these preprocessing stages for data augmentation and a 3D CNN based on DenseNet architecture for classification, the implementation by Zhang, Lin, and Wang achieved an accuracy score of 92.4%, a sensitivity score of 87.0%, and a specificity score of 96.0%. Despite not having an identical experimental setup, our proposed data-centric approach scored similarly while having a much smaller and less complex CNN architecture. Complex architectures, such as the compared method, are often less favorable as they become increasingly tedious to optimize and train, making them less comprehensible to medical experts. In healthcare, it is important for an implementation to match or complement the opinion of a human expert. Our data modification methodology conducting simple changes allowed us to improve the quality of data, avoiding this reliance on complex models to learn difficult features and reduce interpretability. Our solution, using a simple 5-layered CNN trained on the preprocessed data, provided results that are balanced between the identification of nodules and non-nodules, having similar sensitivity and specificity scores.

Our results stand as a proof-of-concept of the potential of building data-centric AI. By exploring the area of data modification, data-centric approaches

can be refined, and the quality of data can be improved until it reaches sufficient performance for real-life applications in healthcare.

5 Conclusions

This study demonstrated the benefits of modifying CT scan images to achieve higher-quality input for lung cancer classification. Due to the extensive nature of these data modification stages, the changes improved performance in all measured metrics. The study contributes to the concepts of data-centric AI by extensively reviewing available methodologies and presenting a simple, but effective data-centric approach for computational intelligence in healthcare. This approach acts as a proof-of-concept that employing simple implementations of image preprocessing, image segmentation, and feature extraction, allow for attaining the most important features of CT scans and inputting those into a simple CNN model to achieve good classification results. Exploring concepts of model hyperparameter optimization (i.e., model-centric AI) also indicated that achieving higher results for such a problem was more effective with our proposed data-centric solutions.

While this study promotes the concept of data-centric AI, and its importance in achieving high performance, we noted that our study was limited to the exploration of a single dataset, experimenting with few data processing techniques for lung cancer classification. Nevertheless, the results demonstrated that data-centric AI is worth investing time in to gain informative insights about medical images and how their quality will impact a model's learning capabilities. Future work would explore other image applications in healthcare, particularly where data is often limited.

References

1. ArnavJain: Candidate generation and luna16 preprocessing. Kaggle (2017). https://www.kaggle.com/code/arnavkj95/candidate-generation-and-luna16-preprocessing#Reading-a-CT-Scan. Accessed 20 Nov 2022
2. Brownlee, J.: Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, pp. 11–12. Machine Learning Mastery (2016)
3. Chaturvedi, P., Jhamb, A., Vanani, M., Nemade, V.: Prediction and classification of lung cancer using machine learning techniques. In: IOP Conference Series: Materials Science and Engineering, vol. 1099, no. 1, p. 012059 (2021). https://doi.org/10.1088/1757-899x/1099/1/012059
4. Gao, J., Jiang, Q., Zhou, B., Chen, D.: Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: an overview. Math. Biosci. Eng. **16**(6), 6536–6561 (2019). https://doi.org/10.3934/mbe.2019326
5. van Ginneken, B., Jacobs, C.: Luna16 part 1/2. Zenodo (2016). https://doi.org/10.5281/zenodo.3723295. Accessed 20 Nov 2022
6. Hamid, O.H.: From model-centric to data-centric AI: a paradigm shift or rather a complementary approach? In: 8th IEEE International Conference on Information Technology Trends (ITT), pp. 196–199 (2022). https://doi.org/10.1109/ITT56123.2022.9863935

7. Ng, A.: A chat with Andrew on MLOps: from model-centric to data-centric AI (2021). https://www.youtube.com/watch?v=06-AZXmwHjo&t. Accessed 20 Nov 2022
8. Polyzotis, N., Zaharia, M.: What can data-centric AI learn from data and ml engineering? In: 35th NeurIPS Conference on Neural Information Processing Systems, pp. 1–2 (2021). https://doi.org/10.48550/ARXIV.2112.06439
9. Rajkomar, A., Dean, J., Kohane, I.: Machine learning in medicine. N. Engl. J. Med. **380**(14), 1347–1358 (2019). https://doi.org/10.1056/NEJMra1814259
10. Vas, M., Dessai, A.: Lung cancer detection system using lung CT image processing. In: 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1–5 (2017). https://doi.org/10.1109/ICCUBEA.2017.8463851
11. Vero, A., Srinivasan, A.: Deep learning for lung cancer detection and classification. Multimed. Tools Appl. **79**, 7731–7762 (2020). https://doi.org/10.1007/s11042-019-08394-3
12. Whang, S.E., Roh, Y., Song, H., Lee, J.G.: Data collection and quality challenges in deep learning: a data-centric AI perspective. arXiv preprint arXiv:2112.06409 (2021). https://doi.org/10.48550/ARXIV.2112.06409
13. Zhang, A., Xing, L., Zou, J., Wu, J.C.: Shifting machine learning for healthcare from development to deployment and from models to data. Nat. Biomed. Eng. **149**, 104420 (2022). https://doi.org/10.1038/s41551-022-00898-y
14. Zhang, G., Lin, L., Wang, J.: Lung nodule classification in CT images using 3D densenet. J. Phys: Conf. Ser. **1827**(1), 012155 (2021). https://doi.org/10.1088/1742-6596/1827/1/012155