Biomedical Signal Processing and Control 69 (2021) 102931

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

# 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

ARTICLE INFO

*Keywords:*
Computer aided diagnosis
Deep learning
Osteosarcoma
Histological image
Transfer learning

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.

## 1. Introduction

Primary bone tumors account for 5–10% of all new pediatric cancer diagnoses. Osteosarcoma is the most common form of malignant primary bone tumor under the category of bone tumors. 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. Common symptoms include the patient's increasing pain, which often affects patients' movement and joint functions. The early phase of osteosarcoma, if not treated, often results in a wide range of metastasis to other parts of the body such as at lungs, other bones and soft tissues [3].

Histological biopsy test, X-ray test, and magnetic resonance image consist of essential diagnosis of osteosarcoma. Currently, the diagnosis of osteosarcoma includes an initial detailed medical history taking and physical examinations [4,5]. The presenting symptoms that may direct to osteosarcoma typically include deep-seated, constant, gnawing pain and swelling at the affected site. Pain in multiple areas may portend skeletal metastasis; therefore, they should be investigated appropriately [5]. Beyond the examination, a further evaluation of potential osteosarcoma includes the following procedures:

(1) X-ray of the entire affected bone: It is one of the most common ways to diagnose potential tumors. However, the diagnosis of suspicious tumors often requires further confirmation [3]. (2) Magnetic resonance imaging (MRI) scan of the entire affected bone: Doctors use MRI scans frequently for diagnosing joint and bone problems. MRI creates pictures of soft tissue parts of the body that are sometimes hard to see using other imaging tests [1]. (3) Laboratory test, which is a percutaneous image guided-biopsy [5]. Other tests can suggest that cancer is present, but a biopsy can make a diagnosis. One drawback is that the preparation of histological specimens is time-consuming. For example, accurate detection of osteosarcoma malignancy requires the preparation of at least 50 histology slides to represents a plane of a large three-dimensional tumor [2].

The Biopsy is a vital and time-consuming step to determine the

* Corresponding author.
*E-mail addresses:* anisuzz2@uwm.edu (D.M. Anisuzzaman), barzekar@uwm.edu (H. Barzekar), ltong@uwm.edu (L. Tong), jakeluo@uwm.edu (J. Luo), yuz@uwm.edu (Z. Yu).

https://doi.org/10.1016/j.bspc.2021.102931
Received 30 June 2020; Received in revised form 7 June 2021; Accepted 24 June 2021
Available online 6 July 2021

presence of malignant tissue. The increasing number of cancer incidence results in countless laboratory tests, which often overwhelms pathologists. Meanwhile, patient-specific treatment options make diagnosis and treatment of cancer more complex than ever before [6].

To address these limitations, automatic analysis of microscopic image has been the center of cancer diagnosis in recent years [5]. Computer-Aided Detection (CAD) technology offers a solution for radiologists and pathologists to automatically detect malignancies based on biopsy images. However, the application of CAD is not practical before the 2000s because of relatively low detection accuracies [7]. The poor performance of made clinical implementation impractical, until the recent advances in computerized image detection [8].

Recent advances enabled the trend of turning histological slides into digital image datasets, in which machine learning can intervene on digital images to address the limitation of inaccurate diagnosis. In 2017, the US Food and Drug Administration (FDA) announced the approval of the first whole slide imaging (WSI) system for primary diagnosis in surgical pathology [9]. With the advent of WSI, digital pathology has become a part of the routine procedure in clinical diagnosis yet leaving new questions and possibilities.

Due to the rise of cancer incidence and patient-specific treatment options, diagnosis, and treatment of cancer are becoming more complex [10]. Pathologists must spend an extremely long time examining a large number of slides; Therefore, detecting the nuances of histological images can be difficult [11]. The 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. 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 [12]. CAD technology that integrates powerful algorithms, such as deep learning algorithm, that is able to accurately recognize tumor malignancy.

The primary goals of this study are:

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

(2) To explore a suitable deep learning framework for accurate detection, and discover possible features that contribute to performance.

To achieve these primary goals, histological medical image analysis based on transfer learning was applied to the pathology archives at Children's Medical Center dataset [13]. Two modified transfer learning approaches including VGG 19 [14] and Inception V3 [15] models were applied to the data. The novelty of the model is applying the models to different categories of the dataset and using the whole tile image as input.

## 2. Literature review

CAD technology offers a solution for radiologists and pathologists to automatically detect malignancies [7]. This solution becomes feasible since 2010 thanks to increasing diagnosing accuracy [16]. Remarkable progress has been achieved in medical images, primarily due to the availability of large-scale datasets and machine learning algorithms for pattern recognition in the computer science area [17]. We will discuss the progress in two aspects of advances of computer-aided technology in disease detection: (1) Type of diseases; (2) Detecting algorithms.

### 2.1. Type of diseases related to CAD: Tumor-based and non-tumor-based diseases

Diseases diagnosed based on radiological and histological images often get more help in Computer-Aided technologies. Although all diseases are possible with the help of CAD technology, this technique is specialized in pattern recognition in images. Therefore, CAD technology yields its own strength in interpreting medical images in both the radiological department and laboratory tests. It has relatively high accuracy in both radiological images and histological images.

A wide range of disease-related to these two departments have the potential of applying CAD technology. For example, CAD technology has been widely applied to a variety of medical images for the detection of different diseases. The most common target is different types of tumor recognition, such as breast cancer [18–23], gastric cancer [24–26], skin cancer [27], lung cancer[28,29], brain tumor [30,31], prostate cancer [32], osteosarcoma [33,34].

To detect the malignancy or presence of tumor, reviewing an image-based document is a key step for pathologists to confirm its diagnosis [7]. The CAD technology is often assigned the task of differentiating nuances between images and find unique patterns, thereby classifying malignant tumors from normal images. A major part of current studies focuses on the tumor-based diagnosis, which generates a number of representative training data. The abundance of training data fosters related research, making the detecting accuracy over 90% on most malignant tumors.

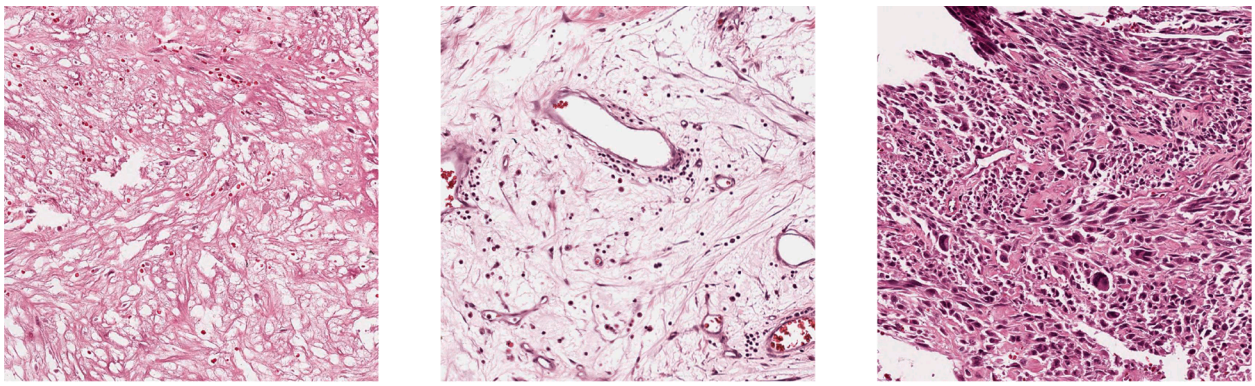
Other non-tumor-related diseases, which often involve the X-ray or histological test, also employ similar technology. Common examples are chest X-ray pneumonia [35], pulmonary edema [36], pulmonary fibrosis [37], gastric endoscopic images for celiac diseases [38], and diabetic retinopathy [39]. However, the lack of representative data restricts the development of more accurate performance. Also, the task of image pattern recognition has more variety than tumor cell recognition, therefore, yielding more challenging tasks. Therefore, the study of non-tumor-related diseases falls behind compared to tumor-related diseases in image classification.

### 2.2. Type of detection algorithms in CAD: Machine learning and deep learning

Advances in both medical images and computers have led to a rise of artificial intelligence in various imaging tasks. The machine-learning-based methods are one of the most important algorithms in a number of techniques in image detection and diagnosis [7]. Deep learning, as a new type of algorithm, is able to achieve higher performance in medical diagnosis and detection. We will discuss the advantages and disadvantages of the two algorithms.

#### 2.2.1. Machine learning

Machine learning algorithms use computer-extracted features or are called learning materials. Various machine learning techniques have been applied in the past, for example, linear discriminant analysis, support vector machines, decision trees, and random forests, and neural networks [40]. For images, the data is often encoded with RGB-encoding schemas, and sometimes with additional features. With appropriate features, the important histological and radiological image information can be integrated into the above algorithms. One significant drawback of machine learning is that it requires handcrafted features. In other words, appropriate feature selection techniques are important and necessary to achieve a great performance. A number of studies have been conducted to explore appropriate feature selection techniques [41,42]. Such analyses have considered both performance and applicability for other similar studies. That is, a computer-derived tumor signature needs to both perform well in its specific task and be generalizable to other cases. Some studies [43,33,34] have achieved a relatively good result in identifying osteosarcoma malignancies, by carefully constructing



(a) Necrotic Tumor (b) Non-Tumor (c) Viable Tumor

**Fig. 1.** Sample images from the dataset.

images to machine learning framework and selecting appropriate features.

#### 2.2.2. Deep learning

Deep learning is a sub-category of machine learning in which multiple-layered networks are used to assess complex patterns within the raw imaging input data [44]. A comprehensive technical review of deep learning structures in medical image analysis is given by Shen et al. [45]. The biggest difference between deep learning and machine learning is the capability of automatic feature selection. Also, a multi-layer structure makes deep learning more suitable for non-linear classification tasks. Therefore, deep learning has theoretically higher performance in classifying tasks.

A number of recent studies have used deep Convolutional Neural Networks (CNNS), as a major enhancement in histological image detection. In 2017, Jongwon et al. [46] did a pilot study on histopathology of breast cancer, which achieves an Area Under the ROC Curve (AUC) value of 0.93 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 et al.'s result [47] shows the state-of-the-art performance using VGG16 and AlexNet models, with an average of 90.961.59% accuracies. This also indicates the suitability of these models for image classification tasks.

### 2.3. Transfer learning: Important technique in deep learning

Deep learning is high in complexity, which also requires a larger number of datasets than traditional machine learning algorithms. However, the lack of histological and radiological images often restricts its development. A typical deep learning model that addresses image classification tasks requires three steps: model building, training on a dataset, and evaluation of performance on specific tasks. The first two steps require additional attention: Firstly, the model building process requires an appropriate network structure; Secondly, a large data set, especially on histological images, is very difficult to find. The transfer learning aims to address these two limitations [48]. A pre-trained model is a saved network that was previously trained on a large dataset, typically on a large-scale image classification task. You either use the pre-trained model as is or use transfer learning to customize this model to a given task. The intuition behind transfer learning for image classification is that if a model is trained on a large and general dataset, this model will effectively serve as a generic model of the visual world. You can then take advantage of these learned feature maps without having to start from scratch by training a large model on a large dataset. In machine learning and deep learning 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 by training our deep learning model from pre-trained networks and apply the parameters to data sets from other domains. In these situations, knowledge-transferring significantly improves learning outputs if done efficiently while minimizing expensive data labeling efforts [49].A few studies have already focused on this area in medical image classifications: De Matos et al. [21] used double transfer learning to classify histopathologic images. Noorul Wahab et al. [22] aimed at a more challenging task of segmentation and detection of mitotic nuclei. They used a similar hybrid CNN model and achieved a 76% AUC value. Other studies include the prediction of pathological invasiveness in lung adenocarcinoma [50], classification of liver cancer histopathology images [51], automated invasive ductal carcinoma detection [52], and skin cancers [27]. In [53], the authors reported 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 tumors, and 92.7% for necrotic tumors. For osteosarcoma studies, researchers employed deep learning techniques [54,55,43] focusing on segmentation and classification of histology tissue in tumor image datasets. A multiple-layered neural network is proficient in image segmentation, as they achieved significantly better performances than machine learning algorithms. In clinical practices, the primary goal is to decrease the mortality of osteosarcoma diagnosis. 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. The adoption of computer-aided technology using CNN can significantly reduce the surgeon's workload and achieve a better prognosis of patients.

## 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 $\times$ 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

**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 |

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 Fig. 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 an image data generator module of "Keras"[56]. 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.

### 3.3. Model Selection

There are 26 deep learning models in Keras Applications that can be used for prediction, feature extraction, and fine-tuning [56]. 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.

From Table 1, we can see that ResNet50 gives the inferior result among all the models. The second most inferior result is given by DenseNet201. Both networks have very deep layers and complex architecture. With the small number of images in our selected dataset, these networks are underfitting during the training process which is reflected in the testing results. InceptionV3 and NASNetLarge give almost the same results for all performance metrics (precision, recall, f1-score, and accuracy). Although they show almost 80% accuracy, they are still far away from the accuracy given by VGG16 and VGG19 networks. VGG16 and VGG19 are the simplest networks among the selected networks, and they both perform quite well with our selected dataset. As VGG19 is an extension of the VGG16 network, later we choose the InceptionV3 network for result comparison with the best model (VGG19) due to their divergent network architecture.

### 3.4. VGG19 model

We have used Keras applications for importing the VGG19 model. Pre-trained weights have been used for model training. We have discarded the fully connected layer along with the output layer of the VGG19 model. Two fully connected layers have been added after the last "maxpool" layer. Convolutional layers are used for feature extraction where fully connected layers are used for classification. These layers learn a non-linear function between the high-level features given as an output from the previous (convolutional) layers. As the dataset is not very big and a maximum of 3-class classification is performed, adding more fully connected layer(s) does not improve the model performance, but instead increases the training time. Dropout layers are used for avoiding over-fitting the training data. We have used "ReLU" (Rectified Linear Unit) activation in the dense layers and the "softmax" activation function in the output layer. ReLU is computationally efficient than the "sigmoid" and "tanh" functions, as it does not need to perform expensive exponential operations. Also, ReLU solves the vanishing gradient problem, as the gradient is either 0 or 1 for this function and it never saturates which means the gradients cannot vanish and be transferred perfectly across the network. The "Softmax" activation function is generally used for multiclassification and its output is a probability distribution, which means the output is mapped to the range of [0,1] and the sum of the total output is 1. Fig. 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 max-pooling layers use 2 × 2 filters. The FC1 and FC2 layers contain 512 and 1024 neurons respectively. Softmax layer's neurons vary 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 an Nvidia GeForce RTX 2080Ti GPU platform.

The loss functions used for binary classification and multiclass

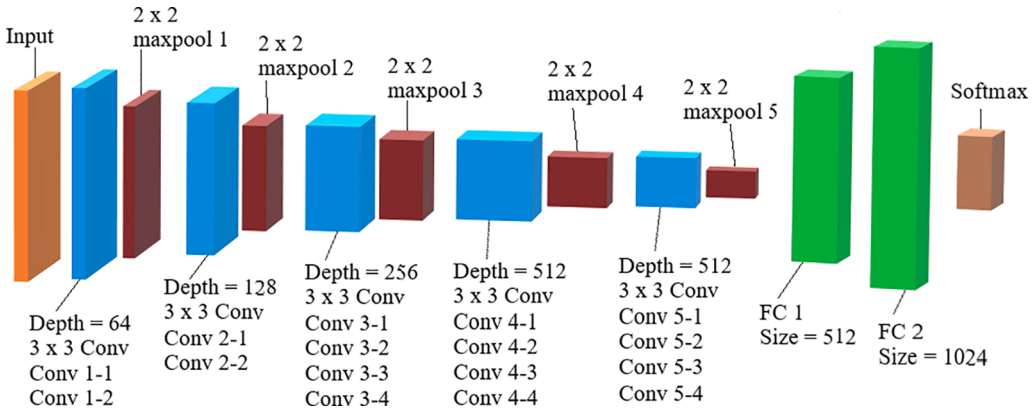



**Fig. 2.** VGG19 Network architecture.

**Table 2**
Summary of all the parameters and hyperparameters used for each model. (Opt: Optimization, lr: learning rate).

| Model | Classification type | Loss function | Opt | lr | Batch Size | | |
|---|---|---|---|---|---|---|---|
| | | | | | Train | Validation | Test |
| VGG19 | Binary | Binary cross-entropy | Adam | 0.01 | 80 | 28 | 16 |
| | Multiclass | Categorical cross-entropy | Adam | 0.01 | 80 | 28 | 16 |
| InceptionV3 | Binary | Binary cross-entropy | Adam | 0.01 | 80 | 28 | 16 |
| | Multiclass | Categorical cross-entropy | Adam | 0.01 | 80 | 28 | 16 |

| 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

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

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, validation, and testing respectively. All models are trained for 1500 epochs, with a callback that stops training when validation accuracy reaches over 0.98. Table 2 shows a summary of hyperparameter tuning used for the model.

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 a confusion matrix, precision, recall, f1 score, and accuracy for all classifications. We also reported the 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 the false positive rate (1-specificity). The area under the curve (AUC) indicates that the classifier gives a randomly chosen positive instance a 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. Fig. 3 shows the confusion matrix for all classifications with both networks.

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

**Table 3**
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 |

## 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 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[14,15]. 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 Fig. 3, we can see that for NT vs VT and NCT vs VT respectively the prediction of non-tumor and the necrotic tumor is performed well by Inception V3. In all other cases, VGG19 works very robustly compared to Inception V3. So, in overall balance, VGG19 beats Inception V3.

From Tables 3 and 4, 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 the VGG19 model. Since the 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 a lower score of the other one. Hence, in balance in overall performance, VGG19 beats inception V3 by a huge margin. From Fig. 4, it is clear that for all classifications VGG19 achieves the highest accuracy.

From Fig. 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

**Table 4**
Precision, Recall, and F1-Score for multicalss classification.

| 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 |

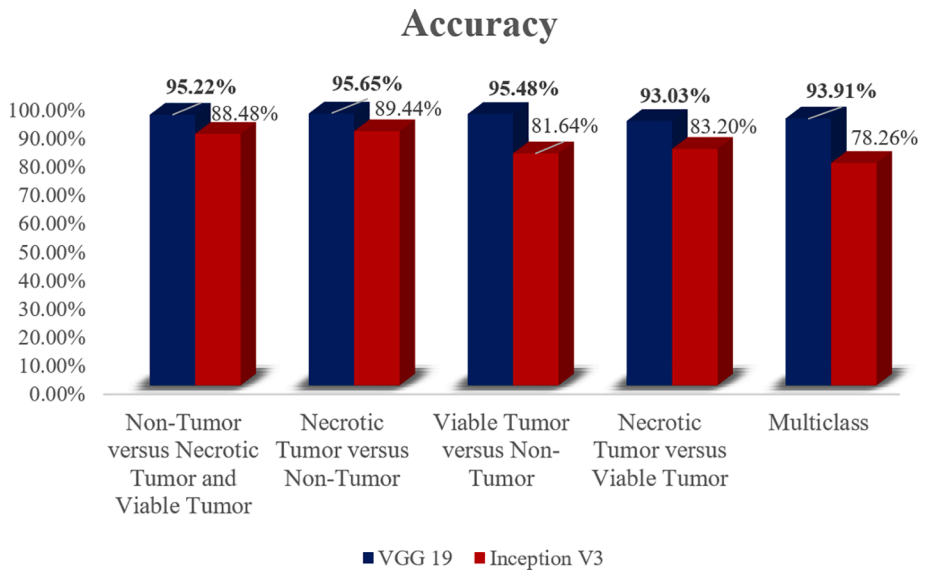



**Fig. 4.** Accuracy scores.

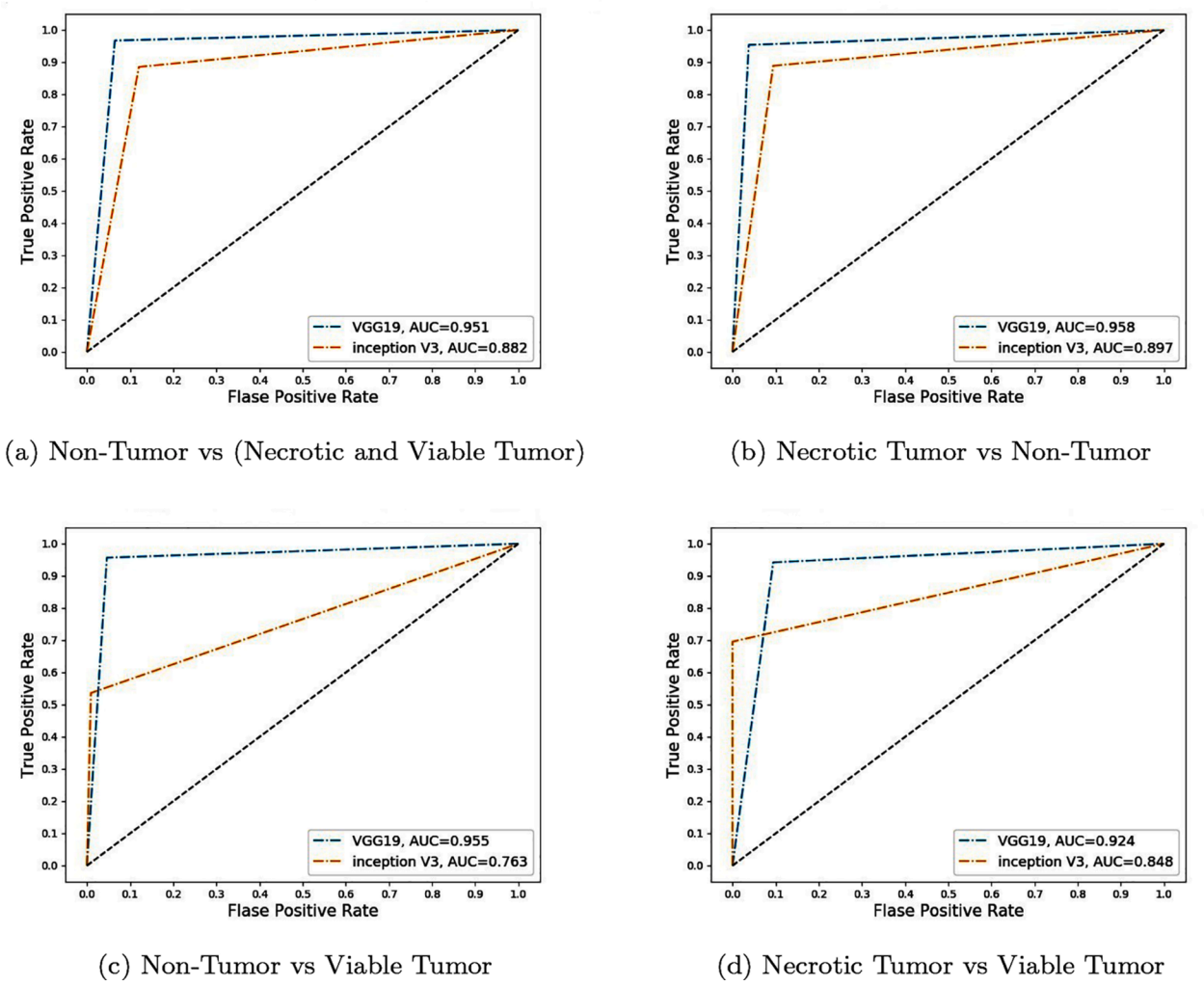



**Fig. 5.** ROC and AUC of all two-class classifications.

analytical discussions, it is safe to say that VGG19 works well for all classifications. While Inception V3 has three types of convolutions ($1 \times 1, 3 \times 3, 5 \times 5$), VGG19 has only one type of convolution ($3 \times 3$). 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 ($3 \times 3$); 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 the input and the corresponding output. To overcome the data limitation problem, we applied the 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 the 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 has been used in VGG19 and Inception architecture in Deep learning to recognize 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

**Table 5**
Result comparison.

| Tumor type | Tile accuracy in % | |
|---|---|---|
| | VGG19 | Arunachalam [57]'s deep learning model |
| Non-Tumor | 95.45 | 89.5 |
| Necrotic Tumor | 94.34 | 91.5 |
| Viable Tumor | 94.26 | 92.6 |

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 5 represents the comparison of these two works. We have done a binary classification for all possible combinations between three classes, where Arunachalam et al. [57]'s deep learning model provides a direct class-specific accuracy. Therefore Table 5 represents our average accuracy for a specific tumor class. For viable tumor the average of VT vs NT and NCT vs VT; for a 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 [57]), as we have used the 1144 whole images for our classification. Table 5 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.

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 the 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.

## CRediT authorship contribution statement

**D.M. Anisuzzaman:** Conceptualization, Methodology, Software, Writing - original draft, Data curation. **Hosein Barzekar:** Conceptualization, Methodology, Software, Writing - original draft. **Ling Tong:** Writing - original draft. **Jake Luo:** Validation, Writing - review & editing. **Zeyun Yu:** Supervision, Validation, Writing - review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] A.J. Chou, D.S. Geller, R. Gorlick, Therapy for osteosarcoma, Pediatric Drugs 10 (5) (2008) 315–327.
[2] C.A. Arndt, W.M. Crist, Common musculoskeletal tumors of childhood and adolescence, New England Journal of Medicine 341 (5) (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 (6) (2002) 1123.
[5] D.S. Geller, R. Gorlick, Osteosarcoma: a review of diagnosis, management, and treatment strategies, Clinical Advances in Hematology & Oncology 8 (10) (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 (9) (2019) 1686–1698.
[7] R.A. Castellino, Computer aided detection (CAD): an overview, Cancer Imaging 5 (1) (2005) 17.
[8] A. Madabhushi, G. Lee, Image analysis and machine learning in digital pathology: Challenges and opportunities, Medical Image Analysis 33 (2016) 170–175. ISSN 1361-8415, 20th anniversary of the Medical Image Analysis journal (MedIA).
[9] A.J. Evans, T.W. Bauer, M.M. Bui, T.C. Cornish, H. Duncan, E.F. Glassy, J. Hipp, R. S. McGee, D. Murphy, C. Myers, et al., US Food and Drug Administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised, Archives of Pathology & Laboratory Medicine 142 (11) (2018) 1383–1387.
[10] 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 (3) (2019) 216–233.
[11] P. Picci, Osteosarcoma (osteogenic sarcoma), Orphanet Journal of Rare Diseases 2 (1) (2007) 6.
[12] 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.
[13] T. The Cancer Imaging Archive, Osteosarcoma data from UT Southwestern UT Dallas for Viable and Necrotic Tumor Assessment, URL: https://doi.org/10.7937/tcia.2019.bvhjhdas., 2019.
[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
[15] 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.
[16] 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 (5) (2016) 1285–1298.
[17] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Computation 29 (9) (2017) 2352–2449.
[18] S.S. Aboutalib, A.A. Mohamed, W.A. Berg, M.L. Zuley, J.H. Sumkin, S. Wu, Deep learning to distinguish recalled but benign mammography images in breast cancer screening, Clinical Cancer Research 24 (23) (2018) 5902–5909.
[19] 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), 2017, pp. 1–4, doi: 10.1109/HealthCom.2017.8210843.
[20] 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 (1) (2018) 18.
[21] J. d. Matos, A. d. S. Britto, L.E.S. Oliveira, A.L. Koerich, Double transfer learning for breast cancer histopathologic image classification, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852092.
[22] 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 (3) (2019) 216–233.
[23] F. Gao, T. Wu, J. Li, B. Zheng, L. Ruan, D. Shang, B. Patel, SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis, Computerized Medical Imaging and Graphics 70 (2018) 53–62.
[24] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, et al., Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images, Gastric Cancer 21 (4) (2018) 653–660.
[25] Y. Li, X. Li, X. Xie, L. Shen, Deep learning based gastric cancer identification, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 182–185, doi: 10.1109/ISBI.2018.8363550.
[26] H. Sharma, N. Zerbe, I. Klempert, O. Hellwich, P. Hufnagl, Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology, Computerized Medical Imaging and Graphics 61 (2017) 2–13.
[27] K.M. Hosny, M.A. Kassem, M.M. Foaud, Skin cancer classification using deep learning and transfer learning, in: 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), IEEE, 2018, pp. 90–93.
[28] S. Lakshmanaprabu, S.N. Mohanty, K. Shankar, N. Arunkumar, G. Ramirez, Optimal deep learning model for classification of lung cancer on CT images, Future Generation Computer Systems 92 (2019) 374–382.
[29] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning, Nature Medicine 24 (10) (2018) 1559–1567.
[30] P. Sapra, R. Singh, S. Khurana, Brain tumor detection using neural network, International Journal of Science and Modern Engineering (IJISME) ISSN (2013) 2319–6386.
[31] M.-N. Wu, C.-C. Lin, C.-C. Chang, Brain tumor detection using color-based k-means clustering segmentation, in: Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007), vol. 2, IEEE, 2007, pp. 245–250.
[32] S. Yoo, I. Gujrathi, M.A. Haider, F. Khalvati, Prostate cancer detection using deep convolutional neural networks, Scientific reports 9 (1) (2019) 1–10.
[33] R. Shen, Z. Li, L. Zhang, Y. Hua, M. Mao, Z. Li, Z. Cai, Y. Qiu, J. Gryak, K. Najarian, Osteosarcoma Patients Classification Using Plain X-Rays and Metabolomic Data, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 690–693.
[34] Z. Li, S.R. Soroushmehr, Y. Hua, M. Mao, Y. Qiu, K. Najarian, Classifying osteosarcoma patients using machine learning approaches, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 82–85.
[35] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225.
[36] V.E. Noble, L. Lamhaut, R. Capp, N. Bosson, A. Liteplo, J.-S. Marx, P. Carli, Evaluation of a thoracic ultrasound training module for the detection of

pneumothorax and pulmonary edema by prehospital physician care providers, BMC Medical Education 9 (1) (2009) 1–5.
[37] A. Christe, A.A. Peters, D. Drakopoulos, J.T. Heverhagen, T. Geiser, T. Stathopoulou, S. Christodoulidis, M. Anthimopoulos, S.G. Mougiakakou, L. Ebner, Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images, Investigative Radiology 54 (10) (2019) 627.
[38] J.H. Lee, Y.J. Kim, Y.W. Kim, S. Park, Y.-I. Choi, Y.J. Kim, D.K. Park, K.G. Kim, J.-W. Chung, Spotting malignancies from gastric endoscopic images using deep learning, Surgical Endoscopy 33 (11) (2019) 3790–3797.
[39] S. Wan, Y. Liang, Y. Zhang, Deep convolutional neural networks for diabetic retinopathy detection by image classification, Computers & Electrical Engineering 72 (2018) 274–282.
[40] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H.J. Aerts, Machine learning methods for quantitative radiomic biomarkers, Scientific Reports 5 (1) (2015) 1–11.
[41] A.R. Jamieson, M.L. Giger, K. Drukker, L.L. Pesce, Enhancement of breast CADx with unlabeled data a, Medical Physics 37 (8) (2010) 4155–4172.
[42] A.R. Jamieson, M.L. Giger, K. Drukker, H. Li, Y. Yuan, N. Bhooshan, Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and-SNE, Medical Physics 37 (1) (2010) 339–351.
[43] L. Huang, W. Xia, B. Zhang, B. Qiu, X. Gao, MSFCN-multiple supervised fully convolutional networks for the osteosarcoma segmentation of CT images, Computer Methods and Programs in Biomedicine 143 (2017) 67–74.
[44] Q. Li, W. Cai, X. Wang, Y. Zhou, D.D. Feng, M. Chen, Medical image classification with convolutional neural network, in: 2014 13th international conference on control automation robotics & vision (ICARCV), IEEE, 2014, pp. 844–848.
[45] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annual Review of Biomedical Engineering 19 (2017) 221–248.
[46] 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.
[47] 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 (1) (2018) 1–7.
[48] M. Shaha, M. Pawar, Transfer learning for image classification, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2018, pp. 656–660.
[49] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2009) 1345–1359.
[50] 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 (25).
[51] 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 (6) (2019) 1643–1651.
[52] 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.
[53] 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 (4) (2019), e0210706.
[54] R. Mishra, O. Daescu, P. Leavey, D. Rakheja, A. Sengupta, Convolutional neural network for histopathological analysis of osteosarcoma, Journal of Computational Biology 25 (3) (2018) 313–325.
[55] R. Zhang, L. Huang, W. Xia, B. Zhang, B. Qiu, X. Gao, Multiple supervised residual network for osteosarcoma segmentation in CT images, Computerized Medical Imaging and Graphics 63 (2018) 1–8.
[56] F. Chollet, et al., Keras, URL: https://github.com/fchollet/keras, 2015.
[57] 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 (4) (2019), e0210706.