Automatic Detection of Spine Region Using Multiple Pseudo 3D U-Net Models with Weighted Average Voting and Attention Mechanisms

Kai Yang [1,*] and Masayuki Kikuchi [1,2]

[1] Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Tokyo, Japan
[2] School of Computer Science, Tokyo University of Technology, Tokyo, Japan
Email: yk1355392481@outlook.com (K.Y.); kikuchi@stf.teu.ac.jp (M.K.)
*Corresponding author

***Abstract*—The field of CT imaging has been witnessing significant advancements. However, extracting precise information from complex image data remains a challenging task. This study focuses on automating the extraction of CT images. In our study, we adopt the U-Net architecture, a multi-scale blurring technique on data, to obtain a multi-resolution representation. This method is specifically designed to capture information at various granularities, from more detailed information to broader structures. After applying this multi-step blur, we calculate the difference between adjacent images to take advantage of the change in situation between different resolutions. Although feeding the blurred results Directly into the U-Net model may yield satisfactory results, our approach to computing differences between blurred images focuses on the nuances of these changes. To further enhance accuracy, we focused on ensemble learning, leveraging the weights from the training processes of multiple models to average their output during prediction. The results demonstrated that by adopting our approach, we achieved a Dice accuracy of 96.8% and improved the accuracy of CT image extraction.**

***Keywords*—convolutional neural network, U-Net, medical image processing, spine segmentation, ensemble learning, attention gate**

I. Introduction

The modern healthcare ecosystem relies on medical imaging technology to alleviate the burdens on physicians while enabling access to more accurate and precise diagnostic information. One area of significant concern and challenge is the imaging of the spinal structure, which is central to bodily functions such as movement, support, and protection. The diagnosis of spinal diseases often requires a meticulous examination of cross-sectional images generated through CT scans. However, the manual labeling of slices, a task primarily carried out by radiologic technologists, is labor-intensive and time-consuming. Furthermore, the intricate anatomy of the spinal structure amplifies the complexity of image extraction. Thus, there's a growing need for an automated system capable of adept image extraction to assist the treating physician, reducing the workload and potentially minimizing human error.

Considering the above, our study embarks on a journey into supervised learning, aiming to automate CT image extraction. We employed Convolutional Neural Networks (CNNs) tailored for medical image processing, building upon the foundational work of Shigeta *et al.* [1]. Unlike Shigeta *et al.*, who utilized a U-Net-based CNN learning model to achieve notable accuracies on test data, our innovative approach incorporates an ensemble learning technique. The model proposed by Shigeta *et al.* [1] exploits a stack of CT spine slice data to learn from the x, y, and z-axis directions, simulating a pseudo three-dimensional learning and taking the difference between two neighboring slices for the spinal data according to formula. We further refined this approach by expanded the pseudo three-dimensional convolution learning methodology among voxels in a 3D volume by expanding each voxel coordinate from n × n × n filter and taking the difference between slices n pixels apart to improve the accuracy and efficiency of CT image extraction, ensuring a more effective integration of semi-global information. Furthermore, we utilized the weights from multiple CNN models during the training phase and averaged the outputs during the prediction phase to enhance the accuracy of image extraction.

This study hopes to bridge the existing research gaps, offering a significant leap towards automating the medical imaging process, and potentially contributing to faster, more accurate spinal disease diagnosis and treatment.

II. Related Studies

Ronneberger *et al.* [2] introduced a seminal innovation in the field of biomedical applications in 2015, the U-Net

Manuscript received January 17, 2024; revised February 1, 2024; accepted February 28, 2024; published May 8, 2024.

model. This model is typified by a distinctive U-shaped architecture, comprising two crucial components: an encoder for image convolution and a decoder for deconvolution, predicated on the principles of Fully Convolutional Networks (FCNs). What sets the U-Net model apart is its capacity to maintain information that is frequently lost in the convolution process. This is accomplished through a novel application of skip connections, which in turn enable a more precise and accurate extraction of the region of interest in biomedical imaging scenarios.

Huang *et al.* [3] showed a commendable result of 96% extraction accuracy of lumbar vertebrae using Ada Boost method. At the same time, Wang et al. By leveraging multi-atlas segmentation techniques, we managed an impressive accuracy of 92.7% in whole vertebra extraction [4].

Additionally, Korez *et al.* [5] published a report that revealed an extraction accuracy of 95.1% for healthy vertebrae (including the thoracic and lumbar regions of the spine). This was achieved by the Canny method, which combines certain tricks.

Vania *et al.* [6] conducted experiments utilizing Convolutional Neural Networks (CNN), incorporating redundant class labels. Their approach resulted in a superior extraction accuracy of 94.3% by the Dice coefficient for all spines, offering an excellent performance when compared to existing methods, such as the Level-set method. This result confirms the potential and efficiency of CNN-based approaches in medical image analysis.

In a preceding study, Kamata *et al.* [7] developed a model based on the original U-Net for the automatic extraction of tasks in spinal medical images. This model retained the U-shaped data flow of the original U-Net but underwent certain abridgments and parameter modifications, which has achieved an accuracy of 82.7% on untrained data.

In the model proposed by Shigeta *et al.* [1], the Convolutional Neural Network (CNN) utilizes two-channel image data, in which preprocessing incorporates volumetric shape change information into the grayscale single-channel CT image data. This process generates a change map derived from the CT value difference of corresponding pixels in upper and lower slices. The slice data is normalized to 512×512×512 for learning in the CNN model. The input and output are both two-channel in the U-shaped network, with the CNN image data stack being input for learning in batches of 512 slices per sample. The data output from each learning session is assessed for accuracy using the Dice coefficient, with reference to a labeled data stack that has undergone the same preprocessing as the learning data. The learning process of the CNN is carried out from each of three directions, which make outputs are integrated to generate a single spinal slice data stack (Fig. 1). This integration is achieved the average accuracy result data at 95.1%.

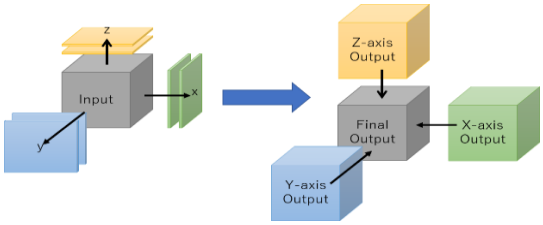

Fig. 1. Pseudo 3D feature learning.

III. Materials and Methods

Our research model integrates the output from a CNN model fed with slices from three axial directions to generate a single vertebral slice data stack. As part of the preprocessing of the slice data, we initially train on the data averaged within the 3×3×3, 5×5×5, and 7×7×7 ranges around the coordinates to identify the best performing model. Next, we use Shigeta *et al.*'s [1] pseudo 3D convolutional learning model for training process. This process has been proven to provide the best performance. These weights were acquired in three axial directions and subsequently reconstructed. We opted to utilize the weights derived from both these methods, thus facilitating pseudo 3D convolutional learning of global information (Fig. 2).

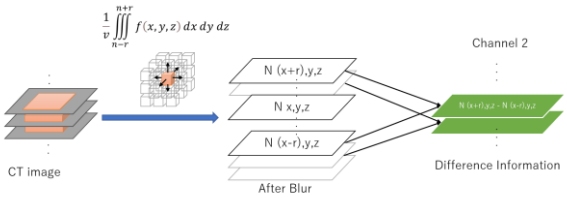

Fig. 2. The preprocessing of local averaging within r×r×r cube region.

CT images inherently exhibit three axes: *x*, *y*, and *z*, collectively representing a 3D data structure. These axes encompass a multitude of individual pixels, each contributing specific radiodensity information at their respective spatial coordinates. Using the definition of these three axes, the above-mentioned average processing within the local area can be expressed mathematically as shown in Eq. (1):

$$D'(\mathrm{i},j,k) = \frac{1}{r^3}\Sigma_{x=i-\frac{r-1}{2}}^{i+\frac{r-1}{2}}\Sigma_{y=j-\frac{r-1}{2}}^{j+\frac{r-1}{2}}\Sigma_{z=k-\frac{r-1}{2}}^{k+\frac{r-1}{2}}D(x,y,z) \quad (1)$$

$$x,y,z \in Z$$

where $D'(i,j,k)$ represents the blurred intensity value at voxel (i,j,k). $D(x,y,z)$ is the original intensity value at voxel (x,y,z), r is a positive integer that defines the cubic neighborhood size, and x,y,z are integers that denote the voxel positions in the original image.

This method effectively integrates a wide range of local information, and due to the function of the averaging process, it leads to a reduction in noise, allowing for the observation of broader patterns of variations. Additionally, this method presents an advantage by preserving the overall context, thus enhancing the potential to identify

meaningful trends within the data. This ability to handle a wide range of variations effectively contributes to improved model performance and robustness against diverse data distributions.

To further enhance the model's ability to extract features, we decided to incorporate an attention gate, as proposed by Abraham *et al.* [8] and Oktay *et al.* [9] (see Figs. 3 and 4).

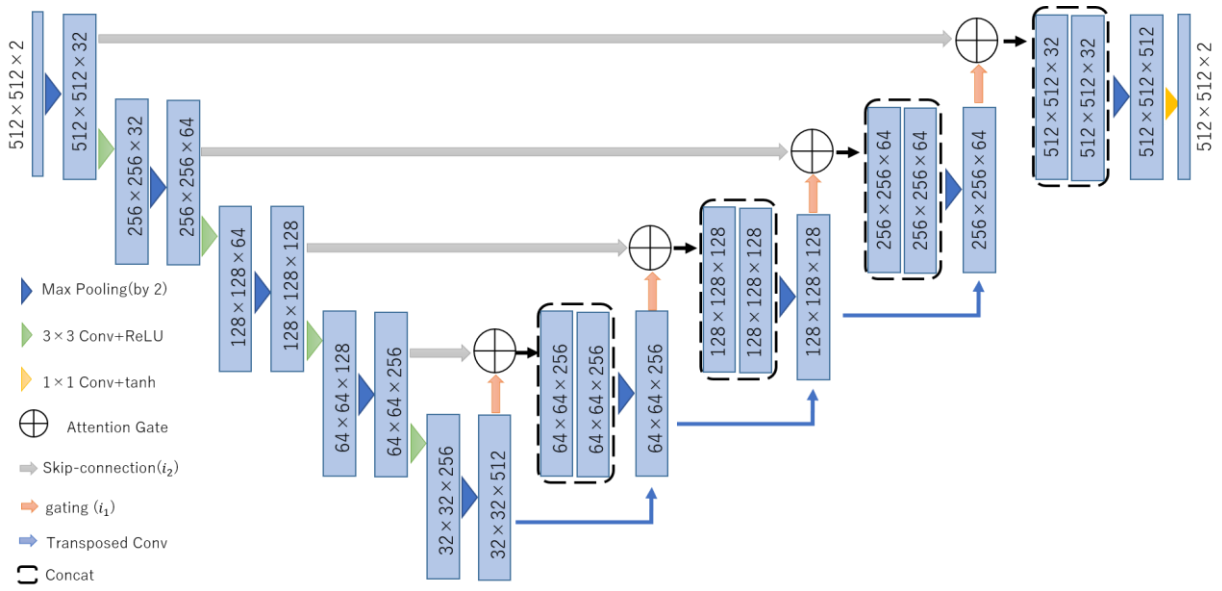

Fig. 3. U-Net model using attention gate.

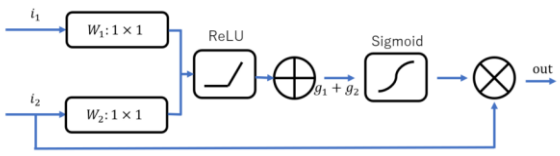

Fig. 4. Attention gates.

We perform convolution operations on i_1 and i_2, where i_1 represents the upsampled component, and i_2 is derived from the skip-connection. W_1 and W_2 are incorporated into the model as attention gate convolution kernels. Specifically, these weights are defined using a Conv2D layer in Keras and using glorot_uniform (Xavier initialization) as the initialization method. In this context, the weight matrices W_1 and W_2 serve as weight matrices in the convolution operation, enabling the learning and application of spatial filters to the input feature maps. These filters capture relevant features for subsequent processing stages, enhancing the representation of the input data. The ReLU activation function is signified by *f*, and the convolution operation is indicated by the asterisk (*). Following these operations, we proceed to sum g_1 and g_2. Thereafter, we execute another convolution operation on g and pass the result through a sigmoid activation function, culminating in the attention weights (a), in which W_3 represents the weight matrix of the convolution operation. b_1, b_2, b_3 are biases, these are adaptation constants. Lastly, the attention weights are multiplied by i_2 to yield the output, where the symbol $\otimes$ denotes element-wise multiplication and out represent the output. This framework incorporates an attention mechanism, efficiently emphasizing the critical features in the input maps. The corresponding mathematical equations are given as follows, from Eqs. (2)–(6).

$$g_1 = f(W_1 * i_1 + b_1) \quad (2)$$

$$g_2 = f(W_2 * i_2 + b_2) \quad (3)$$

$$g = g_1 + g_2 \quad (4)$$

$$a = sigmoid(W_3 * g + b_3) \quad (5)$$

$$out = a \otimes i_2 \quad (6)$$

This approach allows us to adjust the features of i_2 by emphasizing the relevant features and weakening the less important features using the attention factor. As a result, it is easier to perform focused and effective tasks, which can improve model performance.

As mentioned earlier, our study is based on a pseudo-3D convolutional CNN model. When evaluating using test data, we use the weights of both models. The approach is to set equal weights during the voting process.

Our unique approach, which involves applying the obtained weights to perform an average vote, providing numerous benefits. Enhance model diversity, reduce the risk of overfitting, and allow prediction errors to average out. This method effectively integrates the model's predictive capabilities, reducing model complexity and reducing the risk of overfitting (Fig. 5). Therefore, our approach reduces and balances model integration and overfitting to improve the performance of pseudo-3D CNN models in medical image segmentation tasks.

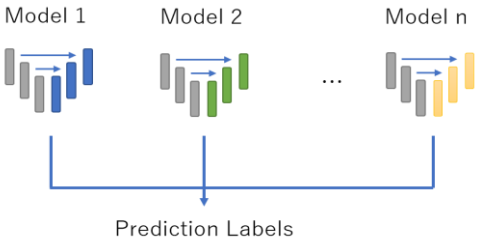

Fig. 5. Proposed average voting.

In the present study, a collective dataset, designated as "Dataset 15: Test set for CSI 2014 Vertebra Segmentation Challenge", encompassing ten exemplary spinal samples (Case 1 through Case 10), was meticulously selected. Additionally, an isolated case manifesting a compression fracture, as documented on SpineWeb [10], was incorporated, culminating in a comprehensive experimental ensemble. Among the healthy vertebral specimens, a subset comprising five cases (Case 1 to Case 5) was delineated for the learning phase, serving as the foundational basis for algorithm training. The remainder consisted of five original vertebrae along with compression fractured vertebrae, confirming the robustness and generalizability of the developed algorithmic framework in identifying and delineating spinal anatomy and pathology. It is treated as validation data for the purpose.

IV. Result and Discussion

Table I presents the degree of accuracy, as denoted by the correct label, achieved in the extraction of the healthy spinal region from the unlearned data, employing the methodology of Shigeta *et al.* [1], the 3×3×3 range, and both 1×1×1 and 7×7×7 ranges with attention learning.

TABLE I. The Degree of Coincidence between Segmented Healthy Spine Region and Correct Label Region for Each Unlearned Spine Data

	Dice Coefficient		
Case No.	**Shigeta's Method**	**3×3×3 range**	**1×1×1 and 7×7×7 with attention**
6	0.937	0.940	0.943
7	0.963	0.968	0.973
8	0.964	0.969	0.974
9	0.957	0.957	0.969
10	0.966	0.969	0.975
Average	0.957	0.961	0.968

When we analyzed Shigeta *et al.*'s [1] method that combined the 7×7×7 range with the 1×1×1 and 7×7×7 range attention mechanisms, we found that the latter combination had the highest concordance rate. We compared Dice Coefficient values of the three methods, namely methods using Tukey's multiple comparison method. As a result, statistically significant differences were found between all pairs among three methods ($p < 0.05$). This result suggests that combining multiple ranges and attention mechanisms may improve accuracy at a single range.

Fig. 6 show the spine extraction results for unlearned healthy spine data (Case No. 10) by each method. It represents the original CT image (Fig. 6(a)), the correct label (Fig. 6(b)), and the extraction result (Fig. 6(c)), respectively.

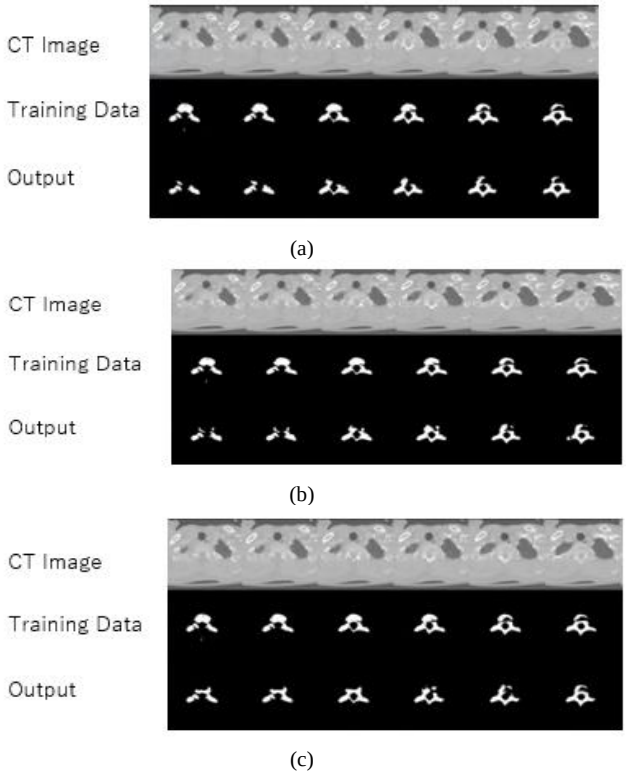

Fig. 6. Spine segmentation results: (a) by Shigeta *et al.* [1].; (b) 3×3×3 range; (c) average votes with 1×1×1 and 7×7×7 ranges using attention gate.

Furthermore, by considering the shortcomings of each method as "individuality" and complementing each other's strengths and weaknesses, extraction performance can be improved. We observed that methods using attention mechanisms are better at finding and highlighting detailed features and outperform other methods in extracting complex anatomical structure. However, the applicability and computational cost of each method requires further research.

Table II shows the degree of accuracy by correct label as the result of extracting the healthy spinal region from the unlearned data by the different range of average labels with attention learning.

TABLE II. The Degree of Coincidence between Segmented Healthy Spine Region and Correct Label Region for Each Unlearned Spine Data

	Dice Coefficient			
Case No.	**1×1×1 with attention**	**3×3×3 with attention**	**5×5×5 with attention**	**7×7×7 with attention**
4	0.951	0.950	0.950	0.953
5	0.962	0.966	0.966	0.965
6	0.939	0.882	0.882	0.916
7	0.970	0.967	0.967	0.967
8	0.968	0.966	0.966	0.970
9	0.959	0.954	0.954	0.952
10	0.972	0.968	0.968	0.972
Average	0.960	0.950	0.950	0.956

Table III shows the Dice scores and related references of prominent spinal segment studies conducted in recent years. As is evident from this table, various algorithms and methodologies have been proposed and their effectiveness and performance verified. Although our proposed method achieved a Dice score of 0.968, which outperforms all other studies listed, these results should be interpreted with caution as different studies used different datasets. is needed. Therefore, although the dataset is not consistent across all studies, our method shows competitive, if not superior, performance. Comparative experiments using common datasets are essential to conclusively establish that our proposed method is superior to previous models. We currently lack comprehensive data to directly compare performance with other studies.

TABLE III. DICE COEFFICIENT FOR SPINAL SEGEMENTATION IN RECENT STUDIES

Team	Dice score
You [11]	0.86
Rehman [12]	0.94
Carson [13]	0.92
Qadri [14]	0.87
Dabiri [15]	0.96
Sekuboyina [16]	0.92
Zhang [17]	0.94
Khandelwal [18]	0.92
Tao [19]	0.95
Chen [20]	0.87
Proposed	0.968

In this study, we introduced diversity through ensemble learning, allowing the model to capture a wider range of features. Additionally, by blurring the image preprocessing step, we aimed to capture more global structure by obtaining a lower-resolution representation. Moreover, rather than directly using the blurred results, we obtained the difference to capture aspects of spatial variation. This approach improved overall prediction accuracy by allowing us to focus on more important structural features while suppressing noise and unnecessary details.

An important problem for future research is the improvement of vertebral compression assessment, which is expected to improve accuracy. Furthermore, techniques used in more advanced ensemble learning have the potential to achieve superior predictive performance for further exploration and optimization.

V. CONCLUSION

In this research, we attempted to apply CNN to medical image processing. Specifically, we utilized pseudo-3D convolutional CNN models, incorporated multiscale blurring techniques on the data to obtain multiresolution representations, and equipped these models with attention mechanisms for CT imaging. We adopted an ensemble approach, specifically a weighted average vote using weights derived from both models during training. For predictions using test data, we leveraged the weights of both models to automatically extract spine regions more accurately. This approach resulted in an average accuracy of 96.8% on the test data.

Our findings show that while three-dimensional feature learning exhibits a certain level of effectiveness in improving extraction accuracy, there is a necessity for further scrutiny of the technique. Future prospects for this research could involve the enlargement of our dataset and the adoption of more advanced ensemble methods, which hold potential for enhancing prediction accuracy. Furthermore, we contemplate incorporating measures to appropriately extract data that deviates from normal vertebral morphology or CT values, as observed in cases of fractures or tumors.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Kai Yang wrote the manuscript as well as implemented the ensemble learning method and made necessary modifications to the integration of attention gates and programming aspects of the study; Masayuki Kikuchi has contributed to the theoretical foundations of semi-global methodology research. He provided valuable support and advice during the experimental phase of our research; All authors approved the final version.

ACKNOWLEDGEMENTS

We thank Kanon Kobayashi for conducting some preliminary experiments to confirm the effectiveness of 3D smoothing for different sizes.

REFERENCES

[1] N. Shigeta, M. Kamata, and M. Kikuchi, "Effectiveness of pseudo 3D feature learning for spinal segmentation by CNN with U-Net architecture," *Journal of Image and Graphics*, vol. 7, no. 3, pp. 107–111, 2019.

[2] O. Ronneberger, P. Fischer, and T. Brox, "UNet: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015*, 2015, pp. 234–241.

[3] J. Huang, F. Jian, and H. Wu, "An improved level set method for vertebra CT image segmentation," *Biomedical Engineering Online*, May 2013.

[4] Y. Wang, J. Yao, H. R. Roth, J. E. Burns, and R. M. Summers, "Multi-atlas segmentation with joint label fusion of osteoporotic vertebral compression fractures on CT," in *Proc. 3rd Workshop & Challenge on Computational Methods and Clinical Applications for Spine Imaging*, 2015, pp. 74–84.

[5] R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "Interpolation-based shape-constrained deformable model approach for segmentation of vertebrae from CT spine images," in *Proc. 2nd Workshops on Computational Methods and Clinical Applications for Spine Imaging*, Boston, 2014, pp. 235–240.

[6] M. Vania, D. Mureja, and D. Lee, "Automatic segmentation of spine using convolutional neural networks via redundant generation of class labels," arXiv preprint, arXiv:1712.01640, November 2017.

[7] M. Kamata, K. Fukushima, H. Shouno, I. Hayashi, and M. Kikuchi, "Automatic detection of spine in CT image by U-Net," in *Proc. Nonlinear Circuits, Communications and Signal Processing*, NCSP2018, Honolulu, Hawaii, USA, 2018, pp. 608–610.

[8] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019.

[9] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," arXiv preprint, arXiv:1804.03999, 2018.

[10] SpineWeb. [Online]. Available: http://spineweb.digitalimaginggroup.ca
[11] X. You *et al.*, "VerteFormer: A single-staged transformer network for vertebrae segmentation from CT images with arbitrary field of views," *Medical Physics*, vol. 50, pp. 6296–6318 2023. https://doi.org/10.1002/mp.16467
[12] F. Rehman *et al.*, "A robust scheme of vertebrae segmentation for medical diagnosis," *IEEE Access*, vol. 7, pp. 120387–120398, 2019.
[13] B. Carson, "Automatic bone structure segmentation of under-sampled CT/FLT-PET volumes for HSCT patients," M.S. Thesis, University of Oklahoma, 2021.
[14] S. F. Qadri *et al.*, "SVseg: Stacked sparse autoencoder-based patch classification modeling for vertebrae segmentation," *Mathematics*, vol. 10, no. 5, 796, 2022.
[15] S. Dabiri, "Deep learning-based computed tomography image processing," Ph.D. thesis, Simon Fraser University, 2022.
[16] A. Sekuboyina *et al.*, "VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images," *Medical Image Analysis*, vol. 73, 102166, 2021.
[17] D. Zhang, A. Aoude, and M. Driscoll, "Development and model form assessment of an automatic subject-specific vertebra reconstruction method," *Computers in Biology and Medicine*, vol. 150, 106158, 2022.
[18] P. Khandelwal and P. Yushkevich, "Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging," in *Domain Adaptation and Representation Transfer and Distributed and Collaborative Learning*, Cham, Switzerland: Springer, 2020, pp. 73-84.
[19] R. Tao, W. Liu, and G. Zheng, "Spine-transformers: Vertebra Labeling and segmentation in arbitrary field-ofview spine CTs via 3D transformers," *Medical Image Analysis*, vol. 75, 102258, 2022.
[20] Y. Chen, Y. Gao, K. Li, L. Zhao, and J. Zhao, "Vertebrae identification and localization utilizing fully convolutional networks and a hidden Markov model," *IEEE Trans. Med. Imaging*, vol. 39, no. 2, pp. 387–399, 2020.