A Deep Learning Study on Osteosarcoma Detection from Histological Images

D M Anisuzzaman[a], Hosein Barzekar[a,*], Ling Tong[b], Jake Luo[b], Zeyun Yu[a,c]

[a]*Big Data Analytics and Visualization Laboratory, Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA*
[b]*Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA*
[c]*Department of Biomedical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA*

Abstract

In the U.S, 5-10% of new pediatric cases of cancer are primary bone tumors. The most common type of primary malignant bone tumor is osteosarcoma. The intention of the present work is to improve the detection and diagnosis of osteosarcoma using computer-aided detection (CAD) and diagnosis (CADx). Such tools as convolutional neural networks (CNNs) can significantly decrease the surgeon's workload and make a better prognosis of patient conditions. CNNs need to be trained on a large amount of data in order to achieve a more trustworthy performance. In this study, transfer learning techniques, pre-trained CNNs, are adapted to a public dataset on osteosarcoma histological images to detect necrotic images from non-necrotic and healthy tissues. First, the dataset was preprocessed, and different classifications are applied. Then, Transfer learning models including VGG19 and Inception V3 are used and trained on Whole Slide Images (WSI) with no patches, to improve the accuracy of the outputs. Finally, the models are applied to different classification problems, including binary and multi-class classifiers. Experimental results show that the accuracy of the VGG19 has the highest, 96%, performance amongst all binary classes and multiclass classification. Our fine-tuned model demonstrates state-of-the-art performance on detecting malignancy of Osteosarcoma based on histologic images.

Keywords: Computer aided diagnosis, Deep learning, Osteosarcoma, Histological Image, Transfer learning

1. Introducation

Primary bone tumors account for 5-10% of all new pediatric cancer diagnoses. Osteosarcoma is the most common form of malignant primary bone tumor. Despite the limited approximately 1,000 new cases every year in the United States, the prognosis of osteosarcoma remains a challenging issue [1]. There are two age peaks of incidence among patients, with a peak age of children under age 10, and adolescents at age 10-20 [2]. Osteosarcoma cancer usually occurs in the metaphysis of long bones on lower limbs, consisting of 40-50% of the total cases [1]. The symptoms of osteosarcoma usually begin with mild localized bone pain, redness, and warmth at the site of the tumor. Patients experience increasing pain, which often affects patients' movement and joint functions. If the early phase of osteosarcoma is not treated, it is expected to see a wide range of metastasis such as at lungs, other bones and soft tissues [3].

Histological biopsy tests, X-ray tests and magnetic resonance images are essential diagnosis to of osteosarcoma. Currently, the diagnosis of osteosarcoma includes a detailed history taking and physical examinations [4, 5]. The presenting symptoms typically include deep-seated, constant, gnawing pain and swelling at the effected site. Pain in multiple areas may portend skeletal metastasis; therefore, they should be investigated appropriately [5].Beyond the examination, the standard studies for evaluation of potential osteosarcoma are laboratory tests, an X-ray of the entire affected bone, a magnetic resonance imaging (MRI) scan of the entire affected bone, a chest X-ray, a chest computed tomography (CT) scan, a whole-body technetium bone scan, and a percutaneous image-guided biopsy [5].

*Corresponding author
Email addresses: anisuzz2@uwm.edu (D M Anisuzzaman), barzekar@uwm.edu (Hosein Barzekar), ltong@uwm.edu (Ling Tong), jakeluo@uwm.edu (Jake Luo), yuz@uwm.edu (Zeyun Yu)

Although the biopsy-based methods can effectively discover the malignancy, limitations in histological-guided biopsies and MRI scans have limited detecting capacity. Additionally, the preparation of histological specimens is time-consuming. For example, an accurate detection of osteosarcoma malignancy requires preparation of at least 50 histology slides to represent a plane of a large three-dimensional tumor [2].

Due to the rise of cancer incidence and patient-specific treatment options, diagnosis and treatment of cancer are becoming more complex [6]. Pathologists must spend an extremely long time examining a large number of slides. Detecting the nuances of histological images can be difficult [7]. Misdiagnosis often occurs due to the extensive work that decreases the accuracy of diagnosis. The osteoblasts' morphology has little difference in differentiated cells, which makes the image barely distinguishable. Also, the biopsy is a vital and time-consuming step to determine the presence of malignant tissue. Meanwhile, Computer-Aided Detection (CAD) technology offers a solution for radiologists to automatically detect malignancies [8].

To address these limitations, microscopic image-based analysis has been the foundation of cancer diagnosis in recent years [5]. However, it was not practical before the 2000s because of relatively low detection accuracy. The poor performance of CAD made clinical implementation impractical until the recent advances in computerized image detection [9].

Recent advances enabled the possibility of turning histological slides to digital image datasets, in which machine learning can intervene on digital images to address some of the limitations. With the advent of whole slide imaging (WSI), digital pathology has become a part of the routine procedure in clinical diagnosis. The emergence of digital pathology provides new chances of developing new algorithms and software. A histological image can be quantified in such a system in order to improve the pathological procedures. The system digitizes glass slides with stained tissue sections at very high-resolution images, which makes computerized image analysis viable [10].

The primary goals of this study are:

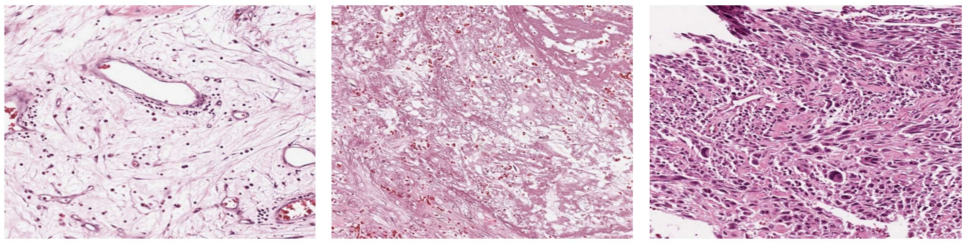
1) To demonstrate that the development of deep learning-based tools is capable of detecting the osteosarcoma malignancy with high accuracy based on a public dataset. The purpose is to successfully distinguish the typical patterns of non-tumors, necrotic tumors and viable tumors.

2) To explore a suitable deep learning framework for accurate detection and discover possible clues that contribute to performance. To achieve the goals, histological medical image analysis based on transfer learning were applied to the pathology archives at Children's Medical Center's dataset [11]. Two modified transfer learning approaches including VGG19 [12] and Inception V3 [13] models were applied to the data. Compared to the previous results, we achieved an overall 2% improvement in accuracy. The novelty of the model is that is being applied to different categories of the dataset and using the whole tile image as the input.

2. Literature Review

Computer-aided technology in radiological and biopic detection becomes viable since 2010 [14]. Remarkable progress has been achieved in medical images, primarily due to the availability of large-scale datasets and deep convolutional neural networks (CNNs) in the computer science area [15]. This technology has been widely applied to a variety of medical images for the detection of different diseases, such as chest X-ray pneumonia, breast cancer, pulmonary edema, pulmonary fibrosis, gastric endoscopic images for celiac diseases and gastric cancer [6, 16]. The x-ray and biopsy for osteosarcoma share a similar pattern with these diseases; Therefore, it is practically feasible to use CNN to detect the early stage tissue morphological change. To decrease the mortality, it is imperative to prevent the early stage tumor from metastasis. Early automatic detection can not only decrease the chance of misdiagnosis but also serve as an assistant tool for the surgeon's preference to determine if metastasis has occurred. We believe the adoption of computer-aided technology using CNN can significantly reduce the surgeon's workload and achieve a better diagnosis of patients.

Several state-of-the-art studies based on deep learning has been recognized as a recent major enhancement in histological image detection; However, most efforts of image detection are focused on histological images of breast cancer. In 2017, Jongwon [17] did a pilot study on histopathology of breast cancer, which achieves an AUC value of 93% on microscopic biopsy images in classifying benign or malignant tumors. They show that transfer learning is a viable and pre-trained model that is useful in classifying histological images. Erkan's



(a) Non tumor (b) Necrotic tumor (c) Viable tumor

Figure 1: Sample Images from the dataset

result [18] shows the state-of-the-art performance using VGG16 and AlexNet models, with an average of 90.96±1.59% accuracy. This also indicates the suitability of these models for image classification tasks.

Other examples of image classification in recent years use similar methods: Jonathan De Matos [19] used double transfer learning to classify histopathologic images. Noorul Wahab [20] aimed at a more challenging task of segmentation and detection of mitotic nuclei. They used a similar hybrid CNN model and achieves 76% AUC value. Other examples include the prediction of pathological invasiveness in lung adenocarcinoma [21], Classification of Liver Cancer Histopathology Images [22] , and Automated invasive ductal carcinoma detection [23].

In Harish Babu Arunachalam's study [24], the article reports the first fully automated tool to assess viable and necrotic tumor in osteosarcoma using histological images and deep learning models. The goal is to label the diverse regions of tissue into a viable tumor, necrotic tumor, and non-tumor. They employed both machine learning and deep learning models. The ensemble learning model achieved an overall accuracy of 93.3% with class-specific accuracies of 91.9% for non-tumor, 95.3% for viable tumor, and 92.7% for necrotic tumor.

In machine learning and data mining algorithms, the main premise is that training and potential data should be in the same space and distribution. The problem arises when we have no access to enough training data in the specific research domain. Hence, we can obtain the basic parameters for training our deep learning model from pre-trained networks applied to larger data sets from other domains. In these situations, knowledge-transferring significantly improves learning outputs if done efficiently while minimizing expensive data labeling efforts [25].

3. Methodology

3.1. Dataset

The dataset used in the study was obtained from the work of Arunachalam et al. where they provided a data set of osteosarcomas and conducted a variety of machine learning and deep learning techniques. Tumor samples from the Children's Medical Center, Dallas, were collected from the pathology reports of the osteosarcoma resection for 50 patients treated between 1995 and 2015. They selected 40 WSIs of the digitized images representing tumor heterogeneity and response properties in the study. In each WSI, 30 1024×1024 pixel image tiles were randomly selected at the 10X magnification factor. 1,144 of the resulting 1,200 image tiles, such as those that fall into non-fabric, ink marks regions, and blurry images were chosen after removing irrelevant tiles. Moreover, they generated 56,929 patches of 128×128 pixels. Some sample dataset images are shown in Figure 1.

3.2. Data Preprocessing

Original images of 1024×1024 pixels were used for model training, validation, and evaluation. We split the datasets into training, validation, and testing images at a ratio of 70%, 10%, and 20% respectively. The data are then augmented by using image data generator of "Keras"[26]. In this step, all image intensities are first rescaled to the range of 0 to 1, and then the following augmentations have been performed: rotation, width shift, height shift, vertical flip, and horizontal flip. Due to memory limitations, we down sampled the original images by passing the input shape of 375×375, rather than 1024×1024.

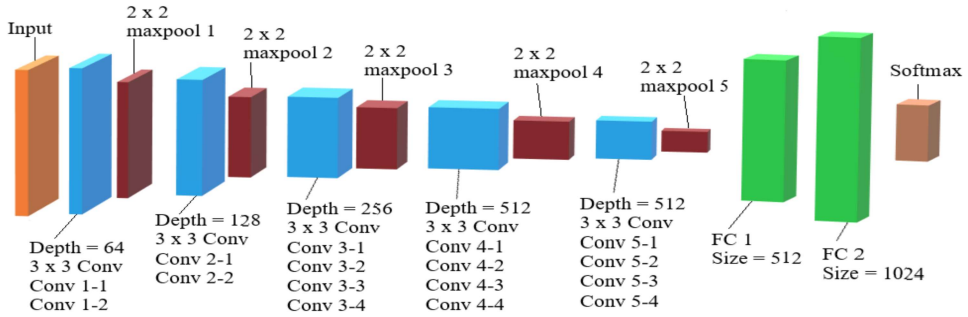


Figure 2: VGG19 Network Architecture

3.3. Model Selection

There are 26 deep learning models in Keras Applications that can be used for prediction, feature extraction, and fine-tuning [26]. Six of these models are applied for multi-class classification and among them we have chosen the best model for our experiment depending on the test accuracy. Table 1 shows the test results of these models. VGG19 gives the best result among these models and we choose this model for future experiments.

Table 1: Multi-class Result of Various Models

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1-Score	Accuracy
VGG16	0.89	0.88	0.88	0.883
VGG19	0.94	0.94	0.94	0.939
ResNet50	0.22	0.47	0.30	0.470
InceptionV3	0.81	0.78	0.79	0.783
DenseNet201	0.61	0.58	0.56	0.583
NASNetLarge	0.80	0.79	0.79	0.791

3.3.1. VGG19 Model

We have used Keras applications for importing VGG19 model. Pre-trained weights have been used for model training. We have discarded the fully connected layer along with output layer of the VGG19 model. We have added two fully connected layers after the last "maxpool" layer. Dropout layers are used for avoiding over-fitting the training data. We have used "Relu" activation in the dense layers and "softmax" activation function in the output layer. Figure 2 shows the VGG19 model architecture. All the "Conv 1-1" to "Conv 5-4", and "maxpool 1" to "maxpool 5" use pre-trained weights. We have added the FC1, FC2, and softmax layers to this network. As shown in the figure, all the convolution layers use 3×3 filters, and all the maxpooling layers use 2×2 filters. The FC1 and FC2 layers contain 512 and 1024 neurons respectively. softmax layer's neurons varies depending on our classification task. For binary and multi-class classification, it contains two and three neurons respectively.

4. Experimental Results

4.1. Setup

With our dataset containing three classes, we performed four binary classifications and a multiclass (three classes) classification. In each classification, we applied two models: VGG19 and Inception V3. Inception V3 has been used for model comparison. The models are written in the Python programming language in the Keras deep learning framework. The models are trained and tested on a Nvidia GeForce RTX 2080Ti GPU platform.

The loss functions used for binary classification and multiclass classifications are binary cross entropy and categorical cross entropy respectively. In both types of classification, Adam optimizer is applied for minimizing the loss function by updating the weight parameters. The learning rate is set to Keras's default 0.01. Batch size is set to 80, 28, and 16 for training,

Actual Class \ Predicted Class	NT	NCT+VT
NT	**101**	7
NCT+VT	4	**118**

(a) NT vs. NCT+VT with VGG 19

Actual Class \ Predicted Class	NT	NCT+VT
NT	95	13
NCT+VT	14	108

(b) NT vs. NCT+VT with Inception V3

Actual Class \ Predicted Class	NCT	NT
NCT	**51**	2
NT	5	**103**

(c) NCT vs. NT with VGG 19

Actual Class \ Predicted Class	NCT	NT
NCT	48	5
NT	12	96

(d) NCT vs. NT with Inception V3

Actual Class \ Predicted Class	NT	VT
NT	103	5
VT	3	**66**

(e) NT vs. VT with VGG 19

Actual Class \ Predicted Class	NT	VT
NT	**107**	1
VT	32	37

(f) NT vs. VT with Inception V3

Actual Class \ Predicted Class	NCT	VT
NCT	48	5
VT	4	**65**

(g) NCT vs. VT with VGG 19

Actual Class \ Predicted Class	NCT	VT
NCT	**53**	0
VT	21	48

(h) NCT vs. VT with Inception V3

Actual Class \ Predicted Class	NCT	NT	VT
NCT	**48**	3	2
NT	2	**103**	3
VT	2	2	**65**

(i) multiclass with VGG 19

Actual Class \ Predicted Class	NCT	NT	VT
NCT	44	6	3
NT	12	93	3
VT	19	7	43

(j) multiclass with Inception V3

Figure 3: Confusion matrixes of all classifications. Here, NT = Non-Tumor, NCT = Necrotic Tumor, and VT = Viable Tumor

validation, and testing respectively. All models are trained for 1500 epochs, with a callback that stops training when validation accuracy reaches over 98%.

Two-class classifications are evaluated on the following datasets: 1.) Non-Tumor (NT) versus Necrotic Tumor (NCT) and Viable Tumor (VT), 2.) Necrotic Tumor versus Non-Tumor, 3.) Viable Tumor versus Non-Tumor, and 4.) Necrotic Tumor versus Viable Tumor. We also performed the multiclass classification among the three classes: NT, NCT and VT. To evaluate our model performance, we presented confusion matrix, precision, recall, f1 score, and accuracy for all classifications. We also reported receiver operating characteristic (ROC) curve and area under the curve (AUC) for all the two-class classifications.

Precision measures the percentage of correctly classified images in that specific predicted class, and recall measures the percentage of correctly classified images in the ground truth. F1 score is the weighted average of precision and recall. Accuracy measures the percentage of correctly classified (predicted) images among all the predictions. The receiver operating characteristic (ROC) curve shows the diagnostic ability of a binary classifier system for different thresholds. This curve plots the true positive rate (sensitivity) against false positive rate (1-specificity). The area under the curve (AUC) indicates that the classifier gives a randomly chosen positive instance higher probability than a randomly chosen negative instance.

4.2. Results

The evaluation metrics for all the classifications with two models are briefly presented in the following sections. Figure 3 shows the confusion matrix for all classifications with all three networks.

Table 2 and 3 show the precision, recall, and f1 score for all the binary and multiclass classifications with each of the present networks. Figure 4 shows the accuracy of the classifiers for all the classifications.

5. Discussion

Osteosarcoma is a common tumor in pediatric cases of cancer which requires extensive work of pathologists in order to confirm the case. While other medical images have already performed computerize analysis, osteosarcoma histological image is rarely mentioned in classification using deep learning models. We believe it is possible to make use of computer-aided technology to help classify and recognize the image of a malignant tumor. In this study, a

Table 2: Precision, Recall, and F1-Score for binary classes

Non-Tumor versus Necrotic Tumor and Viable Tumor						
	Non-Tumor			Necrotic and Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG19	0.96	0.94	0.95	0.94	0.97	0.96
Inception V3	0.87	0.88	0.88	0.89	0.89	0.89
Necrotic Tumor versus Non-Tumor						
	Necrotic Tumor			Non-Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG19	0.91	0.96	0.94	0.98	0.95	0.97
Inception V3	0.8	0.91	0.85	0.95	0.89	0.92
Viable Tumor versus Non-Tumor						
	Non-Tumor			Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG19	0.97	0.95	0.96	0.93	0.96	0.94
Inception V3	0.77	0.99	0.87	0.97	0.54	0.69
Necrotic Tumor versus Viable Tumor						
	Necrotic Tumor			Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG19	0.92	0.91	0.91	0.93	0.94	0.94
Inception V3	0.72	1	0.83	1	0.7	0.82

Table 3: Precision, Recall, and F1-Score for Multicalss

Multiclass									
	Necrotic Tumor			Non-Tumor			Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
VGG19	0.92	0.91	0.91	0.95	0.95	0.95	0.93	0.94	0.94
Inception V3	0.59	0.83	0.69	0.88	0.86	0.87	0.88	0.62	0.73

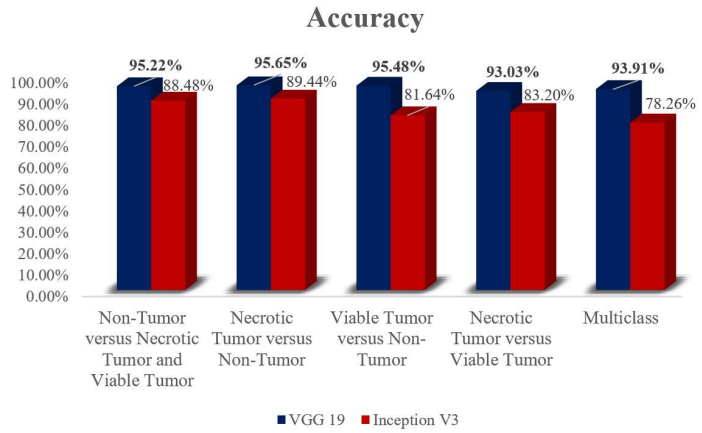
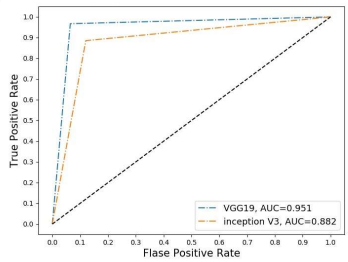
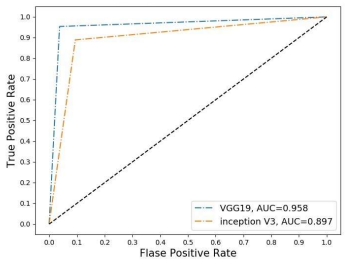


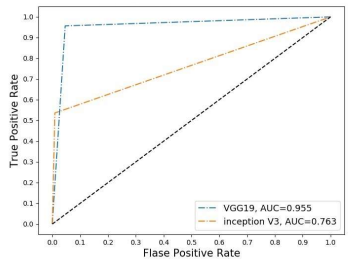
Figure 4: Accuracy Scores



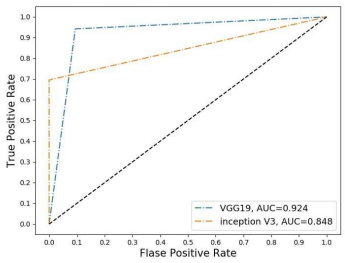
(a) Non-Tumor vs (Necrotic and Viable Tumor)



(b) Necrotic Tumor vs Non-Tumor



(c) Non-Tumor vs Viable Tumor



(d) Necrotic Tumor vs Viable Tumor

Figure 5: ROC and AUC of all two-class classifications

deep learning-based technique has been used for image classification to detect the histologic images to identify malignancy of osteosarcoma. Our study provides an option of using a computer to accelerate the diagnosis and detection of osteosarcoma malignancy. Furthermore, we apply and compare two popular network architectures VGG19 and Inception V3[12, 13]. Thus, we obtain higher performance than prior studies with the same dataset. We have configured and tested models with custom layers to achieve the best performance.

From Figure 3, we can see that for NT vs VT and NCT vs VT respectively the prediction of non-tumor and necrotic tumor is performed well by Inception V3. In all other cases VGG19 works very good compared to Inception V3. So, in overall balance VGG19 beats Inception V3.

From Table 2 and 3, we can see that for VT vs NT and NCT vs VT cases precision of viable tumor and recall of necrotic tumor and non-tumor are high for Inception V3. But the interesting fact is that all the f1 scores are higher for VGG19 model. Since f1 score indicates the weighted average of precision and recall, a higher f1 score means precision and recall are close to each other for VGG19, where for inception V3 only a single metric is higher (either precision or recall) indicating lower score of the other one. Hence, in balance in overall performance, VGG19 beats inception V3 by a huge margin. From Figure 4, it is clear that for all classifications VGG19 achieves the highest accuracy.

From Figure 5, we can see that VGG19 has the highest AUC value for all binary (two-class) classifications. The AUC values are impressive (0.95, 0.96, 0.96, and 0.92 for non-tumor versus necrotic tumor and viable tumor, necrotic tumor versus non-tumor, viable tumor versus non-tumor, and necrotic tumor versus viable tumor classifications respectively), which assures us with great reliability. So, from all the above analytical discussion, it is safe to say that VGG19 works well for all classifications. While Inception V3 has three types of convolutions (1×1, 3×3, 5×5), VGG19 has only one type of convolution (3×3). Instead of going deeper, Inception V3 goes wider on an image feature searching. As our dataset contains biopsy images in which some parts may only contain some specific features of a specific class (necrotic or viable), some of the inception kernels may not provide good features and in the concatenation layer, the performance may decrease. In VGG19, the kernel size is always the same (3×3); which may lead to better classification accuracy specifically for our dataset. This dataset has a small number of images (1144), which is not suitable for deep learning models. Deep learning demands lots of data to learn the connection between given input and corresponding output. To overcome the data limitation problem, we applied transfer learning approach. Both VGG19 and inception V3 are pre-trained with the imagenet dataset, where all the low-level features (edge, curve etc.) are trained with imagenet dataset and we

transfer that learned weights to our dataset. The fully connected layers and output layers are replaced in both models and trained with our dataset.

To the best of our knowledge, this is the first pipeline that have been used in VGG19 and Inception architecture in Deep learning to recognize the osteosarcoma malignancy. The adjusted model can identify the minimal differences of images to predict the early signs of cancer. If the pipeline was deployed in various medical facilities, our model could help pathologists as an adjunct tool reducing their extensive work.

The best accuracy is achieved by the VGG19 model compare to Arunachalam et al.'s deep learning model (a CNN model with three pairs of convolutions and pulling layers for sub-sampling, and two fully connected multi-layer perceptron). Table 4 represents the comparison of these two works. We have done a binary classification for all possible combinations between three classes, where Arunachalam et al. [24]'s deep learning model provides a direct class specific accuracy. Therefore Table 4 represents our average accuracy for a specific tumor class. For viable tumor the average of VT vs NT and NCT vs VT; for necrotic tumor the average of NCT vs NT and NCT vs VT; and for non-tumor the average of NT vs NCT and VT, NCT vs NT, and VT vs NT is represented. The comparison is done on the whole images (tile accuracy [24]), as we have used the 1144 whole images for our classification. Table 4 shows a better performance of non-tumor than other classes, which may be caused by the imbalance data in each class. This dataset contains 536, 345, and 263 whole images of non-tumor, viable tumor, and necrotic tumor respectively.

Table 4: Result Comparison

Tumor type	Tile accuracy in %	
	VGG19	Arunachalam[24]'s deep learning model
Non-Tumor	95.45	89.5
Necrotic Tumor	94.34	91.5
Viable Tumor	94.26	92.6

Limitations include the lack of evaluation from pathologists. Even though our model reaches a high performance, it is suggested that the tool should be used under a pathologist's supervision. A further study is to compare our model's performance with expert pathologists. The comparison can make sure this tool can detect new malignant cases in clinical practices. Besides, the existing data set might not indicate the future histological images from patients, therefore, the generalizability of our model might be problematic. To address this issue, it would be helpful to be adopted in medical facilities to assess its performance.

6. Conclusion

Within the area of medical image processing, it is important to automate the classification of histological images by computer-aided systems. It is difficult and time-consuming to carry out a microscopic examination of histological images. Automatic diagnosis of histology alleviates the workload and enables pathologists to focus on critical cases. In this work, we used two pre-trained networks from Keras library, including VGG19 and InceptionV3. Regularization and optimization techniques were performed to avoid variance. The analyses were performed in two different ways, one binary classification, and the other one multi-class classification. VGG19 model achieved the highest accuracy in both binary and multi-class classifications, with an accuracy of 95.65% and 93.91% respectively. Furthermore, the highest F1 score in binary class belonged to the Necrotic Tumor versus Non-Tumor, 0.97. Our study compared to the previous study on the same data have outperformed both binary and multi-class. And finally, this study was the first usage of VGG19 and Inception V3 on the Osteosarcoma dataset, and the same framework can also be applied for other types of cancer.

References

[1] A. J. Chou, D. S. Geller, R. Gorlick, Therapy for osteosarcoma, Pediatric Drugs 10 (2008) 315–327.

[2] C. A. Arndt, W. M. Crist, Common musculoskeletal tumors of childhood and adolescence, New England Journal of Medicine 341 (1999) 342–352.

[3] P. P. Lin, S. Patel, Osteosarcoma, in: Bone Sarcoma, Springer, 2013, pp. 75–97.

[4] J. C. Wittig, J. Bickels, D. Priebat, J. Jelinek, K. Kellar-Graney, B. Shmookler, M. M. Malawer, Osteosarcoma: a multidisciplinary approach to diagnosis and treatment, American family physician 65 (2002) 1123.

[5] D. S. Geller, R. Gorlick, Osteosarcoma: a review of diagnosis, management, and treatment strategies, Clin Adv Hematol Oncol 8 (2010) 705–718.

[6] S. Wang, D. M. Yang, R. Rong, X. Zhan, G. Xiao, Pathology image analysis using segmentation deep learning algorithms, The American journal of pathology 189 (2019) 1686–1698.

[7] P. Picci, Osteosarcoma (osteogenic sarcoma), Orphanet journal of rare diseases 2 (2007) 6.

[8] R. A. Castellino, Computer aided detection (cad): an overview, Cancer Imaging 5 (2005) 17.

[9] A. Madabhushi, G. Lee, Image analysis and machine learning in digital pathology: Challenges and opportunities, Medical Image Analysis 33 (2016) 170 – 175. 20th anniversary of the Medical Image Analysis journal (MedIA).

[10] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, J. Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Scientific reports 6 (2016) 26286.

[11] T. The Cancer Imaging Archive, Osteosarcoma data from ut southwestern ut dallas for viable and necrotic tumor assessment, 2019. URL: `https://doi.org/10.7937/tcia.2019.bvhjhdas`.

[12] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[14] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning, IEEE transactions on medical imaging 35 (2016) 1285–1298.

[15] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural computation 29 (2017) 2352–2449.

[16] A. Serag, A. Ion-Margineanu, H. Qureshi, R. McMillan, M.-J. Saint Martin, J. Diamond, P. O'Reilly, P. Hamilton, Translational ai and deep learning in diagnostic pathology, Frontiers in Medicine 6 (2019).

[17] J. Chang, J. Yu, T. Han, H.-j. Chang, E. Park, A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer, in: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), IEEE, 2017, pp. 1–4.

[18] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, Ü. Budak, Transfer learning based histopathologic image classification for breast cancer detection, Health information science and systems 6 (2018) 18.

[19] J. de Matos, A. d. S. Britto, L. E. Oliveira, A. L. Koerich, Double transfer learning for breast cancer histopathologic image classification, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[20] N. Wahab, A. Khan, Y. S. Lee, Transfer learning based deep cnn for segmentation and detection of mitoses in breast cancer histopathological images, Microscopy 68 (2019) 216–233.

[21] M. Yanagawa, H. Niioka, A. Hata, N. Kikuchi, O. Honda, H. Kurakami, E. Morii, M. Noguchi, Y. Watanabe, J. Miyake, et al., Application of deep learning (3-dimensional convolutional neural network) for the prediction of pathological invasiveness in lung adenocarcinoma: A preliminary study, Medicine 98 (2019).

[22] C. Sun, A. Xu, D. Liu, Z. Xiong, F. Zhao, W. Ding, Deep learning-based classification of liver cancer histopathology images using only global labels, IEEE Journal of Biomedical and Health Informatics 24 (2020) 1643–1651. doi:10.1109/JBHI.2019.2949837.

[23] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, U. R. Acharya, Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images, Pattern Recognition Letters (2020).

[24] H. B. Arunachalam, R. Mishra, O. Daescu, K. Cederberg, D. Rakheja, A. Sengupta, D. Leonard, R. Hallac, P. Leavey, Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models, PloS one 14 (2019) e0210706.

[25] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.

[26] F. Chollet, et al., Keras, https://github.com/fchollet/keras, 2015.