University of Louisville
ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

12-2021

Lung nodules identification in CT scans using multiple instance learning.

Wiem Safta
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd

 Part of the Other Computer Engineering Commons

Recommended Citation

Safta, Wiem, "Lung nodules identification in CT scans using multiple instance learning." (2021). *Electronic Theses and Dissertations.* Paper 3793.
Retrieved from https://ir.library.louisville.edu/etd/3793

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

LUNG NODULES IDENTIFICATION IN CT SCANS USING MULTIPLE INSTANCE LEARNING

By

Wiem Safta
B.E. Telecommunications Engineer, Higher School of Communication of Tunis, 2014

A Dissertation
Submitted to the Faculty of the
J. B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Computer Science and Engineering

Department of Computer Science and Engineering
University of Louisville
Louisville, Kentucky

December 2021

LUNG NODULES IDENTIFICATION IN CT SCANS USING MULTIPLE INSTANCE LEARNING

By

Wiem Safta
B.E. Telecommunications Engineer, Higher School of Communication of Tunis, 2014

A Dissertation Approved on

November 22, 2021

---

by the Following Dissertation Committee:

Hichem Frigui, Ph.D, Dissertation Director,

Amir Amini, Ph.D.

Juw Won Park, Ph.D.

Nihat Altiparmak, Ph.D.

Olfa Nasraoui, Ph.D.

DEDICATION

This dissertation is dedicated to my Father's soul. Dear Dad, you always encouraged me to pursue my studies with passion and perseverance, so I hope you are proud of me wherever you are.

ACKNOWLEDGMENTS

All deepest thanks are due to GOD ALMIGHTY for answering my prayers and enabling me to accomplish this thesis.

This thesis wouldn't have been possible without the support of many people. I want to express my sincere gratitude to my advisor Prof. Hichem Frigui for giving me the precious opportunity to carry out my Ph.D. thesis under his guidance. His intellectual creativity and analysis have been of great inspiration for my work. Without his continuous support and great insights, the accomplishment of this Ph.D. wouldn't have been possible.

I also would like to thank Prof. Amir Amini. His knowledge about medical imagining and tremendous help in understanding some basic concepts have been essential for accomplishing this work.

I address, likewise, my deepest gratitude to Prof. Olfa Nasraoui, Prof. Nihat Altiparmak, and Prof. Juw Won Park for being on my dissertation committee with enthusiasm and interest in my research despite their other responsibilities and commitments.

I would also like to thank my colleague Ben Veasey for getting and processing the needed data to accomplish this work.

I am grateful to my colleague, Ali shahrjoo for his help, encouragement and advises.

I want to thank my colleagues in the Multiple Research Lab for their help and support.

I also would like to thank my husband, Ahmed Shaffie. His support, help, encouragement, and love made this journey worth living. Thank you for being the best part of me.

Thanks to my friends Fadoua khmaissia and Pegah Saghab, for standing by my side during this process and encouraging me to hold on whenever it seemed too hard to continue.

Last but not least, I would like to sincerely thank all my family members: my mother Taouhida Bouzgarrou and my brothers Mohammad Nidhal Safta and Mohammad Houssaine Safta, for their incredible patience, encouragement, and unconditional love.

ABSTRACT

LUNG NODULES IDENTIFICATION IN CT SCANS USING MULTIPLE INSTANCE LEARNING

Wiem Safta

November 22, 2021

Computer Aided Diagnosis (CAD) systems for lung nodules diagnosis aim to classify nodules into benign or malignant based on images obtained from diverse imaging modalities such as Computer Tomography (CT). Automated CAD systems are important in medical domain applications as they assist radiologists in the time-consuming and labor-intensive diagnosis process. However, most available methods require a large collection of nodules that are segmented and annotated by radiologists. This process is labor-intensive and hard to scale to very large datasets. More recently, some CAD systems that are based on deep learning have emerged. These algorithms do not require the nodules to be segmented, and radiologists need to only provide the center of mass of each nodule. The training image patches are then extracted from volumes of fixed-sized centered at the provided nodule's center. However, since the size of nodules can vary significantly, one fixed size volume may not represent all nodules effectively.

This thesis proposes a Multiple Instance Learning (MIL) approach to address the above limitations. In MIL, each nodule is represented by a nested sequence of volumes centered at the identified center of the nodule. We extract one feature vector from each volume. The set of features for each nodule are combined and represented by a bag. Next, we investigate and adapt some existing algorithms and develop new ones for this application.

We start by applying benchmark MIL algorithms to traditional Gray Level Co-occurrence Matrix (GLCM) engineered features. Then, we design and train simple Convolutional Neural Networks (CNNs) to learn and extract features that characterize lung nodules. These extracted features are then fed to a benchmark MIL algorithm to learn a classification model. Finally, we develop new algorithms (MIL-CNN) that combine feature learning and multiple instance classification in a single network. These algorithms generalize the CNN architecture to multiple instance data.

We design and report the results of three experiments applied on both generative (GLCM) and learned (CNN) features using two datasets (The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [1] and the National Lung Screening Trial (NLST) [2]). Two of these experiments perform five-fold cross-validations on the same dataset (NLST or LIDC). The third experiment trains the algorithms on one collection (NLST dataset) and tests it on the other (LIDC dataset). We designed our experiments to compare the different features, compare MIL versus Single Instance Learning (SIL) where a single feature vector represents a nodule, and compare our proposed end-to-end MIL approaches to existing benchmark MIL methods. We demonstrate that our proposed MIL-CNN frameworks are more accurate for the lung nodules diagnosis task. We also show that MIL representation achieves better results than SIL applied on the ground truth region of each nodule.

TABLE OF CONTENTS

LIST OF TABLES

TABLE. PAGE

LIST OF FIGURES

CHAPTER I
INTRODUCTION

Lung cancer is the leading cause of death in US with a percentage of 24% for men and 23% for women of estimated related cancer deaths in 2019 [3]. Low-Dose Computed Tomography (LDCT) screening was proven to be adequate in diminishing cancer related deaths by 20% as it assists radiologists to detect lung cancer at an early stage [2]. However this task is still time consuming and rigorous as lung nodules are sometimes similar in shape and texture whether they are benign or malignant. These challenges led researchers toward developing automated Computer Aided Diagnosis (CAD) systems with minimum user intervention to assist radiologists in the diagnosis and prediction process.

Current CAD systems can be divided into two major categories. The first one is based on the traditional learning paradigm and consists of two sequential steps. First salient features such as Gray Level Co-Occurrence Matrix (GLCM), Grey-Level Run Length Matrix (GLRLM), Gray-Level Size-Zone Matrix (GLSZM) [4] are extracted from the specified regions of interest. Second, a classification algorithm such as Support-Vector Machine (SVM) [5] or Linear Discriminant Analysis (LDA) [4], is then used to label the region as benign or malignant. This category of CAD systems has proved to be efficient in lung nodules diagnosis [4, 6–9]. However it relies on a large collection of training samples. These samples need to be segmented and annotated by radiologists. This process is labor intensive and hard to scale to very large datasets.

The second category of CAD systems is based on deep learning algorithms. This approach combines the feature extraction and classification into a single task. Sample deep learning algorithms that have been applied to this application include 2D Convolutional Neural Networks (CNN) [10], 3D CNN [11] and multiview 2D CNN [12]. Other hybrid

CNN versions have been developed in more recent works [5, 13].

CNN based methods have proved to be more successful and useful in the lung nodules classification task for two main reasons. First, instead of relying on hand-crafted features, they learn them during the training phase. Second, data labeling is less rigorous as only the center of the nodule needs to be identified and no nodule segmentation is needed. One potent drawback of CNN methods is that they process a volume with fixed size for all nodules. This may not be optimal as the size of nodules can vary significantly. Figures 1 and 2 display sample nodules with sizes that vary significantly from one sample to another.

One way to adapt learning algorithms to objects with varying sizes, without annotating the bounding box of each object, is to have multiple instances representing each object with different sizes and use Multiple Instance Learning (MIL) algorithms.

MIL was proposed by Dietrich et al. in 1997 for drug activity prediction [14]. Since then, it has been used in many domains such as computer security [15], image retrieval [16], image categorization [17], face detection [18], text categorization [19] and computer aided diagnosis. For instance, for medical applications the authors in [20] used MIL for breast cancer diagnosis and in [21] for classification of diabetic retinopathy images.

In traditional learning, also called Single Instance Learning (SIL), every sample is represented by a single feature vector. On the other hand, in MIL, each sample is represented by multiple feature vectors. Each feature is referred to as instance and the collection of instances representing one object is called a bag. Most MIL algorithms are defined for two class problems: positive class and negative class. Early MIL algorithms characterize the positive class by a single target concept. A bag is labeled positive if at least one of its instances belongs to the target concept and labeled negative if none of its instances are close to the target concept. In this approach, instances within a bag are assumed to be independent and to have identical distribution. Thus, any potential relations among the instances are ignored. Another category of MIL algorithms, refereed to as generalized MIL [22], allows for multiple target concepts. Here, a bag is positive only if all target concepts are

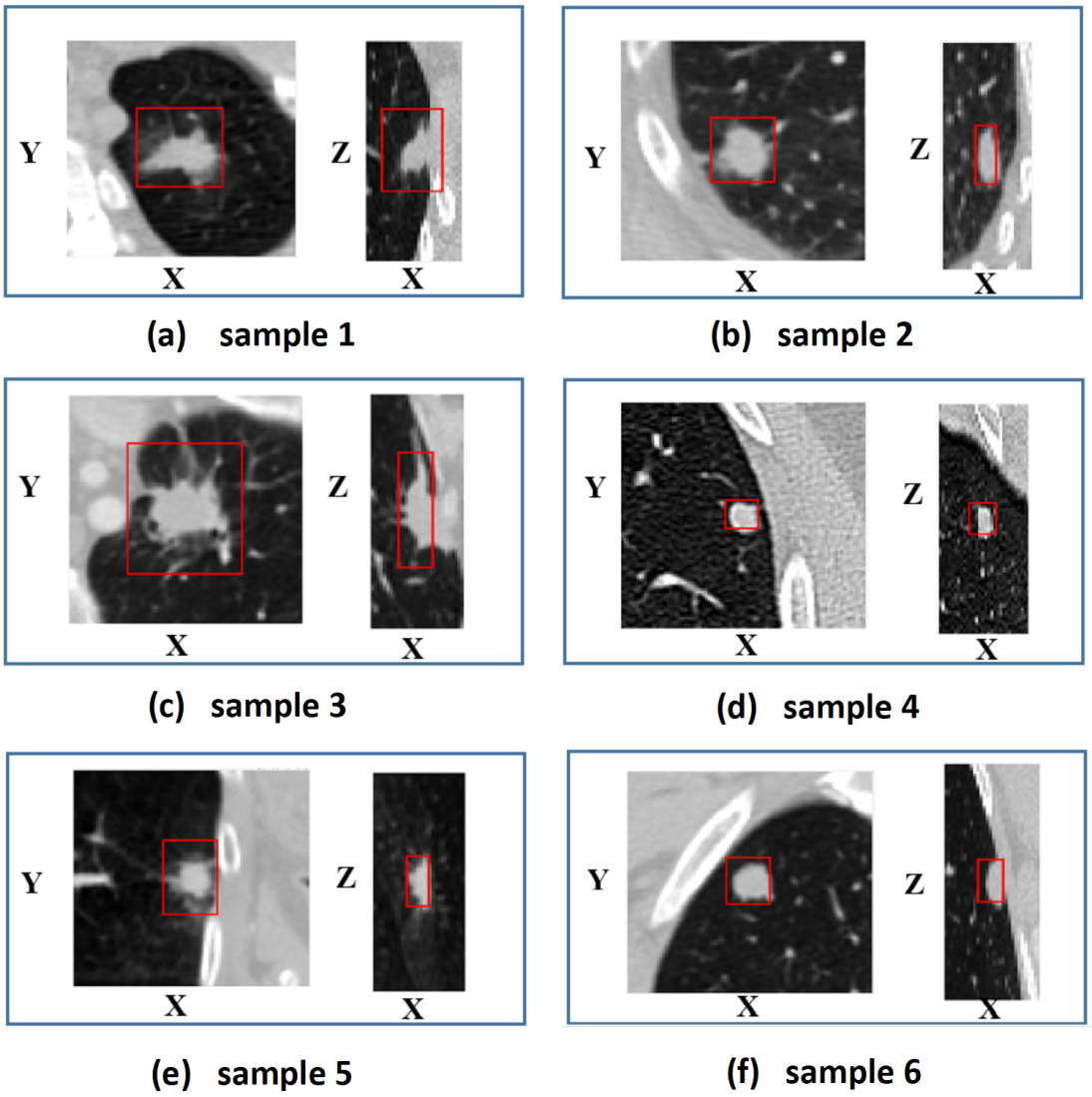


FIGURE 1: 2D views of six malignant nodules identified by radiologists with significant variations in size. The rectangles refer to the ground truth regions.

verified (i.e., the bag has instances from each concept). For both categories of MIL algorithms, the general idea is to learn target concepts/criteria from training samples. These concepts are then used to predict the label of instances within the bag and to predict the label of a new test bag.

In this thesis, we propose to investigate and analyse the application of MIL to the lung nodules diagnosis task. We propose representing each nodule by a bag of instances.

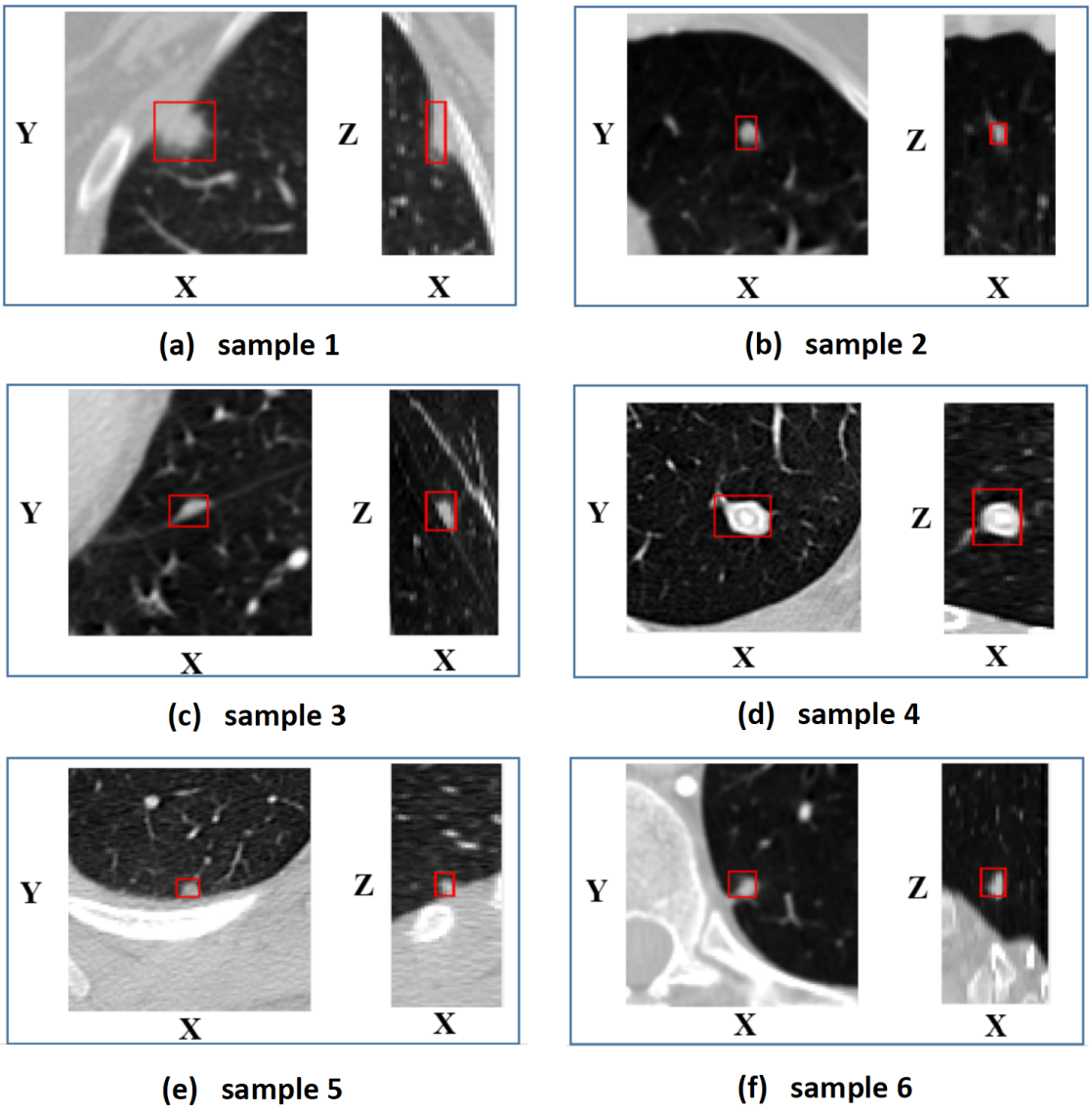


FIGURE 2: 2D views of six benign nodules identified by radiologists with significant variations in size. The rectangles refer to the ground truth regions.

Each instance corresponds to a feature vector extracted from a different volume. For instance, instead of representing samples 3 and 4 in Figure 1 (c) and Figure 1 (d) by one fixed volume, we represent them by a nested sequence of volumes (instances). We postulate that the MIL algorithm will learn to use one (or a few) of the large volumes to represent sample 3 and to use one (or a few) of the small volumes to represent sample 4. A small cube will

be suitable to represent small nodules and the large ones will be more suitable for large nodules. The collection of cubes of each nodule will form a bag and the feature representative of each bag, in the training data, will be labeled as benign or malignant according to the available ground truth, but the labels of individual instances (i.e., the true volume) are unknown.

We consider various feature representation methods and various characterization algorithms. We analyze and compare the results and identify the optimal setting for the considered application. Our first approach is based on traditional features (i.e., GLCM features) extraction methods followed by traditional MIL classification algorithms. We propose and evaluate multiple GLCM+MIL frameworks.

Our second approach is based on a nodule identification algorithm called CNN+MIL that combines CNN features and MIL techniques to take advantages of their strengths. First, a CNN is trained and then used to extract features for each instance within the bag. Then, an MIL algorithm is used to represent each object by a bag of instances (CNN features) and to learn a classification model. We propose and evaluate few CNN+MIL variations.

Our third approach is based on methods that combine CNN and MIL into a single end-to-end network. We refer to this approach as MIL-CNN. We develop and design four MIL-CNN algorithms.

Our research and experiments are designed to answer the following Research Questions :

- ***RQ1***: What is the best MIL algorithm for this application ?
- ***RQ2***: What are the main nodule parameters that influence the classification ?
- ***RQ3***: What is the best feature representation ?
- ***RQ4***: Can an MIL algorithm learn to identify the positive instances (among a bag with large number of candidates) that are close to the ground truth ?

- ***RQ5***: What is the best learning approach for this application: Single Instance Learning (SIL) applied to the ground truth region or Multiple Instance Learning (MIL) applied to multiple regions around the ground truth ?

The remainder of this thesis is structured as follows: Chapter II reviews previous work conducted in the lung nodules classification task and MIL algorithms. Chapter III describes our lung nodules classification framework based on engineered features and Multiple Instance Learning, our lung nodules classification framework based on CNN features and Multiple Instance Learning, and our lung nodules classifications based on end-to-end Multiple Instances Convolutional Neural Networks (MIL-CNN). We introduce the way we process our dataset, our used datasets to validate our results and outline our Research Questions in chapter IV. In chapter V, we present and analyze our experimental results and answer the outlined research questions. Finally, in chapter VI, we conclude by summarizing our work and outlining potential future work.

CHAPTER II
RELATED WORK

In this chapter, we review existing approaches in areas that are highly relevant to our proposed Computer Aided Diagnosis (CAD) system for lung nodules classification along with traditional Multiple Instance Learning (MIL) models. We start with outlining different categories of CAD systems. Then, we outline few MIL algorithms that we adapt to our application.

Lung nodules diagnosis is typically the last component of the CAD system and one of the most crucial ones. It is a two-class classification task that classifies lung nodules into malignant or benign after segmenting them from the lung regions to assist radiologists in the diagnosis process. Many CAD systems have been developed during the past few decades and are mainly divided into two categories: one that relies on handcrafted (or engineered) feature extraction followed by classification; and one that relies on deep learning algorithms that perform feature extraction and classification simultaneously using one network architecture.

Many metrics have been utilized to validate both types of CAD systems. These include: accuracy, specificity, sensitivity, Area Under Curve (AUC) of the ROC curve, precision, recall and F-score.

A. Classification of Lung Nodules based on engineered features

Most CAD systems from this category have two sequential steps: First, features are extracted from the lung nodules. Then, traditional classifiers are trained using the extracted features to discriminate between the different classes. Methods from this category vary de-

pending on the type of extracted features (i.e., shape, appearance, and/or texture) and on the used classifiers (i.e., Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), RandomForest, etc.). Initially, researchers have mainly focused on shape and appearance features as they hypothesised that the difference in size between both nodules categories is enough to discriminate between them (malignant nodules tend to be larger than benign nodules). Some of these works include those of Farahani et al. [23] who used five generative morphological features (compactness, eccentricity, circularity, roundness, ellipticity) provided by the radiologists. Farahani et al. [23] classified the extracted features using an ensemble based approach that integrates multiple classifiers including Neural Networks, K-nearest Neighbors (KNN) and SVM.

Another approach that uses the two sequential steps was proposed by Shewaye et al. [24]. Here, the authors extracted a collection of geometric and histogram features including nodule area, approximate nodule perimeter, nodule diameter, nodule aspect ratio and gray scale histogram. After features extraction, both linear classifiers (SVM and Logistic Regression) and non-linear discriminant classifiers (Random Forest, KNN and AdaBoost) were used for classification. Adaboost was shown to be the optimal classifier among all considered classifiers.

In [25], Likhitkar et al. proposed a CAD system that enhanced the images, segmented the lungs contour, extracted features and classified them into benign versus malignant. The last two steps that focused on lung nodules diagnosis consisted of using the shape, growth rate, boundary of the nodule and density as input to an SVM classifier to discriminate between benign and malignant classes.

Firmino et al. [26] proposed an approach based on features provided by radiologists, including spiculation, margin, calcification, lobulation, sphericity, texture and internal structure. These features were validated using Gaussian Naive Bayes (NB), Fisher's Linear Discriminant (FDA) and SVM. The authors considered learning five different classes (nodules highly unlikely of being malignant, nodules moderately suspicious of being malig-

nant, nodules highly unlikely of being malignant, nodules with indeterminate malignancy and nodules highly suspicious of being malignant). Among all considered classifiers, SVM gave the best classification results using both cross-validations and leave-one-out strategies.

The method proposed by Jeeva et a1. [27] used as features the perimeter, centroid, mean, irregularity index, area and eccentricity. These features are then input to a feed forward network for classification.

Chen et al. [28] used a boosted ensemble algorithm (XGBoost) on clinical and radiographic features such as nodule size, age, solid nodule, speculation, lobulation, etc.

In more recent years, researchers noticed that both density and texture of lung nodules are not uniform, which may create a variation in Hounsfield Units (HU) values associated with the Computed Tomography (CT) imaging modality. Consequently, they exploited this variation by focusing on texture features. Some of the methods that are based on texture features include those of Han et al. [6] who developed a CAD system using Gabor [29], Haralick [30] and Local Binary Patterns (LBP) [31] texture features. In this approach, the three different texture features were extracted from 2D slices from the lung nodules ROIs. Then, an SVM classifier was used to classify each feature. Using a standard CT image collections, authors reported that the Haralick features achieved the best diagnosis results. The authors also extended their work to 3D volumes and showed that more accurate results can be obtained.

Nishio et al. [32] proposed a CAD system that classified lung nodules into benign or malignant using a variant of the Local Binary Pattern features along with the XGBoost and SVM classifiers.

In [33], Rodrigues et al. proposed an algorithm based on Structural Co-occurrence Matrix (SCM). They created multiple configurations of their algorithm by considering different filters to preprocess images before features extraction. Three different classifiers (KNN, SVM and multi-layer perception) were used to evaluate the considered features and compare their performance to LBP, statistical moments, Gray Level Co-occurence Matrix

(GLCM) and central moments.

Huang et al. [34] investigated the use of fractal texture features, from Fractional Brownian Motion (FBM) model. Here, also the SVM was used for classification.

Wang et al. [35] also investigated lung nodules classification using Haralick texture features, and considered four different classifiers (Extreme Learning Machine (ELM), Probabilistic Neural Network (PNN), SVM and Multilayer Perceptron (MLP)). They reported that the ELM classifier gave the best diagnosis results. To make use of uncertain data (i.e., data with ambiguous labels), the authors developed a semi supervised version of the ELM classifier. They showed that the semi supervised version outperformed all four supervised methods.

De et al. [36] proposed a method based on different features for lung nodules diagnosis. They utilized phylogenetic diversity with particular indices that include extensive quadratic entropy, pure diversity indexes, average taxonomic distinctness, intensive quadratic entropy and total taxonomic distinctness. A genetic algorithm was used for classification.

As outlined above, both appearance and texture features have proved to be useful in the lung nodules classification task. The success of such features individually made researchers turn toward combining them in the same framework in an attempt to further increase the accuracy of the diagnosis. Some of these architectures include those of Huang et al. [37] and Tu et al. [38], where the authors used a large number of features and applied p-value and t-test statistical tests, along with a forward search algorithm, to select a small subset of effective features.

Another CAD system was proposed by Shaffie et al. [8]. This group fused Higher-Order Markov Gibbs Random Field (MGRF) appearance features to retrieve the spatial inhomogeneities of the lung nodules, spherical Harmonics to describe the shape of nodules and volumetric features to describe the nodules size. They fed these features to a deep autoencoder (AE) classifier.

More architectures that combine shape and texture features include those of Dhara et al. [7] who extracted 2D and 3D texture based features as (HoG and GLCM) and 2D and 3D shape features (sphericity, spiculation, lobulation, area, perimeter, etc.) and classified them using SVM.

Gong et al. [4] proposed a CAD system that combined texture features (Neighborhood Gray Tone Difference Matrix (NGTDM), GLCM, Gray Level Size Zone Matrix (GLSZM) and GLRLM (Gray Level Run Length Matrix)) and shape features (volume and surface area) and Histogram features. They analysed these features using naive Bayes (NB), SVM and LDA classifiers with the Relief-F [39] feature selection algorithm.

Ma et al. [40] utilized radiomics features including heterogeneity, information in multi-frequencies, shape and intensity and classified them using the Random Forest algorithm.

B. Classification of Lung Nodules based on deep learning

Deep learning of features along with classification have proven to be very successful in diverse domains and applications. Contrary to traditional approaches that treat features extraction and classification as two independent and sequential steps, in deep learning feature learning and classification are performed simultaneously. In particular, Convolutional Neural Networks (CNN) have proven to be very effective in image recognition tasks.

A typical architecture of CNN usually consists of convolutional layers, pooling layers and fully-connected layers. A typical CNN architecture is shown in Figure 3.

B.1. Convolutional layer

What distinguish CNN from all other artificial neural networks are the convolutional layers. The input data is convolved using many kernels to extract features, then the output is entry to an activation function that produces feature maps. These feature maps are then

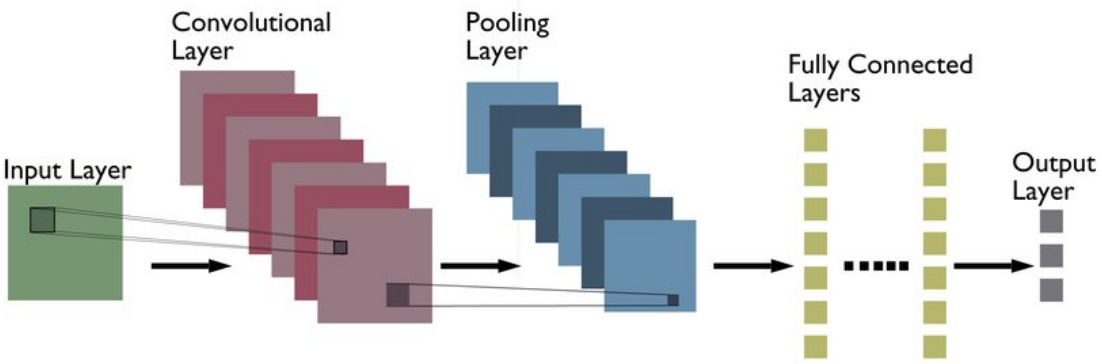


FIGURE 3: Typical CNN architecture.

fed to the next hidden layer, l. The output of each layer, x_i^l, is computed using the output of the preceding layer x_i^{l-1} using:

$$x_i^l = f(\sum_i x_i^{l-1} * k_{ij}^l + b_j^l) \tag{1}$$

In (1), x_i^l represents the i^{th} input feature map to layer l, k_{ij}^l the kernel joining the j^{th} feature map of the output layer to the i^{th} feature map of the input layer and b_j^l is the bias.

Once the process is repeated multiple times across the convolutional layers of the CNN, progressive levels of features are learned. These extracted features can be classified into two types: low-level and high-level features. Low-level features are associated with the early layers of the CNN, and usually describe corners, edges, lines, etc. High-level features are associated with the deeper layers of the CNN and describe the details and more salienet features informations associated with each object of the image.

B.2. Activation function

An activation function is applied to the convolved output of each layer. Typical activation functions used in CNN include ReLu, sigmoid and tanh. These activation functions are illustrated in Figure 4.

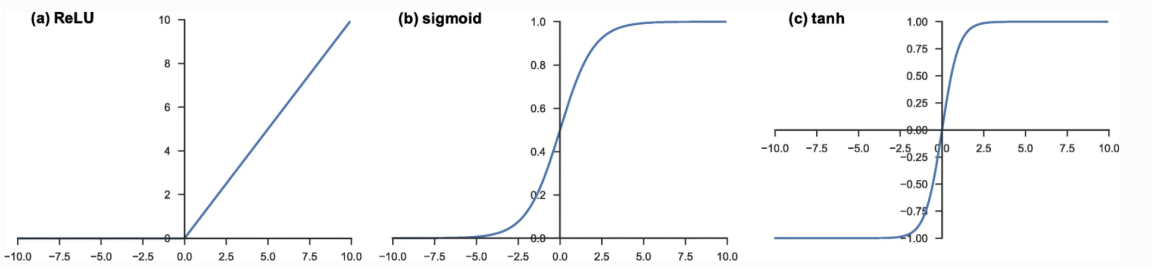


FIGURE 4: Typical activation functions used in CNN.

B.3. Pooling layer

Pooling layers provide a downsampling operation to reduce the in-plane dimensionality of the feature maps. The most common pooling operations are max pooling and average pooling.

B.4. Fully Connected Layers

The output feature maps of the final convolution or pooling layer is flattened (transformed into a one-dimensional array), and connected to one or more fully connected layers, also known as dense layers. In these dense layers, every input is connected to every output by a learnable weight. When the features extracted by the convolution layers are created, they are mapped by a subset of fully connected layers to the final outputs of the network (i.e., the probabilities for each class in classification tasks). The final fully connected layer typically has the same number of output nodes as the number of classes.

Many researchers have focused on CNN architectures based networks and adapted them to the lung nodules diagnosis task. Architectures used for this application include 2D CNN [10], multi-view 2D CNN [41] and 3D CNN [11,42], In addition, several architectures such as residual blocks [43], transferable multi-model ensemble [44], optimal deep neural networks [13] and deep neural networks [45] have been investigated.

Shivan et al. [46] compared multiple pre-trained CNNs including Googlenet [47],

AlexNet [48], ResNet50 [49] and ResNet18 [50] to classify lung nodules. They reported that AlexNet model along with transfer learning achieved the best classification results.

In [51], the authors compared three different deep learning algorithms. These include Stacked Denoising Autoencoder (SDAE) [52], CNN and Deep Belief Networks (DBNs) [53]. They concluded that DBN achieved the best classification results.

In [54, 55], the authors used a CNN architecture with data augmentation generating additional images with similar characteristics as pulmonary nodules using Generative Adversarial Networks (GANs) [56]. They reported that using GAN allowed to enlarge the dataset and to achieve better classification results than other architectures.

Hua et al. [57] analysed CNN and Deep Belief Network [58] and compared them to two traditional generative features followed by classification. They showed that deep learning based methods achieved better classification results.

Liu et al. [12] proposed a multiview convolutional neural networks for lung nodules classification where features are extracted from multiple views of the same sample. They showed that their method is more efficient than the single view method.

Another group [10] proposed a hierarchical learning framework based on Multi-scale Convolutional Neural Networks (MCNN). They extracted discriminative features from succeeding layers from multi scale nodule patches and concatenated the response neuron activations obtained at the last layer from each input scale. They showed that their method is effective in classifying malignant and benign nodules without nodule segmentation.

Zhang et al. [59] proposed a Multi-Level Convolutional Neural Network (ML-CNN) for the lung nodules classification task. To optimize the hyperparameter configuration, the authors used a non-stationary kernel-based Gaussian surrogate model. They stated that their algorithms perform better than several hyperparameter optimization methods and manual tuning.

Despite the effort of some researchers to develop 2D CNN architectures with multi

views or multi level features to optimize the accuracy of their diagnosis, single-or multi-view 2D images are still incapable to exploit the complete 3D information provided by lung nodules [11]. To overcome these disadvantages, some works developed architectures that take as input 3D volumes instead of 2D patches. An example of these frameworks is of Nasrullah et al. [60]. The input of this framework were volumes where nodules are located in the center. To learn higher level features, the authors used a CNN along with 3D MixNet blocks and Gradient Boosting Matching (GBM).

Another work proposed by Zhang et al. [61] was based on a 3D CNN architecture. The main idea behind this work is to build a framework capable of classifying both large and small nodules. They proved that their framework is efficient in classifying both large lung nodules with diameter between 10 and 30 mm and small nodules with diameter (≤ 10 mm).

Polat et al. [62] proposed to use both Straight 3D CNN with conventional softmax and hybrid 3D CNN with Radial Basis Function (RBF)-based SVM for the lung nodules classification task. They compared their models to 3D-GoogleNet and 3D-AlexNet CNN architectures. Both methods achieved better classification results than the traditional CNN architectures and the 3D CNN with Radial Basis Function (RBF)-based SVM achieved the best diagnosis results.

Deep learning models evolved during the past few years and many new architectures improved the accuracy in multiple domains. As these models get more efficient, researchers started adopting them to CAD systems. Some of these works include those of Nishio et al. [32] who extracted features from 2D patches using pooling operations, image convolutions and principal component.

Lakshmanaprabu et al. [13] proposed an Optimal Deep Neural Network (ODNN) for the lung nodules diagnosis task. Another CAD system for the lung nodules diagnosis have been proposed by Xie et al. [44]. This group utilized the transferable multi-model ensemble (TMME) to classify nodules into malignant or benign. In their later work [63],

they proposed a multi-view knowledge-based collaborative (MV-KBC) deep model for the lung nodules diagnosis task. In another work [64], they proposed to fuse shape (Fourier shape descriptor to study the heterogeneity of nodules), texture (GLCM features) and deep model-learned information (deep convolutional neural network) at the decision level to classify lung nodules.

Ren et al. [65] proposed a Manifold regularized Classification Deep Neural Network (MRC-DNN). Another group [66] proposed a cascaded architecture for the lung nodules classification task. In their method, the authors used transfer learning to identify images that contain nodules, then they classified them.

Veasey et al . [67] proposed a deep Convolutional Neural Network (CNN) with recurrent neural network framework that utilises pre-trained 2-D convolutional feature extractors on a standard dataset. They stated that their architecture achieved similar performance to a 3-D CNN that involves half the number of parameters. The same group proposed a convolutional attention based network that uses pre-trained 2-D convolutional feature extractors and evaluated it for single- and multitime-point classification. Their results showed that the proposed method achieved better results than a 3-D network with less than half the parameters on single-time-point classification and achieved better performance on multi-time-point classification.

Other methods that are based on deep learning include [68–72] where authors used DenseNet and adaptive boosting, 3D Dual Path Networks, transfer learning, a malignancy evaluation network and a joint radiology analysis and an end-to-end dense convolutional binary-tree network respectively for the lung nodules diagnosis task.

C. Multiple Instance Learning

In Multiple Instance Learning (MIL), each object is represented by a bag of multiple instances. Labels are provided at the bag level but labels of instances within each bag are not available. Most MIL algorithms assume that a bag is labeled positive (+1) if at least

one of its instances is positive and negative (-1) if all of its instances are negative. Let Y be the label of a bag X, defined as a set of N instances, $X = \{x_1, x_2, ..., x_N\}$. Each instance x_i has an unknown label y_i. The label of the bag is given by:

$$I = \begin{cases} +1 & if \quad \exists y_i : y_i\text{=+1} \\ -1 & if \quad \forall y_i : y_i\text{=-1} \end{cases} \tag{2}$$

The above standard assumption has been used in most early methods [14, 73, 74], and also in some recent ones [75, 76]. It is based on the assumption that it is not necessary for each instance in the bag to be labeled positive for the bag to be labeled positive.

Several MIL algorithms have been proposed, and a review of many of them can be found in [75, 77, 78]. In the following, we outline few MIL algorithms that we have adapted to our application.

C.1. Support Vector Machines for Multiple-Instance Learning (MI-SVM)

MI-SVM [73] extends the SVM algorithm to the MIL paradigm. Initially, we train a standard SVM classifier with an initial ensemble of positive instances (i.e., initialize positive instances by taking a random number of instances from each bag) and negative instances. Using this learned initial model, a confidence value is assigned to each instance in each bag. Then, few instances from each bag are selected as positive (based on confidence values). In the next iteration, the algorithm is optimized to find a maximal margin hyperplane separating the newly selected positive instances from the all other instances. The above steps are repeated until the same positive instances are identified in two consecutive iterations or a predefined maximum number of iterations is reached. The final model is then used to label the test bags.

C.2. Expectation Maximization - Diverse Density (EM-DD)

The EM-DD [79] defines an initial target point h obtained by trying points from positive bags, then repeatedly performs the following two steps that combine Expectation Maximization with Diverse Density to search for the maximum likelihood hypothesis. In the first step (E-step), the current target h is used to pick one instance from each bag, which is most likely (given the used generative model) to be the one responsible for the label given to the bag. In the second step (M-step) a gradient ascend search (using quasi-newton [80]) is used to find a new h_i that minimizes the Diverse Density DD(h). Next we let $h = h_i$ and repeat the two steps. These steps are repeated until the algorithm converges.

C.3. Multi-instance learning based on Graphs (mi-Graph)

In mi-Graph [81], each bag is represented by a graph where instances represent the nodes of the graph. The edges of the graph are based on the distance between instances. If the distance between two instances is smaller than a threshold, an edge will link these instances. Otherwise, no direct edges will connect the nodes. After representing each training bag by a graph, different classifiers could be used. Examples include a K-nearest neighbor classifier that employs graph edit distance [82], or using a graph kernel [83] to capture the similarity among graphs and then solve the classification problems by using kernel machines such as SVM.

C.4. Multi-instance Fisher Vector (miFV)

In miFV [84], the instances of all bags are first clustered into several "groups", and then mapped into a new feature vector representation (i.e., Fisher Vector [85]) with the bag-level label. The mapped vectors are then fed to a standard supervised learner (i.e., an SVM, ANN, etc.), to learn a classification model. Similar to the training bags, the testing bags are first mapped into feature vectors using the same mapping function, and a bag-level

label is predicted using the learned classifier.

C.5. Multi-instance vector of locally aggregated descriptors (miVLAD)

MiVLAD [86] clusters the instances of the training bags into K clusters using the K-means clustering algorithm. Then, each instance is assigned to the closest cluster [86]. At this stage, a mapping function maps a bag into a feature vector based on the clusters assigned to its instances. The new mapped feature vectors are then fed to a standard supervised learner (i.e., an SVM, ANN, ect.) to learn a classification model. Similar, to the training bags. the testing bags are mapped to feature vectors using the same mapping function and a bag-level prediction is computed using the learned classifier.

C.6. mi-Net: Instance-Space MIL Algorithm

Mi-Net [87] is an MIL algorithm based on neural networks [87]. A bag of instances is fed sequentially into a succession of L (four) fully connected (FC) layers with ReLU activation [87]. Each instance feature is denoted by x_{ij}^{L-2} in the $(L-2)^{th}$ layer and its instance probability is denoted by p_{ij}^{L-1} where p_{ij}^{L-1} is a scaler between [0,1]. In the last layer, there is a Multiple Instance Pooling (MIP) layer which takes instance probabilities as input and outputs a bag probability, denoted as: $P^L(X_j)$. In other words, instance scores from four FC layers are learned then aggregated into bag scores to predict the label of the bag via MIP layer. The first $(L-2)$ FC layers can learn some more discriminative instance features compared to the original features x_{ij}. The last connected layer has one neuron with a sigmoid activation that is used to predict the positiveness of the instances. The MIP method satisfies the MIL standard assumption and a bag is positive if it contains at least one instance with large positiveness. Similarly, a bag will be labeled as negative if all of its instances have low positiveness.

MI-Net [87] is a variation of mi-Net that focus on learning a bag representation,

rather than predicting instance probability. In MI-Net, the MIP layer aggregates the discriminative features learned from the first three fully connected layers into one feature vector as a bag representation. The last FC layer with only one neuron and sigmoid activation takes the bag representation as input and predicts the bag probability.

In [87], the authors proposed a variation of MI-Net that integrates deeply-Supervised Nets (DSN) structure. In this variation, deep supervisions are added to each middle FC layer that can learn instance features. During training, the supervision is added to each level and during testing, the mean score for each level is computed.

The authors in [87] proposed another variation of MI-Net that uses deep residual learning. While the original residual learning [88] learns representation residuals using convolution, batch normalization and ReLU, the bag representation in MI-Net learns residuals via fully connected layers, ReLU and MIP.

CHAPTER III
MULTIPLE INSTANCE LEARNING FOR LUNG NODULES CLASSIFICATION

The objective of this thesis is to research the possibility of using MIL for lung nodules diagnosis. First, we design a multiple instance representation of the nodules using a nested sequence of volumes centered at the identified center of the nodule. We extract one feature vector from each volume. The set of features for each nodule are combined and represented by a bag. Next, we investigate and adapt some existing algorithms and develop new ones to this application. We start by applying benchmark MIL algorithms to traditional engineered features. Then, we design and train simple CNNs to learn and extract features that characterize lung nodules. These extracted features are then fed to various benchmark MIL algorithms to learn classification models. Finally, we develop new algorithms that combine feature learning and multiple instance classification in a single network. These algorithms generalize the CNN architecture to multiple instance data.

A. Lung nodules classification based on engineered features and Multiple Instance Learning

Lung nodules classification, like most traditional learning algorithms, requires labeled training data to learn the parameters of the prediction model. For this application, the contour of the nodules needs to be delineated by radiologists so that descriptors, that capture salient features of the nodules, can be extracted and used to train a classifier that predicts the final diagnosis. This labeling process is labor intensive for radiologists since they need to check multiple slices for the same nodule sample and delineate the nodule in every slice to get the final 3D contour. In addition, when the nodule is hard to delineate (for example, when the nodule is small or the nodule is sub-solid), different radiologists

may identify different contours for the same nodule. This will create a certain ambiguity for researchers as they will have this question in mind: To which extend is the segmentation accurate and reliable ? Figure 5 illustrates a sample nodule where three radiologists disagree when delineating the contour of the nodule at different slices. For the first slice, the contour identified by radiologist 1 (shown in red) is at least three times larger than the contour identified by radiologist 2 (shown in blue). Similar labeling can also be observed for slices 2 and 3. For slices 4 and 5, radiologists 2 and 3 did not even detect the nodule.

In addition to the difficulty to robustly segment nodules, some researchers [89, 90] have concluded that the classification results can be improved when some parenchymal structures are included along with the nodules texture in comparison to using the nodules texture only. Consequently, researchers relaxed the requirement to identify the exact contour of the nodules. Instead, only the center of mass is provided and features are extracted from a 2D or 3D region of fixed size centered at the provided nodule center of mass. This simplification allowed researchers to annotate much larger datasets and train more complex models. Though, in this case another challenge arise since these methods use one fixed bounding box for all samples (typically the one that covers the biggest nodule in the training dataset). This volume size may not be the optimal choice for all nodules in the dataset. Figure 6 displays sample malignant nodules where the box size was fixed to include the largest nodule as illustrated in Figure 6 (a). For a medium size nodule (as the one displayed in Figure 6 (b)), a large area of the box will include background. In this case, the extracted features may capture more information about the background than the nodule. Figure 6 (c) displays an even smaller nodule where the extracted features will more likely capture more background than the nodule in itself.

We propose a solution to adapt learning algorithms to objects with varying sizes, without annotating the bounding box of each object. We represent each nodule by multiple instances (extracted using different bounding boxes) and use Multiple Instance Learning (MIL) algorithms. In MIL, each sample is represented by multiple feature vectors. Each

feature is referred to as instance and the collection of instances representing one object is called a bag. In our considered application, a bag represents the nodule sample and the instances represent features extracted from a nested sequence of volumes around the nodule. The advantage of using MIL resides in the ability of one or few of these nested volumes to capture the nodule's salient features and to be a good representation of the nodule independently of its size. A bag in this case is labeled positive (malignant nodule) if at least one of its nested volumes belongs to the target concept associated with malignant nodules and labeled negative (benign nodule) if none of these nested volumes are close to the target concept.

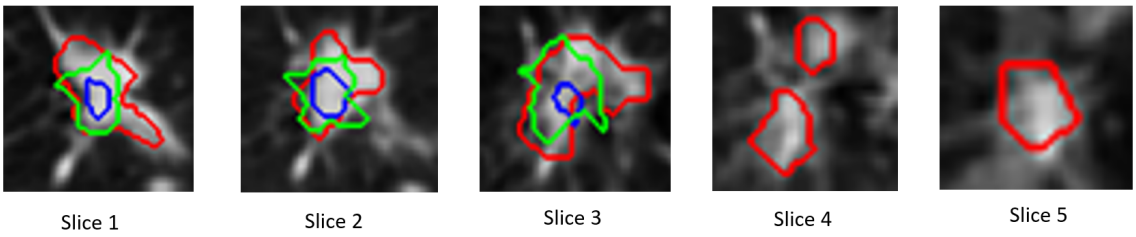
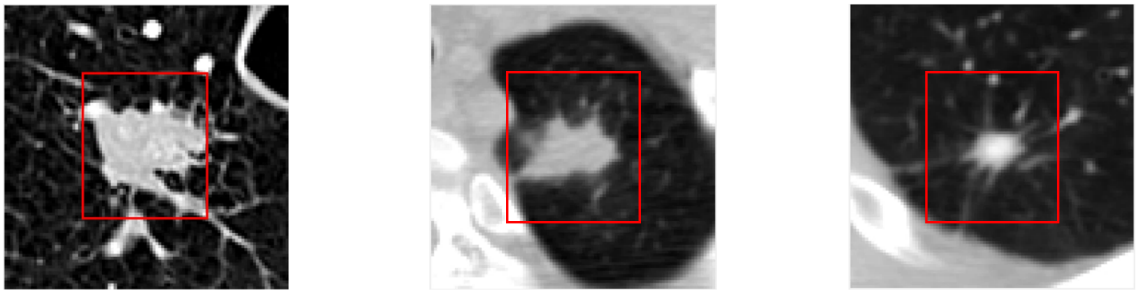


FIGURE 5: X-Y views for a sample nodule delineated by three radiologists. The contour delineated by each radiologist have a distinctive color (red, green, blue). Missing contours in slice 4 and 5 indicate that two of the three radiologists did not even detect the nodule in these slices.

Our proposed approach, described in the rest of this chapter, is general and can be applied to 2D patches as well as 3D volumes. It can incorporate a variety of benchmark features such as GLCM [7], Fisher vector [85], HoG [91], LBP [6] and others [8, 9]. In the following, we outline the main steps of our approach.

A.1. Patch extraction and feature representation

One of our contributions in this work is to demonstrate the advantages of multiple instance data representation compared to traditional single instance representation that extracts a single feature from one fixed region for the lung nodules' diagnosis task. Thus, in



(a) (b) (c)

FIGURE 6: Three nodules of different sizes captured by a box of fixed size: (a) the box size was fixed to include this large nodule, (b) for a medium size nodule, a large area of the box includes background and (c) the small nodule occupies only a fraction of the box area.

our analysis we include a comparison of our MIL approach to traditional single instance approaches using the same feature descriptors.

Traditional single feature extraction requires the annotation of each training sample by a bounding volume. Only voxels within this region are used to extract a single feature. Figure 7(a) illustrates this SIL paradigm. This approach is efficient and can be very effective when the bounding box captures the nodule with little background. However, as outlined earlier, accurate bounding boxes require extensive analysis by radiologists.

For our proposed multiple instance setting, we assume that only the center of mass of each training sample is known and that information about the spatial extent of the nodule in all three directions is not available. We only assume that the region of interest is larger than $V_{\min}$ and smaller than $V_{\max}$, where $V_{\min}$ and $V_{\max}$ are estimated based on prior knowledge (i.e., the minimum and maximum nodule sizes in benchmark labeled datasets). To take into account the ambiguity and size variation of each sample, first we fix a set of N_v nested volumes such that $V_{\min} = V_1 < V_2 < .... < V_{N_v} = V_{\max}$. Then, from each volume V_i, we extract a feature vector f_i using any of the standard feature extraction methods such as GLCM [7], Fisher vector [85], etc. Each f_i will be treated as an instance, and the N_v instances are combined into a set, or a bag, to represent the nodule sample. Our rationale

is that out of the N_v extracted features, one or few of them will capture the nodules salient features effectively [92, 93]. Figure 7(b) illustrates the proposed MIL paradigm.

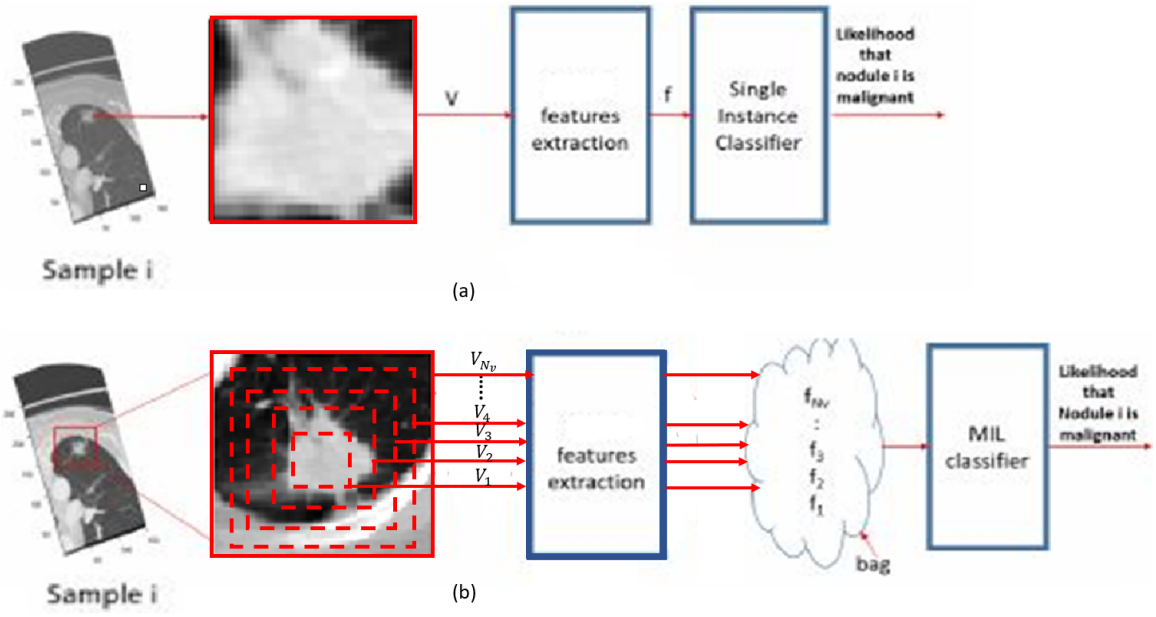


FIGURE 7: Frameworks for feature extraction and classification of nodules using: (a) single instance learning where features are extracted from a single region of interest, and (b) MIL paradigm where features are extracted from multiple instances (i.e., multiple dashed rectangular regions of interest).

A.2. Bag classification

In our work, we test and evaluate a few MIL benchmark algorithms using our extracted multiple instance data. These algorithms are divided into three main categories: (1) instance space MIL algorithms (i.e., EM-DD [79], MI-SVM [73]) that assign confidence values to instances and aggregates these confidences into a bag probability; (2) bag space MIL algorithms (i.e., mi-Graph [81]) that focus on computing the bag probability directly; and (3) embedded space MIL algorithms (i.e., miFV [84], miVLAD [86]) that map each bag to a single feature vector (a feature vector that summarizes the relevant information from all instances) and use a traditional classifier to predict the bag label.

Instance space MIL algorithms have the advantage of being able to identify the best volume (instance) representative of the nodule sample and provide the user with more explainable results. Bag and embedded space MIL algorithms have proved to be more efficient and effective than instance space MIL algorithms for diverse classification tasks [87]. However, they do not explicitly identify the positive instances within each bag, and thus, they have limited explainability.

B. Lung nodules classification based on CNN features and Multiple Instance Learning

Recently, machine learning based on deep Neural Networks have proved to outperform traditional learning methods for many tasks [94–96]. In particular, Convolutional Neural Networks (CNN) are becoming the method of choice to learn classifiers that involve image or video analysis. The main advantage of CNNs is that they do not extract a predetermined set of features. Instead, they process raw images and learn a network that performs both feature extraction and classification simultaneously. In fact, CNNs have proved to be one of the most accurate approaches for lung nodules diagnosis [46, 57, 62]. This has motivated us to develop our second approach for this task where we start by designing and training a CNN. Then, using the learned CNN features, we build a classifier using the same benchmark MIL algorithms that have been tested with engineered features in the previous section.

We designed a simple CNN architecture to classify lung nodules into two classes: benign and malignant. The architecture of this CNN is illustrated on Figure 8. This network has two main components. The first one is for feature extraction and maps an input image I_i to a feature vector X_i. The second component is for classification and predicts a class label for X_i. Both components of the network in Figure 8 are trained simultaneously using benchmark single instance training data to label nodule image patches as either malignant or benign (here all training image patches have fixed sizes: M_x by M_y (i.e., $31 * 31$)).

The first component consists of three convolutional layers, two max pooling layers and Rectified Linear Unit activation (ReLu) layers. Mathematically, a CNN can be represented as a function which is the composition of a sequence of functions. Each function represents a layer, which takes the output of the previous layer, to compute the final feature vector output using connection weights for each layer. A convolutional layer maps an input image patch with a set of multi-dimensional filters to obtain an intermediate output for the next layer. The first convolutional layer of a CNN typically extracts low-level features such as nodule edges and corners. Subsequent convolutional layers act as higher-level features extractors that combine the low level features from previous layers. The max pooling layer computes the maximum among each patch of the feature map from the previous layer. For the activation layer, the Rectified Linear Units ReLU(x) =max(x,0) is employed to address the issue of saturation [97].

The second component of the CNN in Figure 8 is a set of two fully connected layers and they are designed to minimize a classification loss function with respect to the network parameters (i.e., weights of the filters) learned from the training data. To minimize the loss function, we use the stochastic gradient descent (SGD) method. Even though the standard gradient descent method is mathematically simple, the computational cost is enormous especially when we consider a large training set. Therefore, in each step we employ SGD, which calculates a random subset of training examples (i.e., a mini-batch) to estimate the mean gradient for all the training examples. The technique of batch normalization [98] is also used to accelerate the convergence.

After training the network in Figure 8, we are interested only in its feature extraction component. That is, we feed an image patch I_i of size M_x by M_y and map it to a feature vector X_i. We explore this feature extraction component, optimized for nodule image discrimination, to extract features for our multiple instance classification. Similar to the architecture we proposed in Figure 7(b), we fix a set of N_v nested volumes such that $V_{\min} = V_1 < V_2 < .... < V_{N_v} = V_{\max}$. Then, for each volume V_i, we learn a CNN feature vector

f_i. Each f_i will be treated as an instance, and the N_v instances are combined into a set, or a bag, to represent the nodule sample.

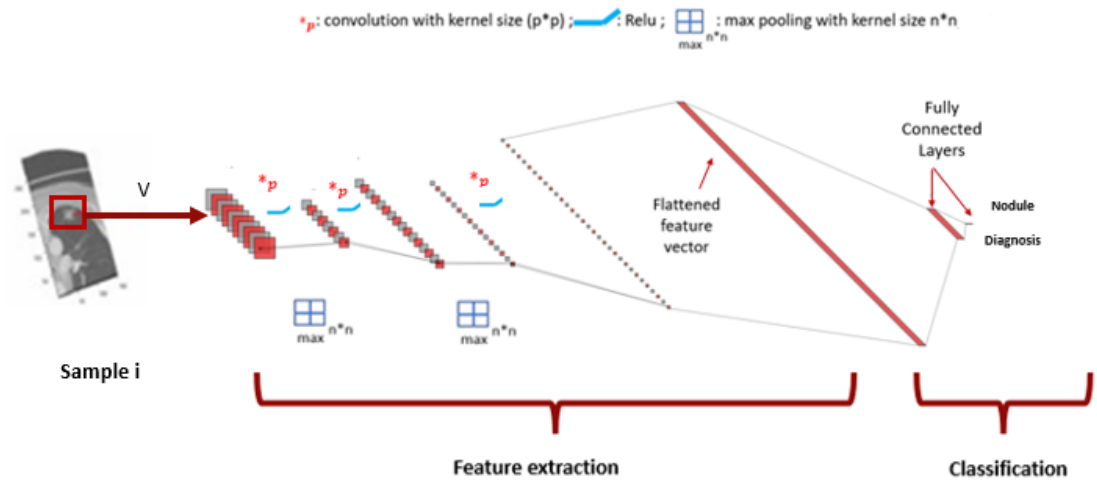


FIGURE 8: Feature extraction and classification using CNN.

C. Lung nodules classification based on end-to-end Multiple Instance Convolutional Neural Networks (MIL-CNN)

Deep learning convolutional networks derive more discriminative features during training. These networks have improved the classification results of multiple machine learning tasks over standard classification algorithms. Therefore, we hypothesize that the integration of deep neural networks with MIL will improve the results over classical MIL algorithms. While our previous approach, described in the previous section, extracts CNN features and trains MIL classifiers in two independent steps and still requires the bounding box of each sample (to train the CNN), the new algorithms proposed in this section combine CNN feature learning and MIL classification into a single network. In particular, we propose four different algorithms that integrate CNN feature extraction and multiple instance Neural Networks in an end-to-end manner. These MIL-CNN have one common architecture, where the same CNN layers are used to extract features from each volume (instance). These feature vectors are input to multiple Instance Neural Networks that compute

the final prediction associated with each bag.

As illustrated in Figure 8, a typical CNN architecture is composed of two main blocks. The first block works as a feature extractor and the second block works as a classifier. Our proposed MIL-CNN framework is illustrated in Figure 9. It processes N_v volumes and extracts their CNN features sequentially before proceeding to the classification block. The N_v extracted feature vectors (instances) are then combined in a bag and passed to the multiple instance classification block.

Let x_{ij} denote the CNN feature vector associated with the ith nested volume $i = \{1, ..., N_v\}$ of nodule sample j. Let y_j denote the class label of nodule j ($y_j = 1$ for malignant and $y_j = 0$ for benign). In case of Single Instance Learning (SIL) strategies, we extract features from a single instance volume and we use tuples (x_j, y_j) for training. On the other hand, for MIL, each nodule j is represented by a bag of features $\{X_j = x_{1j}, x_{2j}, ..., x_{N_v j}\}$, where N_v is the total number of nested volumes used to represent the nodule. In this case we use tuples (X_j, y_j) for training. We should note here that labels y_j are assigned to the bag label. That is, labels for individual instances x_{ij} are not known.

We propose four variations of MIL architectures on the top of the CNN by introducing a pooling layer, called Multiple Instance Pooling (MIP), and using it to transfer the CNN learned feature vectors (learned from the first CNN block) associated with each bag of nested volumes into a likelihood that the input nodule is malignant.

The response of the last convolutional layer in the CNN block is a 2D matrix of CNN features. Let $\{X_{ij}, i = 1, ..., N_v, j = 1, ..., N_s\}$ represent these features where N_s is the total number of samples (nodules) used for training and N_v is the number of nested volumes extracted from each sample. Let L represent the total number of layers and $H^l(.), l = 1, .., L-1$ represent a non linear transformation computed after each layer, that is, $x_{ij}^l = H^l(x_{ij}^{l-1})$, where x_{ij}^{l-1} and x_{ij}^l refer to the output values of instance x_{ij} at layer $(l-1)$ and (l) respectively. $H^l(.)$ is associated with composite operations of fully connections and rectified linear units (ReLU).

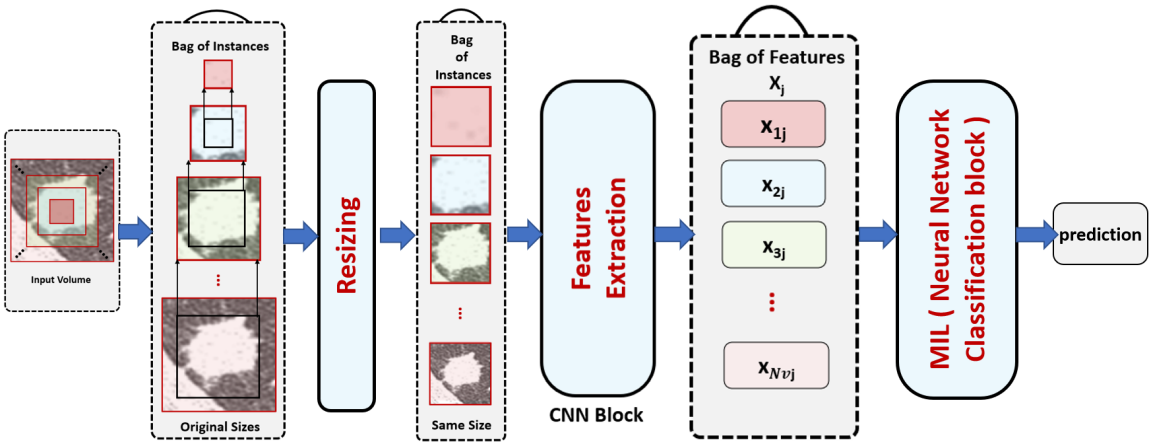


FIGURE 9: General architecture of the proposed MIL-CNNs.

C.1. **Instance-Space Multiple Instance Learning CNN algorithm (Mi-CNN)**

One of our research tasks consists of investigating if MIL algorithms are capable of identifying the positive instances that capture the ground truth region of each nodule without requiring detailed annotation by radiologists. This task requires a class of MIL algorithms (i.e., instance space) that can compute the probability of each instance within the bag. In this section, we propose our instance space MIL-CNN algorithm.

The fully connected classification framework in the instance space, denoted Mi-CNN, computes one final score (or probability) for each instance i in bag j. Thus, the last fully connected layer, denoted $L-1$, of this network has one node that outputs the confidence of x_{ij} in the malignant class. Let p_{ij}^{L-1} refer to the confidence value of this instance. The last layer of the Mi-CNN (denoted L) is a max pooling layer that combines the instance probabilities p_{ij}^{l-1} to compute the bag probability P_j^L using:

$$P_j^L = M^L(p_{ij}^{L-1}) = max_{i=1..N_v}(p_{ij}^{L-1}) \tag{3}$$

In other words, the last layer, M^L, of the Mi-CNN network is responsible for aggregating the learned instance probabilities into a final bag probability, P_j^L.

The Neural Network classification block associated with the Mi-CNN is illustrated in Figure 10. Using this architecture, during training, all samples $x_{ij} \in X_j$ are processed

through all FC layers (i.e., $FC_1, FC_2, .., FC_{L-1}$). The learning of the network weights is achieved by optimizing the cross entropy loss function:

$$\mathcal{L}oss(P_j, y_j) = -\{(1 - y_j)log(1 - P_j) + y_j log(P_j)\} \tag{4}$$

The loss of all N_s bags are accumulated and minimized to train the proposed Mi-CNN by back propagation using Stochastic Gradient Descent (SGD). Thus, the feature of each instance x_{ij} is processed by the network to produce a score p_{ij}. Then, in the last layer, all instances scores are aggregated to output the bag score.

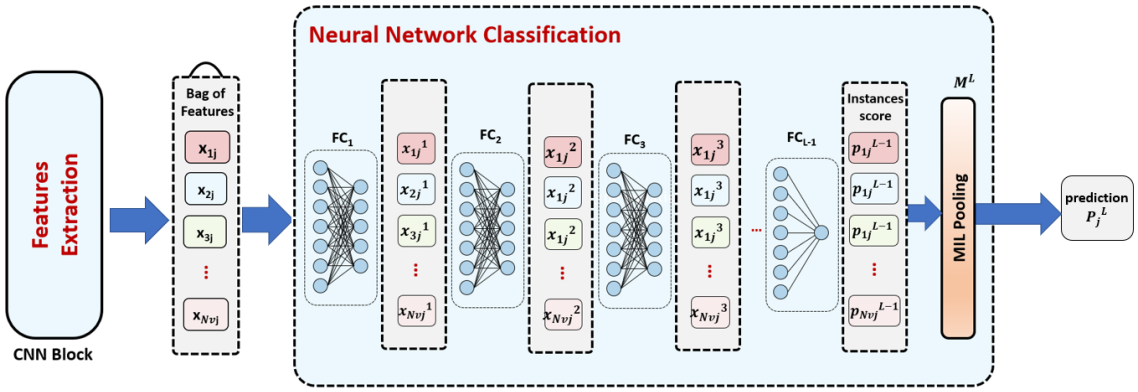


FIGURE 10: The MIL Neural Network Classification block of the proposed Mi-CNN.

C.2. **Embedded-Space Multiple Instance Learning CNN algorithm (MI-CNN)**

The Mi-CNN algorithm first learns instance probabilities. Then, in a final stage these instances are combined to compute the bag score. This instance based algorithm have the advantage of generating more interpretable results. On the other hand, embedded MIL algorithms learn only the final bag score. For applications where individual instance probabilities are not needed, it has been proven that embedded space MIL algorithms perform better than instance space MIL algorithms [22]. This is because embedded methods learn the best feature representation of the whole bag before computing the final bag probability. In the following, we propose an embedded version of our Mi-CNN.

The fully connected classification framework in the embedded space, denoted MI-CNN computes a final score (or probability) of the entire bag j. Layer $L-1$ of the MI-CNN is a max pooling layer that outputs the bag feature vector representative of the entire bag X_j^{L-1} using:

$$X_j^{L-1} = M^{L-1}(x_{ij|i=1..N_v}^{L-2}) = max_{i=1..N_v}(x_{ij}^{L-2}) \tag{5}$$

The last fully connected layer FC_L of the MI-CNN network has one node that is responsible for computing a final bag probability, P_j^L from the learned feature vector X_j^{L-1}.

The Neural Networks classification block associated with the proposed embedded-space MI-CNN variation is illustrated in Figure 11. This network uses the same loss function as the Mi-CNN.

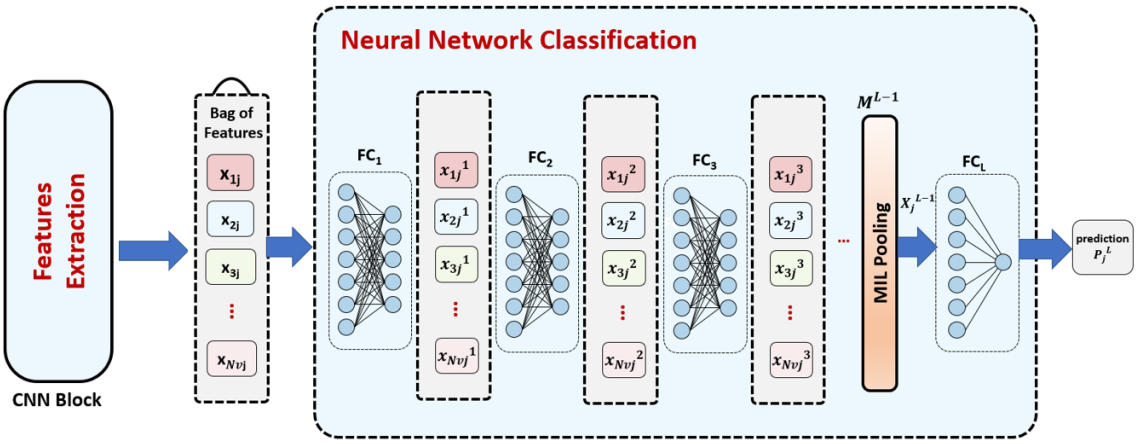


FIGURE 11: The MIL Neural Network Classification block of the proposed MI-CNN.

The only difference between Mi-CNN and MI-CNN is that the latter learns the probability of the bag directly from the learned features without computing the instances score. First, the network computes features x_{ij} for all instances in bag j. Then, the instance features are aggregated in layer $L-1$ to compute one global feature vector X_j in bag j. Finally, in layer L, a score is assigned to each X_j.

We should emphasize here that the Multiple Instance Pooling layer (MIP) in Mi-CNN computes the probability of the bag from the instance probabilities learned from the

previous $(L-1)$ fully connected layers. In contrast, the MIP layer in MI-CNN computes the embedded instance representative of each bag from all bag instances learned from previous $(L-2)$ fully connected layers. This final feature vector is input to a fully connected layer L to output the final probability of the bag.

C.3. Embedded-Space Multiple Instance Learning CNN algorithm (MI-CNN) with Deep Supervisions (DS)

Deeply-supervised nets (DSN) [99], enforce direct and early supervision for both the hidden layers and the output layer. They are a companion objective to the individual hidden layers, which is used as an additional constraint (or a new regularization) to the learning process. Diverse studies [87, 99] proved that discriminative classifier trained on highly discriminative features will result in a better performance than a discriminative classifier trained on less discriminative features. If we take the example of the hidden layer feature maps of a deep network, this observation would suggest that the performance of a discriminative classifier trained using these hidden layer feature maps can serve as a proxy for the discriminativeness and quality of those hidden layer feature maps, and further to the quality of the upper layer feature maps.

In this section, we propose a version of our MI-CNN that incorporates Deep Supervisions. We show that this variation can utilize multiple hierarchies to improve the bag classification accuracy. During training, instance features at early layers can receive better supervision. During testing, we average multiple bag probabilities to get a more robust bag label. The architecture of the fully connected classification framework in the embedded space with Deep Supervisions (DS), denoted MI-CNN-DS is illustrated in Figure 12. MI-CNN-DS computes a final score (or probability) of the entire bag j. A MIP (M^l) is applied to each intermediate FC_l layer in the MI-CNN-DS. M^l is a max pooling layer that aggregates the bag's instance features and outputs a single feature vector to represent bag

j at each level k ($X_j^{l,k}$) using:

$$X_j^{l,k} = M^l(x_{ij|i=1..N_v}^k) = max_{i=1..N_v}(x_{ij}^k), k = \{1, 2, 3, ..L-1\} \tag{6}$$

Each fully connected layer, FC_L, is succeeded by an MIP (M^l) and computes the bag level probability, $P_j^{L,k}$ from the learned feature vector $X_j^{l,k}$ at each level k. The final bag probability P_j^L at the last layer L is computed using:

$$P_j^L = mean_k(P_j^{L,k}), k = \{1, 2, 3, .., L-1\} \tag{7}$$

During training, the loss function of MI-CNN-DS is the sum of all intermediate entropy losses that have been calculated during back-propagation with the stochastic Gradient Descent (SGD). During testing, the average of the bag scores at the $(L-1)$ level is reported. Similarly to MI-CNN, MI-CNN-DS computes the bag scores directly and bypasses the instances sores. The difference between the two variations is that MI-CNN-DS computes the bag score from intermediate fully connected layers and average the final bag score from all computed intermediate scores instead of computing only one final bag probability.

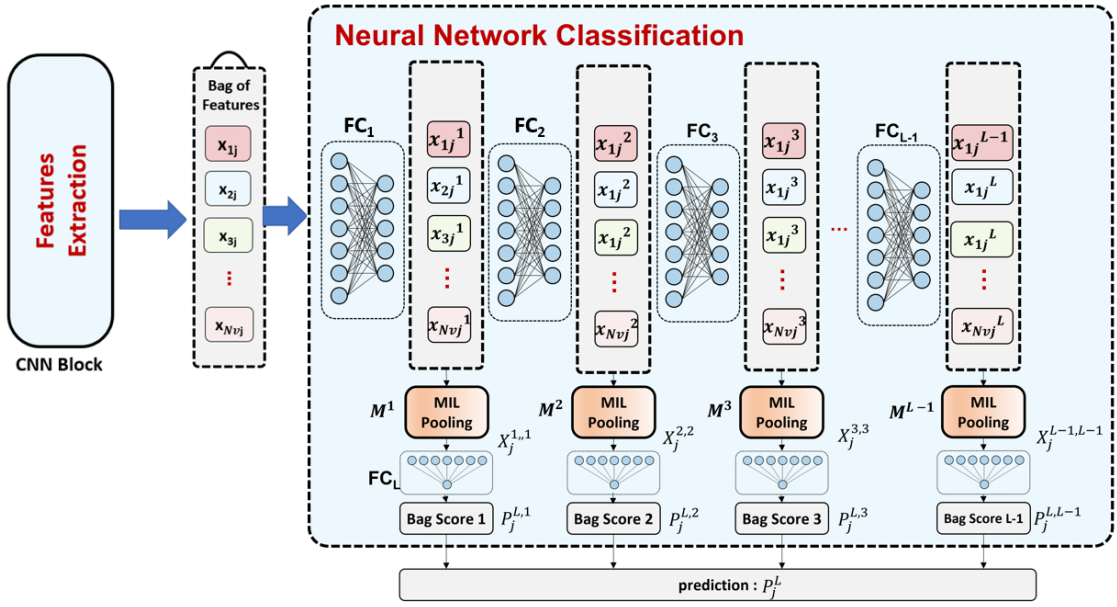


FIGURE 12: The MIL Neural Network Classification block of the proposed MI-CNN-DS.

C.4. Embedded-Space Multiple Instance Learning CNN algorithm (MI-CNN) with Residual Connections (RC)

Deep residual learning [88] showed a significant improvement in image recognition. Thus, for our fourth variation of MIL-CNN, we consider including residual learning. The proposed algorithm, called MI-CNN with Residual Connections (MI-CNN-RC) is illustrated in Figure 13. The MI-CNN-RC computes one final score (or probability) for each bag j. The MIP (M^l), applied after each intermediate FC_l layer in the MI-CNN-RC, is a max pooling layer that outputs X_j^l. These intermediate bag features are combined into a final feature vector that globally represents bag j using:

$$\begin{cases} X_j^1 = M^l(x_{ij|i=1..N_v}^1) \\ X_j^l = M^l(x_{ij|i=1..N_v}^l) + X^{l-1}, l > 1 \end{cases} \tag{8}$$

where $M^l(x_{ij|i=1..N_v}^l) = max_{i=1..N_v}(x_{ij}^l)$. The last fully connected layer FC_L of the MI-CNN-RC network has one node and is responsible for computing a bag level probability, P_j^L, from the learned feature vector X_j^L.

As shown in Figure 13 and Figure 11, MI-CNN-RC follows a similar architecture to MI-CNN. The difference between them is that MI-CNN-RC learns bag representation residuals from intermediate fully connected layers, ReLu and the MIP M^l. These are added together to compose a final feature vector responsible for the bag probability P_j^L.

An alternative approach to our work would be to train N_v CNNs in parallel. Each one of these CNNs will be trained on one specific volume. This method though, is not feasible because we do not have the instances labels and typically, no more than 20% of the instances within every bag are expected to be positive. Thus, about 80% of the instances cannot be labeled using their bag label. In this case, a traditional single instance CNN cannot make reliable predictions as it will be trained with a larger number of incorrectly labeled data.

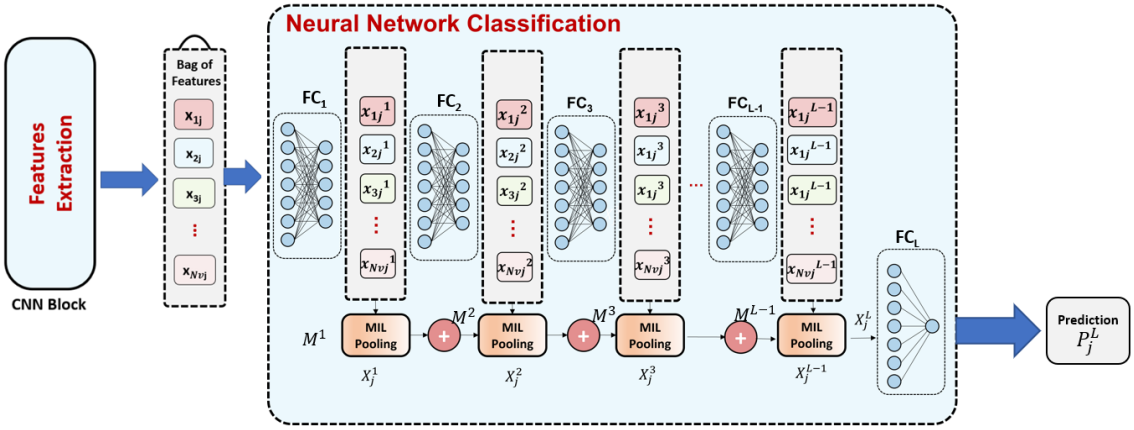


FIGURE 13: The MIL Neural Network Classification block of the proposed MI-CNN-RC.

CHAPTER IV
EXPERIMENTAL APPROACH AND EVALUATION STRATEGY

All proposed frameworks have been tested on three sets of experiments using two datasets: The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [1] and the National Lung Screening Trial (NLST) [2]. We design and report the results of three experiments. Two of these experiments perform five-fold cross-validations on the same dataset (NLST or LIDC). We will refer to these as EXP_{NLST} and EXP_{LIDC}. The third experiment consists of training the algorithms on one collection (NLST dataset) and testing it on the other collection (LIDC dataset). We will refer to this experiment as $Exp_{NLST-LIDC}$.

A. Datasets

A.1. LIDC Benchmark data

The LIDC dataset [1] is a benchmark data for the medical imaging research community. It has been used in most related work referenced in chapter II. This data contains lung cancer screening thoracic Computed Tomography (CT) scans with marked up annotation of each nodule lesion. It is utilized to assess the development of efficient CAD systems for the detection and diagnosis of lung cancer in its early stages. Eight medical imaging companies and seven academic centers collaborated to create this dataset composed of 1018 CT cases that have thoracic CT scan images as DICOM files and an additional informative XML file for each patient. In our experiments, we excluded scans with a slice thickness greater than 2.5 mm, missing slices, and conflicting slice spacing, reducing the data to 888

CT scans. These scans contain nodules with diameter ≥ 3 that have been annotated by four radiologists where the malignancy score is provided with values varying between 1 (the highest probability for the nodule to be benign) and 5 (the highest probability for the nodule to be malignant). Ideally for the data to be more accurate, we should pick only samples where four radiologists agree, however, this condition will reduce the data significantly. As a compromise and to keep a larger subset of the data, we test our algorithms on samples that have been recognized as nodules by at least three radiologists. Moreover, we only use benign nodules that have a median score ≤ 3 and only malignant samples that have a median score > 3. As a result, our cleaned version of LIDC has 260 malignant nodule and 387 benign nodule.

A.2. NLST

In a National Lung Screening Trial (NLST) [2], over 53000 high-risk participants were screened annually for three years. Samples from the screening are denoted by time 0 (T0, for year 1), time 1 (T1, for year 2) and (T2, for year 3). The National Cancer Institute provides data from 15000 of these subjects where most of the CT scans have no nodules. In our experiments, we only use samples from participants that were biopsied during the screening process (biopsy results were used for ground truth annotation). This data is provided with an approximate location of the biopsied nodule. To date, a subset of this data, named by NLSTx [67] has been analyzed by an in-house radiologist to determine the exact location of these nodules and identify a bounding box around each nodule. This annotated data consists of a total of 488 malignant and 1923 benign nodules divided as follows: 803 nodules for T0 (154 malignant and 649 benign), 847 nodules for T1 (205 malignant and 642 benign), and 761 nodules for T2 (129 malignant and 632 benign). As noted, the number of samples for each time is different because some patients got recorded for T0 and T1 only while others were missing from the database.

A.3. Data preparation

All CT scans have been re-sampled with resolution (0.625, 0.625, 2) to ensure the homogeneity of resolution for all CT scans and for both used datasets. After extraction, all considered volumes (instances) have been normalized by clipping the Hounsfield Units (HUs) of all pixels within the volume to the range of (-1000,400) using:

$$I = \begin{cases} 0 & if HU \leq -1000 \\ \frac{HU+1000}{1400} & if -1000 < HU < 400 \\ 1 & if HU \geq 400 \end{cases} \tag{9}$$

Next all volumes are scaled to the (0,1) range using min-max normalization.

B. Evaluation strategy

The objective of this thesis is to develop a robust CAD system for the lung cancer diagnosis task. To achieve this goal, we evaluate, analyze and compare multiple benchmark MIL algorithms using different type of features. We also develop new deep learning algorithms for multiple instance data. We guide our research by formulating and answering five important research questions (RQs). These RQ's are described below.

***RQ1*: What is the best MIL algorithm for this application ?**

The main objective of this thesis is to investigate MIL for the lung nodules diagnosis task. MIL is a good approach for this application since the size of lung nodules can vary significantly. To answer this question, we perform three different experiments by varying the datasets used for training and validation. We experiment with different benchmark MIL algorithms and different types of features. We validate the different methods by comparing their Area Under Curve (AUC) values.

***RQ2*: What are the main nodule parameters that influence the classification?**

The main objective of this question is to study the nodule parameters that can in-

fluence the classification results. In particular, we investigate parameters related to the size and contrast of the nodule as these are important ones used by radiologists in their diagnosis. To answer this research question, we analyze the distribution of parameters related to the size and contrast of nodules. In particular, we compare the distributions for nodules that were classified correctly/incorrectly by the most effective algorithms identified in RQ1.

***RQ3*: What is the best feature representation ?**

The main objective of this question is to determine which feature representation is more suitable for this task. We consider generative features (GLCM) and features learned using a CNN. To answer this question, we analyse and compare the results of MIL algorithms that use both of these features.

***RQ4*: Can an MIL algorithm learn to identify the positive instances (among a bag with large number of candidates) that are close to the ground truth ?**

The main idea behind using MIL for the lung nodules diagnosis task is to overcome the limitation of lung nodule segmentation or applying algorithms using one fixed volume size for all nodules. MIL represents each nodule by a large number of instances extracted from different volumes. The goal of this RQ is to verify if the algorithm can identify a few positive instances that represent the nodule without relying on accurate ground truth. To answer this question, we analyze the results of the most reliable MIL algorithm (identified in previous RQ's) by correlating the confidence values assigned to each instance to their spatial proximity to the actual ground truth.

***RQ5*: What is the best learning approach for this application: Single Instance Learning (SIL) applied to the ground truth region or Multiple Instance Learning (MIL) applied to multiple regions around the ground truth ?**

The main objective of this question is to compare the performance of MIL algorithms and SIL algorithms. SIL algorithms tend to be simpler, but impose more requirements on the ground truth. MIL algorithms, on the other hand, tend to be more complex but provide more flexibility in representing nodules. To answer this question, we analyse the

results of SIL and MIL algorithms and correlate the predictions (and identified instances by MIL) to the ground truth extracted by the radiologists and the consistency among the different radiologists.

CHAPTER V
EXPERIMENTAL RESULTS AND DISCUSSION

In this chapter, we evaluate, analyze, and compare our proposed models introduced in Chapters II and III. We start by describing our experimental settings. Then, we address each research question by describing the conducted experiments and analyzing the results.

A. MIL representation and features extraction

A.1. Multiple instance representation

Our proposed Multiple Instance Learning frameworks for lung nodules classification are applied on 3D volumes. We assume that only the center of mass of each training sample is known and that information about the spatial extent of the nodule in all three directions is not available. In our representation, we only assume that the region of interest is larger than $V_{\min}$ and smaller than $V_{\max}$, where $V_{\min}$ and $V_{\max}$ are estimated based on prior knowledge (The minimum and maximum size of nodules in each dimension are known). To take into account the ambiguity and size variation of each sample, first we fix a set of N_v nested volumes such that $V_{\min} = V_1 < V_2 < .... < V_{N_v} = V_{\max}$. Each volume V_i, is processed for feature extraction as will be described in the following subsection. In our experiments, we set the number of volumes (instances) N_v= 60. These volumes expend from the smallest one, V_{min} with dimensions = $[7, 7, 7]$ to the largest one, V_{max}, $[65, 65, 19]$ in the x, y and z directions. Our choice is based on selecting V_{min} as the minimum bounding box of the smallest benign nodule and V_{max} as the minimum bounding box that covers 90% of all malignant nodules size. In our training data, the remaining 58 sizes are selected

randomly between V_{min} and V_{max} such as x $\in$ {7,11,19,27,41 or 65}, y $\in$ {7,11,19,27,41 or 65} and z $\in${7,11,19,23 or 27}. These dimensions have been selected to allow us to define reasonable intervals between all dimensions in V_{min} and V_{max}. As a result, each nodule will be represented by a bag of 60 instances. Figure 14 illustrates the proposed multiple instance of each nodule representation.

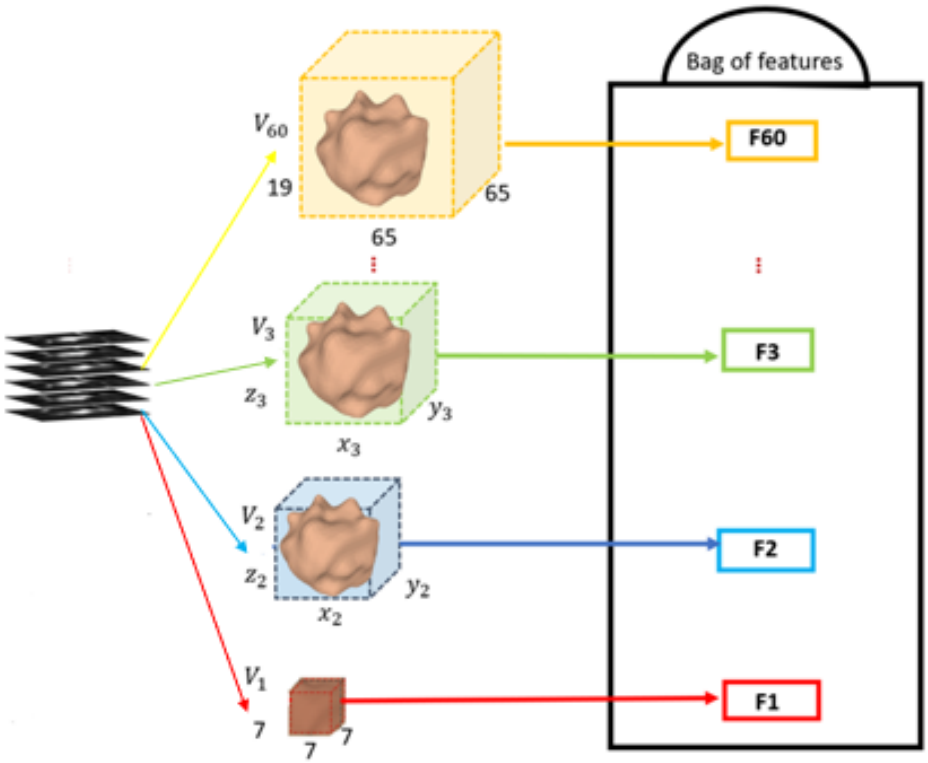


FIGURE 14: Multiple Instance of a given nodule representation.

A.2. Extraction and representation of generative features

As mentioned earlier, our proposed MIL approach can incorporate many benchmark generative features such as GLCM [7], Fisher vector [85], Gray-Level Size-Zone Matrix (GLSZM) [4], etc. In this thesis, we instantiate our approach and validate it using GLCM features [7]. We select these features since they were shown to outperform most other features for lung nodules diagnosis [7]. The GLCM functions characterize the texture of

an image by calculating how often pairs of pixels, with specific values and in a specified spatial relationship, occur in an image. Then, statistical measures are extracted from the co-occurence matrices. Each co-occurence matrix is an N_g x N_g square matrix in size.

In our work, we use the 3D GLCM features introduced by Han et al. [6]. The extraction of these features starts by quantizing the gray values of each input instance (volume (V)) into N_g bins using:

$$V_q(m,n,p) = \lfloor \frac{V(m,n,p)}{max(V)} * N_g \rfloor + 1 \tag{10}$$

where m,n and p define the pixel coordinates of volume V. Next, we construct Co-occurrence matrices (GLCMs) following these definitions:

- Each element i,j of each GLCM define the number of times two pixels of intensities i and j from V_q separated by distance D occur in a specified spatial relationship (direction noted by d). The number of computed GLCMs is directed by the number of considered directions and considered distances. We note $P(i,j) = GLCM(i,j)$.

- Each GLCM is normalized by dividing each element by the sum of all elements using the following formula:

$$P(i,j) = \frac{P(i,j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i,j)} \tag{11}$$

The following quantities are also defined: $\mu_i = \sum_{i=1}^{N_g} i \sum_{j=1}^{N_g} p(i,j)$, $\mu_j = \sum_{j=1}^{N_g} j \sum_{i=1}^{N_g} p(i,j)$, $\sigma_i = \sum_{i=1}^{N_g} (i-\mu_i)^2 \sum_{j=1}^{N_g} p(i,j)$ and $\sigma_j = \sum_{j=1}^{N_g} (j-\mu_j)^2 \sum_{i=1}^{N_g} p(i,j)$.

The elements of the GLCM are considered at this stage as the probabilities of finding the relationship i, j between each pair of pixels separated by distance D in a specific direction d. Finally, a total of 12 Haralick features [7] are extracted from each GLCM in each direction d and distance D, and their mean is taken as the final GLCM feature vector. These features are: energy, entropy, contrast, correlation, homogeneity, variance, sum mean, cluster shade, inverse variance, cluster prominence, max probability and dissimilarity and are defined as:

$$\text{Dissimilarity} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)|i-j| \tag{12}$$

$$\text{entropy} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)log(i-j) \tag{13}$$

$$\text{cluster prominence} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)(i+j-\mu_i-\mu_j)^4 \tag{14}$$

$$\text{cluster shade} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)(i+j-\mu_i-\mu_j)^3 \tag{15}$$

$$\text{sum mean} = \frac{1}{2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)(i+j) \tag{16}$$

$$\text{contrast} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)(i-j)^2 \tag{17}$$

$$\text{energy} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} |P(i,j)|^2 \tag{18}$$

$$\text{homogeneity} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{P(i,j)}{1+|i-j|} \tag{19}$$

$$\text{correlation} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{P(i,j)(i-\mu_i)(j-\mu_j)}{\sigma_i\sigma_j} \tag{20}$$

$$\text{variance} = \frac{1}{2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}[(i-\mu_i)^2 p(i,j) + (j-\mu_j)^2 p(i,j)] \tag{21}$$

$$\text{max probability} = \max\ (\text{GLCM}) \tag{22}$$

$$\text{inverse variance} = \frac{1}{2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{P(i,j)}{(i-j)^2} \tag{23}$$

Figure 15 resume the previously described steps for a sample of 4 x 4 image (of volume V) quantized into 4 gray levels. We set in this example d=1 (as defined in the 3D GLCM neighborhood system). As illustrated by the 3D GLCM neighborhood system, 13 directions in total can be used to define the number of times two pixels of intensities i and j separated by D=1 occurs in each direction d (the 13 considered directions are defined between the center pixel (yellow node) and each of the the numbered nodes). Thirteen co-occurrences in total can be computed for a 3D volume and thirteen 12-haralick feature vectors can be derived. The final GLCM feature vector is the mean between the thirteen 12-haralick feature vectors.

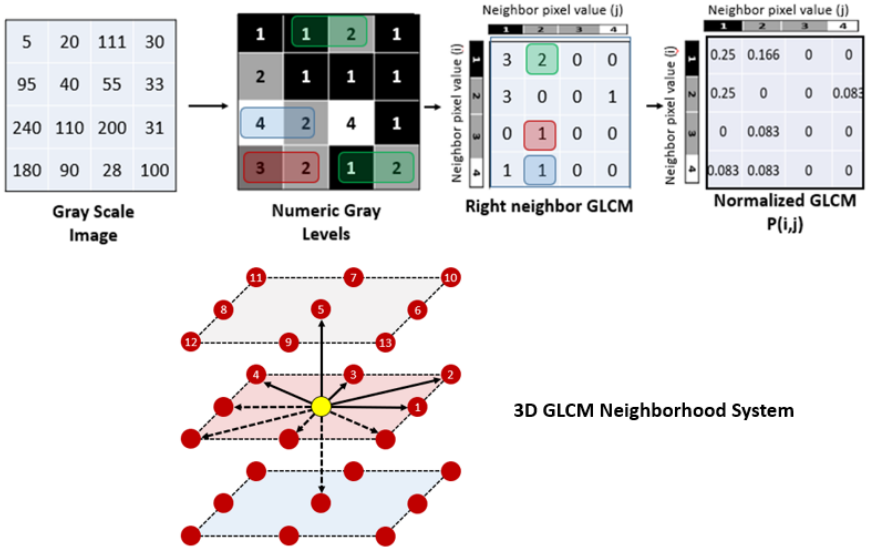


FIGURE 15: The GLCM features extraction process for the example of a 4 x 4 image quantized into 4 gray levels. In this example, D=1 and d=1.

In our experiments, the optimal GLCM feature extraction parameters have been selected empirically using a Grid search on the training data. We varied N_g from 8 to 64 by an increment of 8, d from 4 to 13 and D from 1 to 3, we found that $N_g = 16$, $d = 13$ and $D = 1$ achieved the best classification results.

A.3. CNN features

Deep learning of features has proven to be very successful in diverse domains and applications. In particular, CNNs have proven to be very effective for image recognition and classification tasks, thus our choice to include them in our analysis. For these features, we use the same number of instances per bag and the same size of volumes as those used with GLCM features.

In a typical CNN, input images are required to have the same size. Since we consider different volumes (instances) of various sizes, we resize them using spline interpolation [100]. Resizing can affect the resolution of the images, but since the factor that we are applying is not too large, resizing did not have significant effect on the results. After resizing, all instances are processed using the same CNN architecture and are represented by a feature vector with the same dimension.

The layers of our CNN architectures have been designed to minimize the risk of overfitting and optimize the classification accuracy. Based on these criteria, we fixed the CNN block that is used to extract features from each volume (instance) to include the following sequential convolutional and max-pooling layers. Each instance (volume) of size 31 x 31 x 10 is input to a first convolutional layer composed of 8 filters with dimension (3*3). The resulting feature map then goes through subsampling using a max-pooling layer. The second convolutional layer consist of 16 filters with dimension (3*3). The resulting feature map goes also through subsampling using a max-pooling layer. The third and last convolutional layer consists of 32 filters. At this stage a flattening layer is responsible of outputting a final feature vector of length 512. For regularization, dropout layers [101] with probability 0.25 were implemented after the last covolutional layer. We initialize the network parameters using random initialization and use Adam optimizer [102] to minimize the binary cross entropy loss function with a learning rate set to 0.0001. All networks were trained end-to-end for 200 epochs and updated with a 32-batch size. Figure 16 illustrates the proposed CNN architecture.

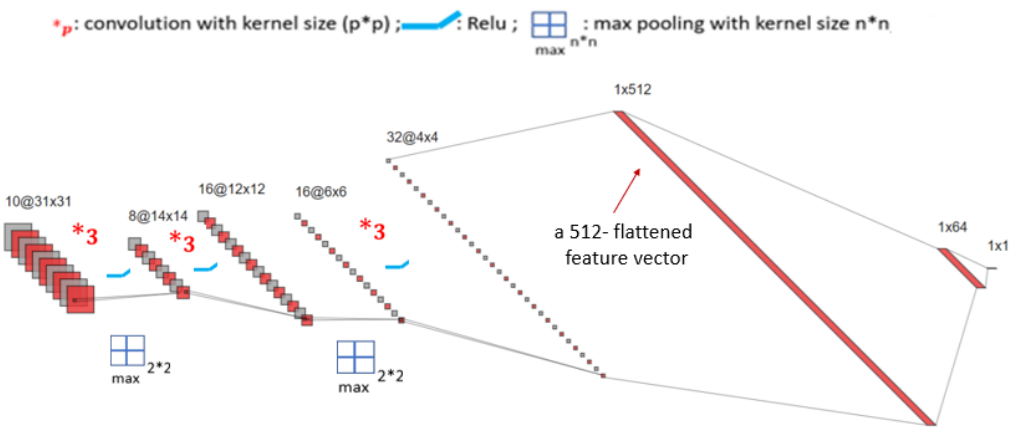


FIGURE 16: Architecture of the CNN block used for feature extraction.

B. Experimental Settings for benchmark MIL algorithms

We propose three different experiments in this thesis: the first one consists of training and testing on LIDC benchmark data using five-fold cross-validations. The second one consists of training and testing on the NLST data also using five-fold cross-validations. The third experiment consists of training on NLST and testing on LIDC.

Our comparative analysis was designed to include multiple benchmark MIL algorithms that belong to different categories (instance space, bag space and embedded space paradigms). Our objective is to experiment with diverse algorithms and identify the category of MIL algorithms that are more effective for the lung nodules classification task. From the category of instance-space MIL, we included the EM-DD [79] and MI-SVM [73]. From the category of bag space MIL, we included the mi-Graph [81], and from the category of embedded space MIL we included the mi-FV [84] and mi-VLAD [86]. In addition, we analyse our proposed MIL-CNN algorithms and compare them to the above benchmark MIL algorithms.

B.1. Parameters setting for MI-SVM

This MI-SVM was outlined in chapter II.C.1. It extends Support Vector Machines (SVMs) to multiple instance data. We selected this benchmark algorithm because the SVM classifier has proven to be more effective than other classifiers for lung nodules classification [26, 33]. Based on this, we hypothesized that MI-SVM should be a good candidate for our considered application. Since the inner structure of MI-SVM involves SVM with RBF kernel, the important parameters for this algorithm are:

- Cost (C): This cost parameter affects the optimization of the fit of the margin that separates both classes while penalizing the samples inside the margin. A low value for Cost will result in a low error while a large value for Cost will result in a high error. However, a low error on the training data does not necessarily lead to low prediction error on test data and may lead to overfitting. The value of Cost needs to be chosen carefully. Thus, there are no established rules to choose the optimal Cost parameter. In general, the optimal choice depends on the used dataset and typically this parameter is learned by applying a grid search.

- Gamma (γ): This parameter is associated with the RBF kernel and controls the degree of similarity between two points. The choice of Gamma controls the complexity of the decision boundary. Similar to the Cost variable, a high value of Gamma may lead to overfitting the training data. Typically, the optimal value of Gamma also depends on the dataset and should be tuned using a grid search.

- The number of important instances selected from each bag after each iteration to be used for training in the subsequent iteration.

In our experiments, using the NLST and LIDC datasets and a grid search, we identified the optimal parameters as: Cost=1, Gamma=10 and number of instances =50. These parameters will be fixed to these values for all experiments reported in this chapter.

B.2. Parameters setting for EM-DD

The performance of most existing MIL algorithms can degrade when the number of instances in each bag is large. The EM-DD [79] algorithm outlined in chapter II.C.2, is relatively robust to the number of relevant attributes in the dataset and scales up well to bags with large number of instances. Since our application involves bags with large number of instances (eg. 60), we opted to include the EM-DD in our comparative analysis.

The most important parameters of this algorithm include the termination criteria for each optimization (M-step), the probability above which a bag is considered as positive, the number of runs that are used to compute these probabilities and the method that is used to compute the final probability from all runs. In our experiments, we use the default parameters recommended by the authors [79]. Specifically, we set the probability threshold above which a bag/instance is classified as positive to 0.5, the M-step termination criterion to 0.1 and the number of runs using different starting points to 10. We report the results averaged over the 10 runs.

B.3. Parameters setting for mi-Graph

MI-SVM and EM-DD (and many other MIL algorithms) treat instances in the bags as independently and identically distributed (i.e.d). However, in most real applications, the instances within each bag are rarely independent and the i.e.d assumption may degrade the performance. In the considered application of lung nodules classification, it is likely that instances extracted from compact volumes around the nodules to be more informative and more important than those extracted from the volumes that are much smaller or larger than the nodule.

The mi-Graph algorithm [81], outlined in chapter II.C.3, does not assume instances within a bag to be independent and represents them by a graph. Thus, we selected this algorithm as another candidate for our analysis. Similar to the MI-SVM algorithm, the

mi-Graph integrates the SVM classifier. Thus, Cost and Gamma are also important parameters for this algorithm. As in MI-SVM, we set these parameters, after a grid search, to Cost=70 and Gamma=2. To build a graph for each bag, the mi-Graph treats each instance as a node and computes the pair-wise distances between instances. If two instances are similar (distance below a threshold), then, the two nodes are connected by an edge. In our experiments, we use the cosine distance and we set the distance threshold to 0.2.

B.4. Parameters setting for miFV and miVLAD

The MiFV [84] and miVLAD [86] algorithms were outlined in chapter II.C.4 and chapter II.C.5. These algorithms have been developed in order to deal with large scale problems in MIL since most existing MIL algorithms (eg. MI-SVM, EM-DD, mi-Graph and many others) can handle only small or moderate-sized data. Our considered datasets are composed of 2411 nodules from the NLST and 647 sample from the LIDC datasets where each nodule is represented by a bag of large number of instances (60). Since these datasets are large (compared to benchmark MIL datasets), we hypothesized that both miFV and miVLAD should be good candidates for our considered application.

The most important parameters for miFV/miVLAD include the Principal Component Analysis (PCA) parameter that uses orthogonal projections to convert a high dimensional vector into a more concise one and the number of centers that represent the number of Gaussian components/number of centroid in K-means clustering. In our experiments, we use the tuned parameters that achieved the best classification results in the original paper [86]. Specifically, we set the PCA value to 1 to keep all initial features without reduction and the number of centers to 2.

C. Parameters Setting for the proposed MIL-CNN algorithms

All of our proposed MIL-CNN architectures (described in chapter III) are composed

of 2 main blocks: a CNN block for feature extraction (described in chapter III section B) and a multiple instance neural network block for classification. The architecture of the neural network block contains four fully connected layers and use the tuned parameters that gave the best classification results in the original paper [87]. The number of neurons in each FC layers of the Mi-CNN, MI-CNN and MI-CNN with Deep Supervisions networks are set to 256,128, 64, and 1 while the number of neurons for the experiments of MI-CNN with Residual Connections are equal to 128, 128, 128, and 1 respectively. Each network is followed by a dropout layer with a 0.5 ratio. For all architectures, the MIL pooling layer is the max pooling layer. The weights of the fully connected layers are initialized randomly. We add the loss to each bag score and train the proposed networks by back propagation using SGD.

D. Experiments designed to investigate research questions RQ1-RQ5

Our objective is to build a lung nodules classification system that can efficiently and reliably classify lung nodules from CT images. Toward this goal, we formulated five research questions to guide us in identifying and validating the best classification method. In this section, we investigate each research question by formulating and conducting appropriate experiments and analyzing the results.

D.1. RQ1: What is the best MIL algorithm for this application ?

To investigate this question, we evaluate and compare all previously selected benchmark MIL algorithms and our proposed MIL-CNN algorithms. These algorithms are based on different approaches and belong to different categories of MIL algorithms. We adapted these algorithms to the lung nodules classification task, and one of our main goals is to identify which MIL category is best suited for this application.

We design and report the results of three experiments applied on both generative

(GLCM) and learned (CNN) features. Two of these experiments perform five-fold cross-validations on the same dataset (NLST or LIDC). We will refer to these as EXP_{NLST} and EXP_{LIDC}. The third experiment consists of training the algorithms on one collection (NLST dataset) and testing it on the other collection (LIDC dataset). We will refer to this experiment as $Exp_{NLST-LIDC}$. Figure 17 displays the AUC results for the GLCM features. As it can be seen, for all algorithms, the highest accuracies are obtained for EXP_{LIDC}. The LIDC and NLST datasets were collected using different sensors and have different resolution. Thus, a direct comparison between EXP_{LIDC} and EXP_{NLST} is not informative. However, we use these three different experiments to join more insights about the behaviour of the different algorithms.

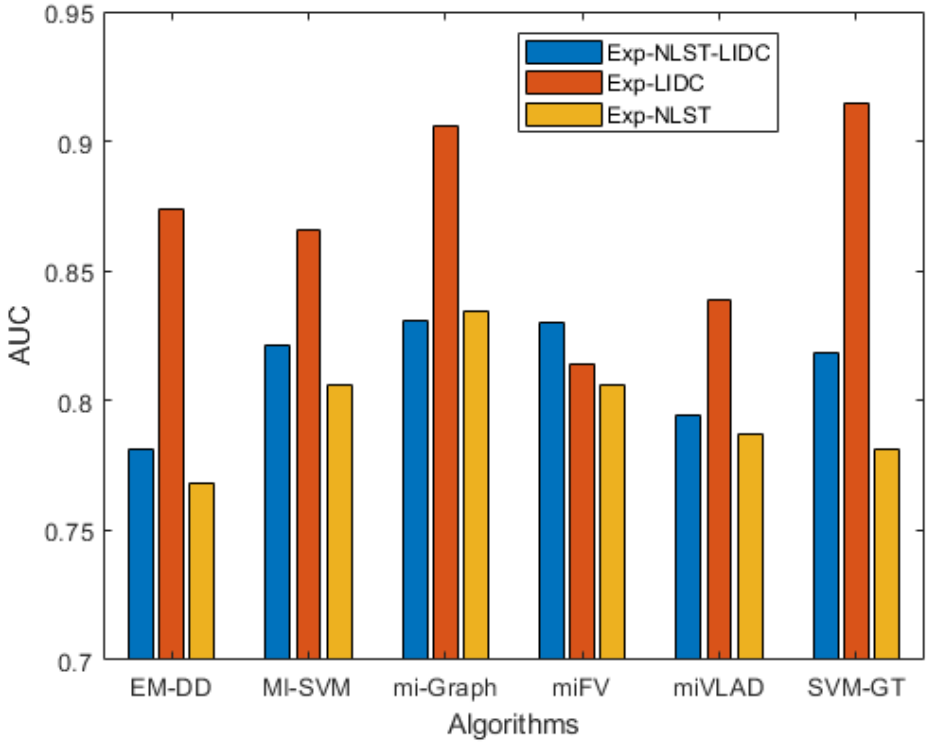


FIGURE 17: Accuracies of the different MIL algorithms for the three experiments using GLCM features. SVM-GT is a single instance classifier that relies on the ground truth bounding box identified for each nodule.

One factor that may affect the performance of the classifier is the size of nodules. To investigate this, in Figure 18 we display the distribution of the nodules' diameter (computed from the mid-slice of each ground truth volume) of both datasets. This figure shows that LIDC dataset samples are more distinct (positive samples tend to be larger than negative samples) when compared to NLST dataset samples. In other words, malignant samples are more extensive in size than benign nodules for the LIDC dataset, where for the NLST data, samples from both classes have comparable sizes. The distributions in Figure 18 justify the main factor of the discrepancy in the AUC in Figure 17.

In general, in machine learning, training on one data and validating on a different data is more challenging, but it is a more reliable indicator of the algorithm's performance. In fact, this is how a learning algorithm will be used in real scenarios. In addition $Exp_{NLST-LIDC}$ is less prone to overfitting. Thus, we use this experiment to rank the different algorithms.

A more detailed comparison between the different algorithms is provided in the following subsections.

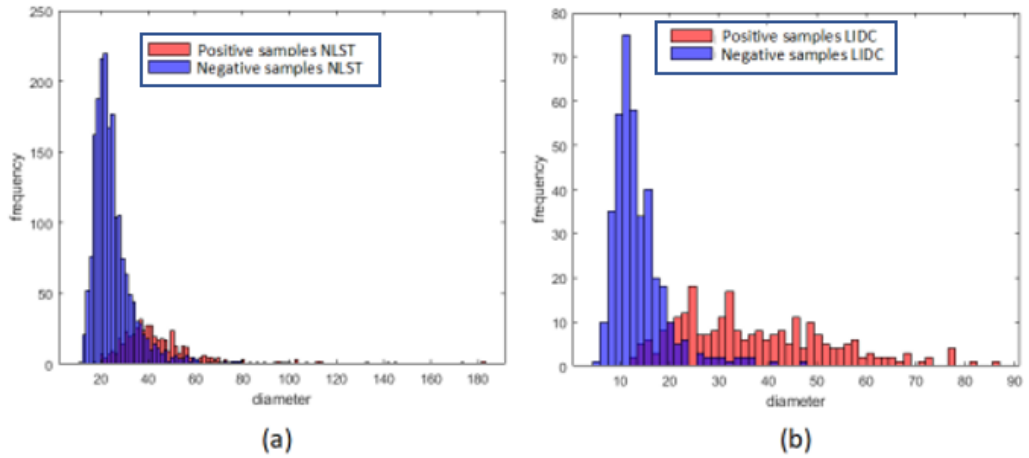


FIGURE 18: Distribution of the samples diameter for: (a) the NLST dataset, and (b) the LIDC dataset.

D.1.a. Comparison of the different MIL algorithms using the GLCM features

As shown in Figure 17, MIL algorithms that achieved the best classification results are based on the SVM classifier (i.e., mi-Graph and MI-SVM). In fact, many researchers used the SVM in their methods to detect and identify multiple types of cancers [103]. As in our results, they also confirmed that SVM based algorithms achieved higher accuracies compared to other classification algorithms.

Among all MIL algorithms used in our experiment and reported in Figure 17, the mi-Graph is the only one that treats instances as non-identical and dependent. This graph-based algorithm considers the relationships among instances to make decisions. This may be the main reason that this method achieved the best classification results with GLCM features compared to all other considered algorithms.

In Figure 17, we also include the results of a single instance classifier (SVM-GT) that uses one single feature vector extracted from the ground truth bounding box. As it can be seen, mi-Graph which doesn't assume known bounding boxes and uses multiple representations, achieves higher accuracy than the SVM classifier applied on the ground truth bounding boxes.

D.1.b. Comparison of the different MIL algorithms that use extracted CNN features

This section reports the results using extracted features with CNN (as described in chapter III, section B). In addition to all benchmark MIL algorithms, we also include the results of a single instance classifier (CNN) that uses one feature vector extracted from the ground truth bounding box. This result is noted by CNN-GT in Figure 19. We also include the results of SVM-GT as described earlier.

For these features, MI-SVM performed better than mi-Graph and achieved the best classification results among all considered MIL algorithms. Possible reasons for SVM to outperform mi-Graph in this case may be due to the robustness of SVM in dealing with high-dimensional data. They may also be due to the sensitivity of the graph construction to

the high dimensionality.

CNN-GT and SVM-GT achieved lower results than some MIL algorithms (i.e., EM-DD, MI-SVM). This may be because the inclusion of some parenchymal structures (when larger volumes are used to extract some of the instance features) has proved to correlate with better classification results [89, 90, 104].

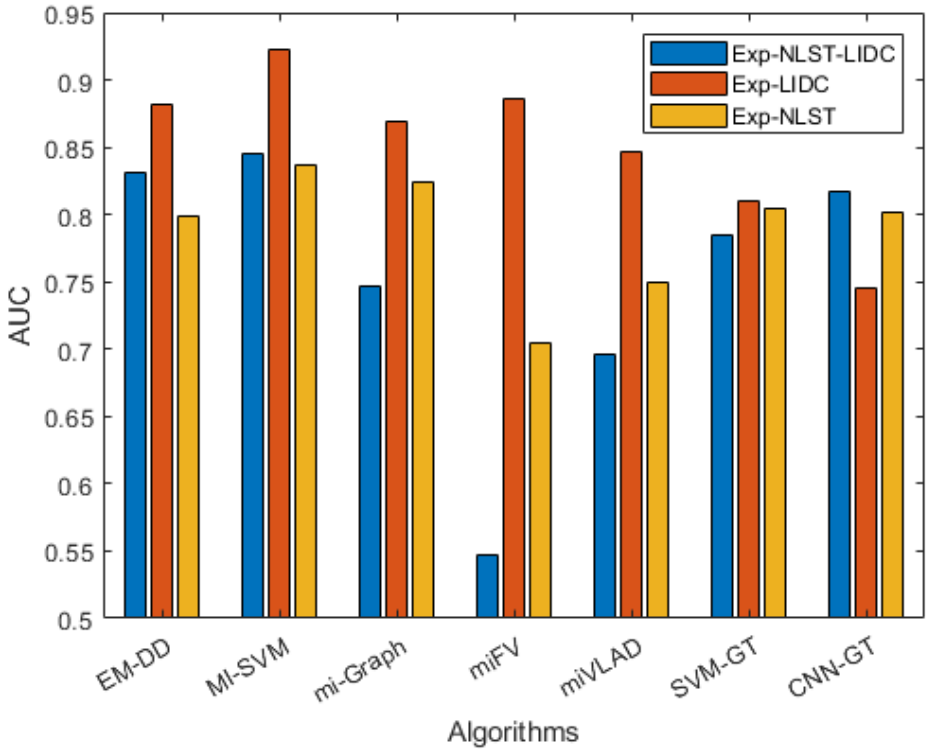


FIGURE 19: Accuracies of the different MIL algorithms for the three experiments using CNN features with traditional MIL algorithms. SVM-GT and CNN-GT are single instance classifiers that use one feature vector extracted from the ground truth bounding box.

D.1.c. Results of the proposed end to end MIL-CNN architectures

In Figure 20, we display the results of the four proposed end-to-end MIL-CNN architectures. For comparison purposes, we also display the CNN-GT and SVM-GT results that use a single instance extracted from the ground truth bounding box as described in previous sections. As it can be seen, the proposed MIL-CNN algorithms achieve higher accuracies

than the single instance classifier. Moreover, MIL-CNN which belongs to the bag space paradigm, outperformed algorithms that belong to the instance level paradigm. This is in agreement to many published comparative studies of benchmark MIL algorithms [22, 105]. These studies stated that bag level and embedded level paradigms perform better than instance ones because they provide an appropriate framework for exploiting information from the whole bag. In addition, the bag space networks (MI-CNN-DS and MI-CNN-RC) have more accurate results when compared to MI-CNN. Typically, Residual Connections are used to reduce the effect of Vanishing Gradient problem on deep networks. However, in our case, they were introduced as a way to learn good instance features since training neural networks using complex Multiple Instance data is a challenging task. As illustrated in Figure 20, the inclusion of Residual Connections achieved better classification results.

MI-CNN-RC learns numerous bag features from all different levels of instance features by MIL pooling. It can combine the various hierarchies using the efficient Residual Connections to get a final and better bag classification accuracy. Thus, we can conclude that the MIL-CNN approach with Residual Connections is the best classifier for the lung nodules classification task.

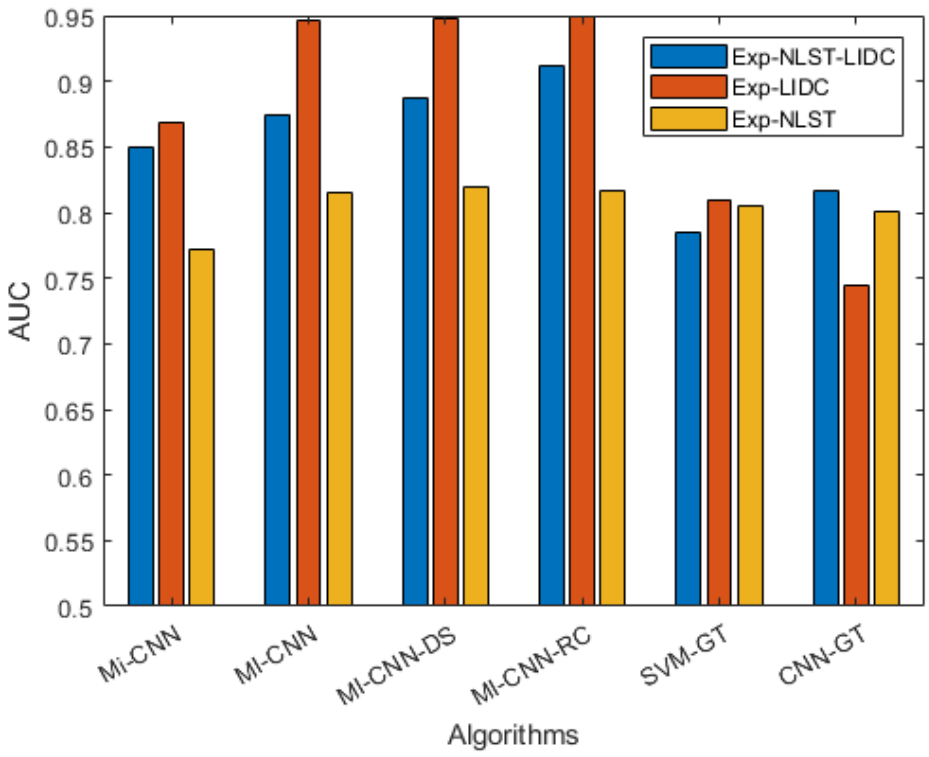


FIGURE 20: Accuracies of the different MIL algorithms for the three experiments using end-to-end MIL-CNN algorithms. SVM-GT and CNN-GT are single instance classifiers that use one feature vector extracted from the ground truth bounding box.

D.2. **RQ2: What are the main nodule parameters that influence the classification ?**

For this task, we analyse the effect of various nodule parameters on the classification accuracy. In particular, we investigate the effect of the nodule's volume, middle slice diameter, and contrast. The contrast represents the difference in intensity between the object (i.e., nodule) and its background. We calculate the contrast between the pixels included in the ground truth bounding box and the pixels included in a larger box that contains more background excluding the bounding box, using:

$$Contrast = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} \tag{24}$$

In 24, μ_1 and σ_1 are, respectively, the mean and standard deviation of the pixels within the ground truth box, while μ_2 and σ_2 are respectively the mean and standard deviation of the pixels within the larger box. These steps are illustrated in Figure 21.

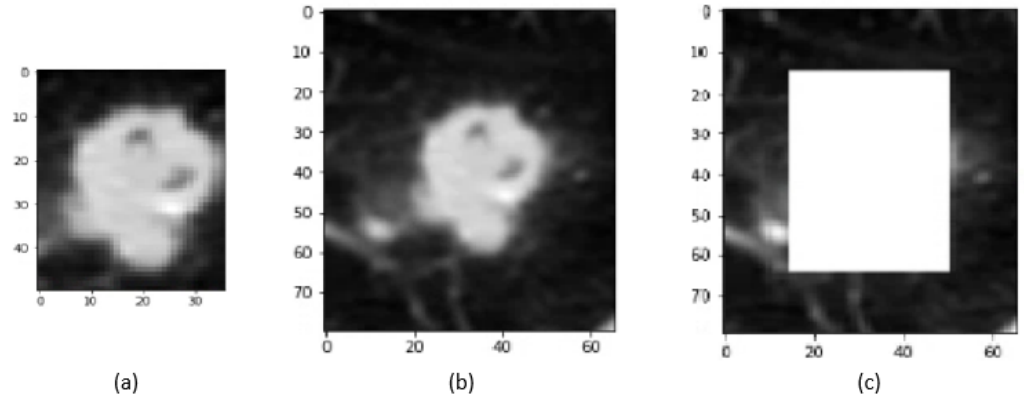


FIGURE 21: Different steps for the contrast computation: (a) the ground truth box that includes the nodule. μ_1 and σ_1 are computed using all pixels within this box, (b) larger box that includes the nodule and more background and (c) pixels used to compute μ_2 and σ_2. These pixels belong to the larger box excluding the ground truth box.

For this analysis, we threshold the output of the classifier and classify each test nodule as benign or malignant. For each algorithm, the threshold is selected to yield the same fixed true negative rate (=80%). For this experiment, we analyse the output of MI-CNN-RC and MI-CNN-DS using the LIDC dataset. Let P^+ denote the true malignant (positive) samples in the validation data, and let N^- denote the true benign (negative) samples. Let $\hat{P}^-$ denote the malignant nodules that are miss-classified by both algorithms as benign. In Figure 22 we display the distributions of the diameter, volume and contrast of all nodules in P^+, N^- and $\hat{P}^-$.

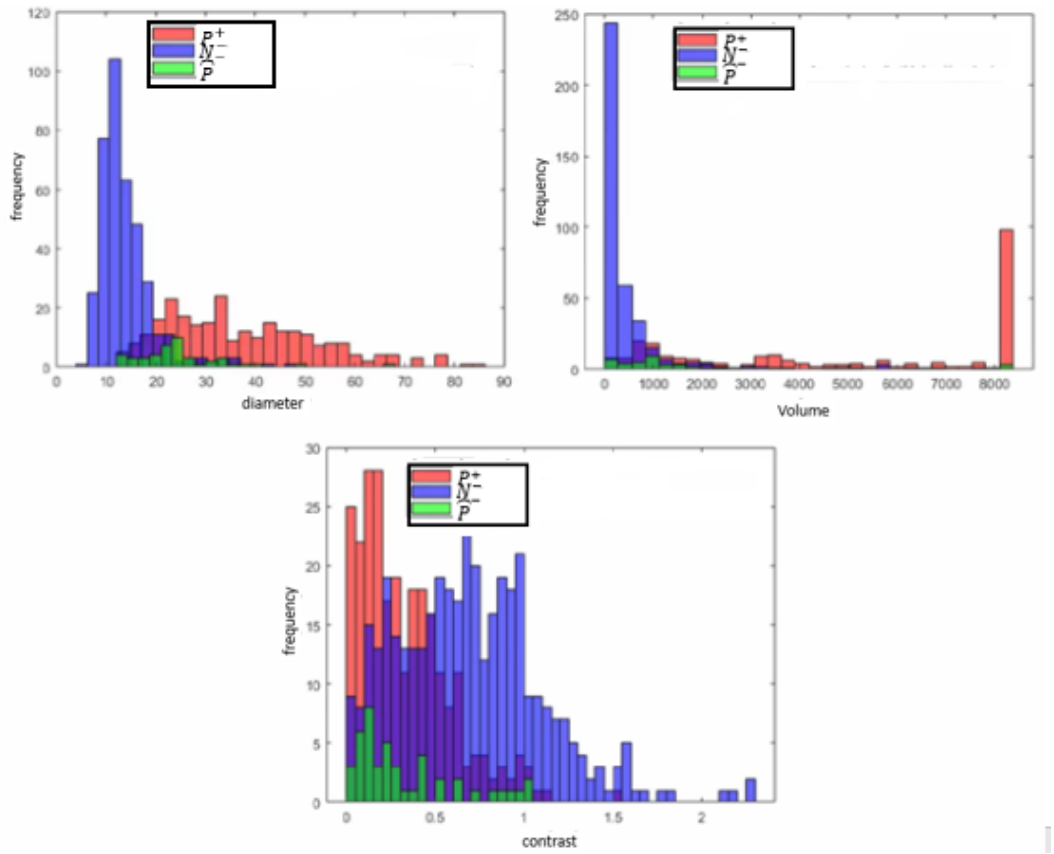


FIGURE 22: Nodules parameters that can influence the classification: (a) 2D diameter of each nodule from the center slice of the 3D volume, (b) volume of each nodule and (c) contrast of the nodules.

For the parameters related to the size (i.e., diameter and volume), we can see that miss-classified positive samples have a distribution closer to the true negative samples than from the distribution of the true positive samples. Thus we can conclude that the size of the nodule is an important factor that can affect the classification. This is in agreement with the radiologists process in analysing CT scans. In fact, if the nodule size is small (less than 4 mm), the radiologist will ask for another scan after six months to check the growth rate of the nodule. On the other hand, if the nodule size is large (more than 20 mm), then the radiologist is almost sure it is a malignant nodule.

For the contrast measure, the miss-classified samples can be close to the distribution of P^+ and N^-. Thus, we may conclude that this measure does not influence the

classification results.

D.3. **RQ3: What is the best feature representation ?**

To investigate this question, we summarize the best AUC results for both features for the train on NLST and test on LIDC experiment in Table 1. As mentioned earlier, MI-CNN-RC has obtained the best AUC results for CNN features, while mi-Graph has the best AUC results for GLCM features. As it can be seen, the CNN features perform significantly better than the GLCM.

Experiment	mi-Graph	MI-CNN-RC
Train on NLST and test on LIDC	0.8308	0.9117

TABLE 1: Best AUC results for the train on NLST and test on LIDC experiment for both GLCM and CNN features.

To compare the two methods, we analyse few miss-classified samples. First, we analyse the false negatives vs. the size of the nodules since as established earlier, this parameter can affect the classifier outcome. For this analysis, we threshold the output of the classifier and classify each test nodule as benign or malignant. For each algorithm, the threshold is selected to yield the same fixed true negative rate (=80%). Let $\hat{P}^{-}_{GLCM}$ and $\hat{P}^{-}_{MI-CNN-RC}$ denote the true malignant nodules that are miss-classified by each algorithm as benign. In Figure 23, we display the frequency of $\hat{P}^{-}_{GLCM}$ and $\hat{P}^{-}_{MI-CNN-RC}$ vs. the 2D diameter of each nodule from the center slice of the 3D volume and vs. the volume of each nodule.

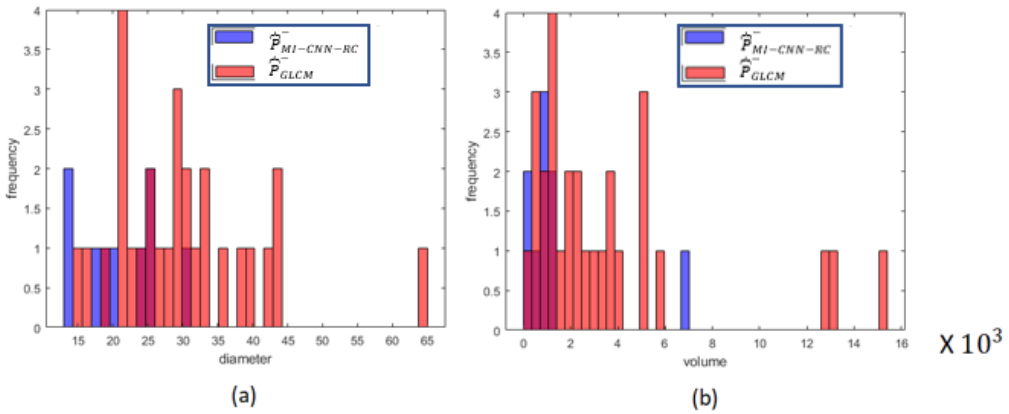


FIGURE 23: False-negative samples for both GLCM and CNN features vs.: (a) the diameter of the nodule of the middle scan, and (b) the volume of the nodule.

First, we note that the CNN-based classifier missed fewer malignant nodules, and all missed nodules are very small. On the other hand, the classifier based on GLCM features miss-classified many samples with large sizes.

One possible explanation for the miss-classification of large nodules by GLCM features may be due to the variation of the spacial distribution of Hounsfield values for few nodules between different nodule types. These variations are expected to be less smooth for malignant nodules than benign ones. To illustrate this, in Figure 24 we display 3D visualization of Hounsfield values for few nodules. In Figure 24 (a) we visualize Hounsfield values for true malignant samples. As it can be seen, there is a large variation within the pixels of the nodules. In Figure 24 (b), we visualize the Hounsfield values for benign nodules where the values are more uniform. In Figure 24 (c), we visualize the Hounsfield values for few of the miss-classified malignant samples. As it can be seen, they tend to have a similar distribution that is more similar to the benign nodules shown in Figure 24 (b).

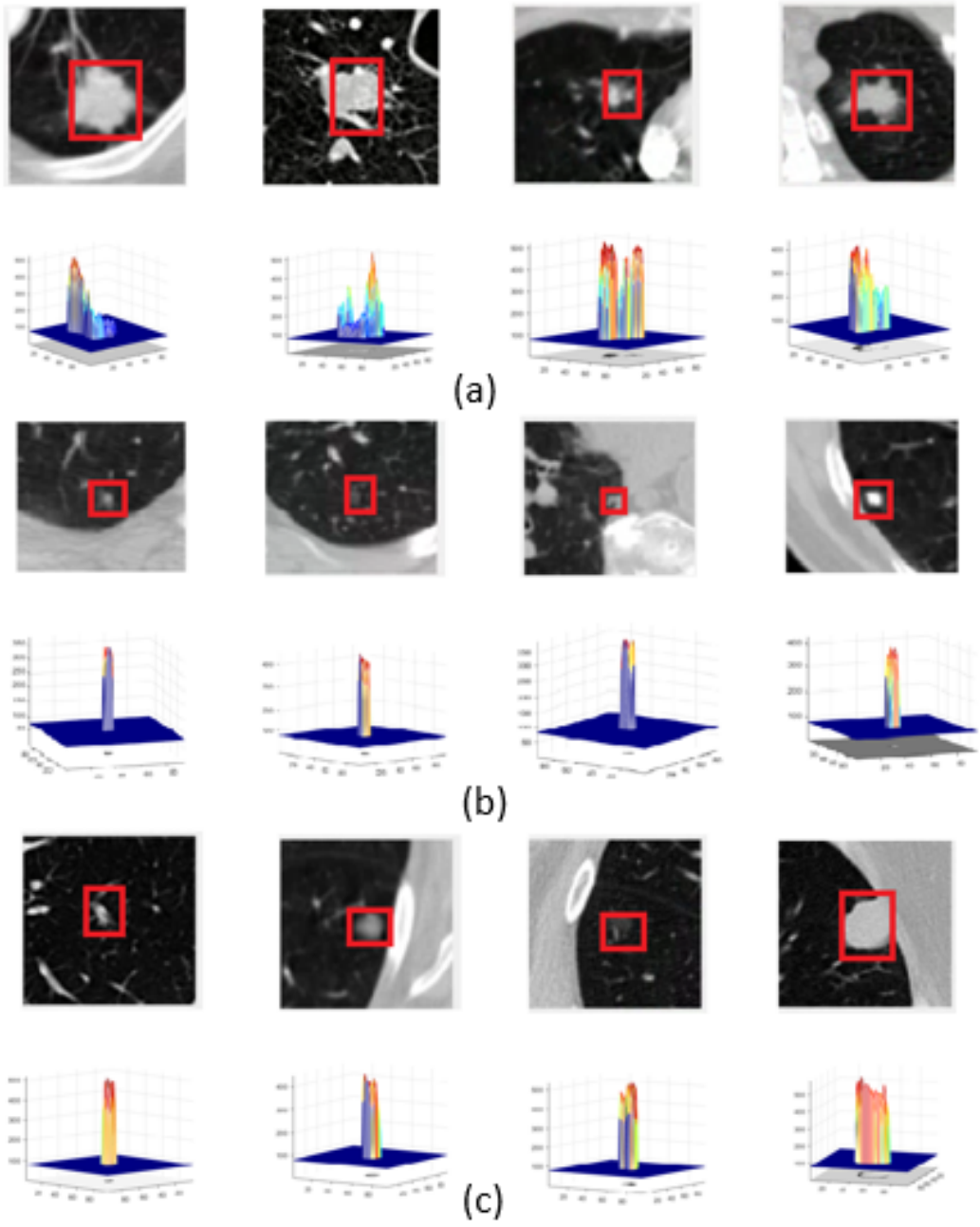


FIGURE 24: Visualization of the Hounsfield values distribution for few nodules that are classified using GLCM features: (a) correctly classified malignant nodules, (b) benign nodules and (c) miss-classified malignant samples.

The learned CNN features achieved better results than the generative GLCM features because the latter ones are more sensitive to image preprocessing and nodules seg-

mentation. This fact may affect the accuracy of the classification of each sample compared to deep learned features that do not rely on segmentation or preprocessing. CNN learned features are extracted automatically to solve a specific task, and filters at multiple layers can be trained to learn various distinctive features for a particular class. This fact may explain why deep learning models are becoming the standard approach for tasks that involve visual image classification. CNN features perform well in the lung nodules classification task because features can be learned at multiple layers to discriminate between different classes: low-level learned features are associated with the early layers of the CNN, and usually extract edges of each nodule. In contrast, high-level learned features are associated with the late layers of the CNN and identify fine details and more in-depth information related to each nodule such as texture and size.

D.4. **RQ4: Can an MIL algorithm learn to identify the positive instances (among a bag with large number of candidates) that are close to the ground truth ?**

To investigate this question, we analyze the spatial similarity between the nodule's ground truth volume radiologists and the volumes identified by our instance space Mi-CNN algorithm as positive instances. For similarity, we use the 3D Intersection over Union (3D-IoU). Intersection over Union (IoU) is a benchmark evaluation metric for many tasks, including object detection, segmentation, and tracking. For example, in image segmentation, IoU is used to quantify how much a segmented region agrees with the provided ground truth region. IoU is computed using:

$$IoU = \frac{A \cap B}{A \cup B} = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{I}{U} \tag{25}$$

where A is the segmented region and B is the ground truth region.

In our experiment, we use the 3D-IoU to identify the instance in each bag that was extracted from the volume that has the highest IoU with the ground truth volume. Let I^* refer to this instance.

We evaluate Mi-CNN algorithm by first ranking instances in each positive bag by decreasing order of confidence values. Next, we check if I^* is one of the top K instances with the highest confidence. We repeat this for all test bags and compute the percentage of time that I^* is among the top K instances. These results are reported in Table 2 for $K = 1..10$.

K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
11.19%	21.64%	26.12%	28.32%	30.52%	31.22%	32%	32%	32.73%	40%

TABLE 2: Percentage of time I^* is among the top K instances of each bag when tested using Mi-CNN.

Examples of nodules where the closest instance to the ground truth region (i.e., I^*) is among the top two positive instances are illustrated in Figure 25.

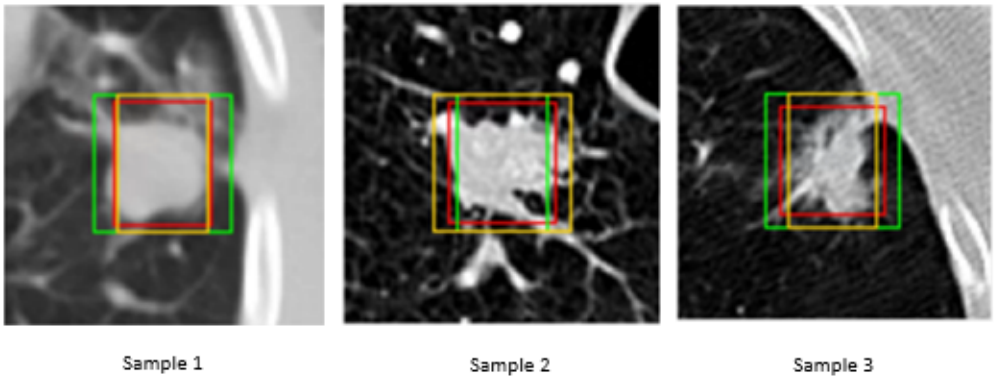


FIGURE 25: Three examples of nodules for which the top two positive instances are the closest to the ground truth region. The red box represents the ground truth bounding box, the yellow box represents the first top positive instance, and the green box represents the second top positive instance.

From Table 2, we notice also that out of 60 instances, only 40% of the time I^* is among the top 10 instances with the highest confidence. In other words, for most of the times, the Mi-CNN classifies the nodule correctly using a volume that is different from the

ground truth volume identified by the radiologist. In the next section, we will justify this behaviour.

D.5. **RQ5: What is the best learning approach for this application: Single Instance Learning (SIL) applied to the ground truth region or Multiple Instance Learning (MIL) applied to multiple regions around the ground truth ?**

As concluded from RQ1, some MIL algorithms achieved higher classification accuracies than SIL algorithms applied on the ground truth region. To justify these results and highlight the usefulness of MIL for the lung nodules classification task, we identify few samples that have been classified correctly by Mi-CNN and miss-classified by CNN applied on the ground truth region. We visualize these samples to identify possible reasons for this behaviour. In Figure 26, we display three such samples. As it can be seen, the two instances closest to the ground truth volume were not among the top instance that have the highest confidence values. Instead Mi-CNN identified other instances that have larger volumes that included parenchymal structures. In fact, multiple studies [89,90,104] proved that the inclusion of parenchymal tissues can improve the classification results compared to using the ground truth volume only, as surrounding texture provide relevant information about the nature of the nodule. In [89], the authors found that texture features extracted from the nodules were not significantly different between malignant and benign cases; however, when the parenchyma was included, texture features were more discriminative. The authors stated that the texture might be quantifying vascularization within the parenchyma, tumor speculation, and parenchymal tissue compression as the lesion invades into the parenchyma.

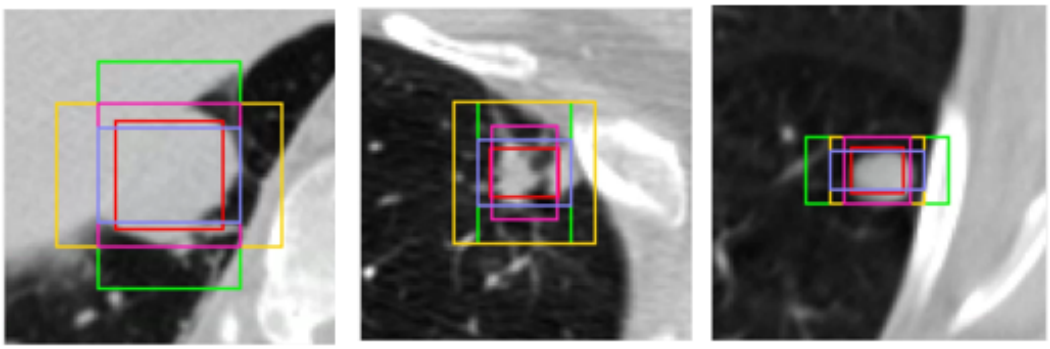


FIGURE 26: Three sample nodules classified correctly by Mi-CNN and miss-classified by CNN applied on the ground truth. For each sample, the red box represents the ground truth bounding box, the yellow and green boxes represent the top two positive instances and the pink and magenta boxes represent the two instances that are closest to the ground truth box.

Another category of nodules that were classified correctly by Mi-CNN and miss-classified by CNN include nodules for which radiologists did not agree in their nodule segmentation and identified different volumes for the same nodule. In this case, the average volume (used as ground truth) may not be the best region of interest. Three samples of these nodules are illustrated in Figure 27.

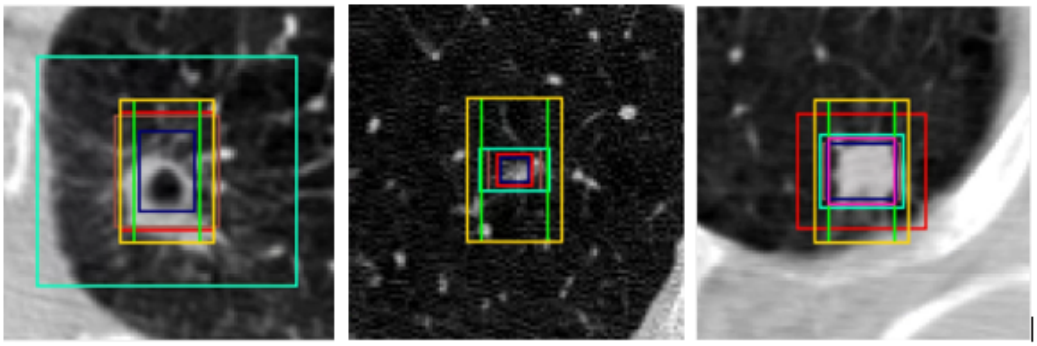


FIGURE 27: Three sample nodules classified correctly by Mi-CNN and miss-classified by CNN applied on the ground truth. The red, cyan and brown boxes represent the ground truth identified by radiologists 1, 2 and 3 respectively, the blue box represents the average volume between all radiologists (used as ground truth) and the yellow and green boxes represent the top two positive instances selected by Mi-CNN.

For sample 1, we notice that boxes that have been identified as positive instances by Mi-CNN (green and yellow) agree with radiologists 1 and 3 (red and brown) and disagree with radiologist 2 (cyan). For sample 2, we notice that boxes that have been identified as positive instances by Mi-CNN (green and yellow) are larger than the ground truth volumes identified by radiologists 1, 2 and 3 (red, cyan and brown) respectively. This confirms that including background with nodules texture can improve the classification results and that radiologists contour may not be the best representation of each nodule. For sample 3, we notice that boxes that have been identified as positive instances by Mi-CNN (green and yellow) are of average sizes compared to ground truths identified by radiologist 2 and 3 (small cyan and brown boxes) and radiologist 1 (large red box).

The previously reported results (AUC in Figure 20) and the above justification indicate that MIL algorithms are more effective than SIL for the lung nodules classification task.

CHAPTER VI
CONCLUSION AND FUTURE WORK

In this thesis, we adapted and investigated the application of Multiple Instance Learning (MIL) in the lung nodules diagnosis task. We explored multiple benchmark MIL algorithms (EM-DD, MI-SVM, mi-Graph, miVLAD, miFV). We also proposed an end-to-end MIL-CNN algorithms (Mi-CNN, MI-CNN, MI-CNN-DS, MI-CNN-RC). We evaluated our models on two different features: learned features (CNN features) and generative features (GLCM features).

We evaluated our models on two different collections (LIDC and NLST) containing malignant and benign nodules. We asked multiple Research Questions and answered them by elaborating on various experiments. We concluded that learned features with our proposed embedded space MIL-CNN algorithms are the most efficient for the studied task. We also concluded that MIL integration could improve the classification results over using SIL on a fixed size for all samples.

One orientation for our future work would be to increase the size of the input samples to the CNN framework and to use a more complex architecture for feature extraction, eventually comparing both frameworks' results.

Future work will also include investing in other deep learning frameworks for the extraction of learned features instead of using CNN features.

Another potential future work may include fusing both types of features and seeing if the fusion between them may improve the results. Some new researches [64, 106, 107] in the lung nodules diagnosis task are moving toward the inclusion of both generative and learned features in the same framework, and this could be a good direction for our work too.

Our proposed approach has the advantage of not requiring the nodule to be segmented and relies only on the center of mass of each nodule. This can still be a limitation as it requires some annotation by radiologists. One way to overcome this limitation is to randomly sample the image space and select center candidates. However, this will require a huge number of instances to increase the chances of selecting the true region of interest. An alternative approach would be to use a screening method [108] that assists in identifying specific locations of interest. These locations can then be considered as instances and grouped in a bag.

Finally, to overcome the potential overfitting issue that arises from using a small number of samples, additional patients should be recruited to validate the classification framework on larger datasets.

REFERENCES

[1] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[2] N. L. S. T. R. Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[3] W. Street. Cancer facts & figures 2019. *American Cancer Society: Atlanta, GA, USA*, 2019.

[4] J. Gong, J.-Y. Liu, X.-W. Sun, B. Zheng, and S.-D. Nie. Computer-aided diagnosis of lung cancer: the effect of training data sets on classification accuracy of lung nodules. *Physics in Medicine & Biology*, 63(3):035036, 2018.

[5] X. Zhao, L. Liu, S. Qi, Y. Teng, J. Li, and W. Qian. Agile convolutional neural network for pulmonary nodule classification using ct images. *International journal of computer assisted radiology and surgery*, pages 1–11, 2018.

[6] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, and Z. Liang. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of digital imaging*, 28(1):99–115, 2015.

[7] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, and N. Khandelwal. A combination of shape and texture features for classification of pulmonary nodules in lung ct images. *Journal of digital imaging*, 29(4):466–475, 2016.

[8] A. Shaffie, A. Soliman, M. Ghazal, F. Taher, N. Dunlap, B. Wang, A. Elmaghraby, G. Gimel'farb, and A. El-Baz. A new framework for incorporating appearance and shape features of lung nodules for precise diagnosis of lung cancer. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 1372–1376. IEEE, 2017.

[9] A. Shaffie, A. Soliman, M. Ghazal, F. Taher, N. Dunlap, B. Wang, V. Van Berkel, G. Gimelfarb, A. Elmaghraby, and A. El-Baz. A novel autoencoder-based diagnostic system for early assessment of lung cancer. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1393–1397. IEEE, 2018.

[10] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015.

[11] R. Dey, Z. Lu, and Y. Hong. Diagnostic classification of lung nodules using 3d neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 774–778. IEEE, 2018.

[12] K. Liu and G. Kang. Multiview convolutional neural networks for lung nodule classification. *International Journal of Imaging Systems and Technology*, 27(1):12–22, 2017.

[13] S. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez. Optimal deep learning model for classification of lung cancer on ct images. *Future Generation Computer Systems*, 92:374–382, 2019.

[14] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[15] G. Ruffo. Learning single and multiple instance decision trees for computer security applications. *University of Turin, Torino*, 2000.

[16] Q. Zhang, S. A. Goldman, W. Yu, and J. E. Fritts. Content-based image retrieval using multiple-instance learning. In *ICML*, volume 2, pages 682–689. Citeseer, 2002.

[17] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5(Aug):913–939, 2004.

[18] C. Zhang and P. A. Viola. Multiple-instance pruning for learning efficient cascade detectors. In *Advances in neural information processing systems*, pages 1681–1688, 2008.

[19] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[20] M. Yousefi. *Computer-aided Detection of Breast Cancer in Digital Tomosynthesis Imaging Using Deep and Multiple Instance Learning*. PhD thesis, Concordia University, 2018.

[21] R. Venkatesan, P. Chandakkar, B. Li, and H. K. Li. Classification of diabetic retinopathy images using multi-class multiple-instance learning based on color correlogram features. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 1462–1465. IEEE, 2012.

[22] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[23] F. V. Farahani, A. Ahmadi, and M. F. Zarandi. Lung nodule diagnosis from ct images based on ensemble learning. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pages 1–7. IEEE, 2015.

[24] T. N. Shewaye and A. A. Mekonnen. Benign-malignant lung nodule classification with geometric and appearance histogram features. *arXiv preprint arXiv:1605.08350*, 2016.

[25] M. V. K. Likhitkar, U. Gawande, and M. K. O. Hajari. Automated detection of cancerous lung nodule from the computed tomography images. *IOSR Journal of Computer Engineering*, 16(1):05–11, 2014.

[26] M. Firmino, G. Angelo, H. Morais, M. R. Dantas, and R. Valentim. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *Biomedical engineering online*, 15(1):1–17, 2016.

[27] J. Jeeva et al. A computer aided diagnosis for detection and classification of lung nodules. In *Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on*, pages 1–5. IEEE, 2015.

[28] K. Chen, Y. Nie, S. Park, K. Zhang, Y. Zhang, et al. Development and validation of machine learning–based model for the prediction of malignancy in multiple pulmonary nodules: Analysis from multicentric cohorts. *Clinical Cancer Research*, 2021.

[29] J.-K. Kamarainen. Gabor features in image analysis. In *2012 3rd international conference on image processing theory, tools and applications (IPTA)*, pages 13–14. IEEE, 2012.

[30] T. Löfstedt, P. Brynolfsson, T. Asklund, T. Nyholm, and A. Garpebring. Gray-level invariant haralick texture features. *PloS one*, 14(2):e0212110, 2019.

[31] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004.

[32] M. Nishio and C. Nagashima. Computer-aided diagnosis for lung cancer: usefulness of nodule heterogeneity. *Academic radiology*, 24(3):328–336, 2017.

[33] M. B. Rodrigues, R. V. M. Da NóBrega, S. S. A. Alves, P. P. Rebouças Filho, J. B. F. Duarte, A. K. Sangaiah, and V. H. C. De Albuquerque. Health of things algorithms for malignancy level classification of lung nodules. *IEEE Access*, 6:18592–18601, 2018.

[34] P.-W. Huang, P.-L. Lin, C.-H. Lee, and C. Kuo. A classification system of lung nodules in ct images based on fractional brownian motion model. In *2013 International conference on system science and engineering (ICSSE)*, pages 37–40. IEEE, 2013.

[35] Z. Wang, J. Xin, P. Sun, Z. Lin, Y. Yao, and X. Gao. Improved lung nodule diagnosis accuracy using lung ct images with uncertain class. *Computer methods and programs in biomedicine*, 162:197–209, 2018.

[36] A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass. Computer-aided diagnosis of lung nodules in computed tomography by using phylogenetic diversity, genetic algorithm, and svm. *Journal of digital imaging*, 30(6):812–822, 2017.

[37] W. Huang and S. Tu. Su-f-r-22: Malignancy classification for small pulmonary nodules with radiomics and logistic regression. *Medical physics*, 43(6Part6):3377–3378, 2016.

[38] S.-J. Tu, C.-W. Wang, K.-T. Pan, Y.-C. Wu, and C.-T. Wu. Localized thin-section ct with radiomics feature extraction and machine learning to classify early-detected pulmonary nodules from lung cancer screening. *Physics in Medicine & Biology*, 63(6):065005, 2018.

[39] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

[40] J. Ma, Q. Wang, Y. Ren, H. Hu, and J. Zhao. Automatic lung nodule classification with radiomics approach. In *Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations*, volume 9789, page 978906. International Society for Optics and Photonics, 2016.

[41] A. Nibali, Z. He, and D. Wollersheim. Pulmonary nodule classification with deep residual networks. *International journal of computer assisted radiology and surgery*, 12(10):1799–1808, 2017.

[42] S. Hussein, K. Cao, Q. Song, and U. Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International conference on information processing in medical imaging*, pages 249–260. Springer, 2017.

[43] M. Al-Shabi, B. L. Lan, W. Y. Chan, K.-H. Ng, and M. Tan. Lung nodule classification using deep local–global networks. *International journal of computer assisted radiology and surgery*, pages 1–5, 2019.

[44] Y. Xie, Y. Xia, J. Zhang, D. D. Feng, M. Fulham, and W. Cai. Transferable multi-model ensemble for benign-malignant lung nodule classification on chest ct. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 656–664. Springer, 2017.

[45] Q. Song, L. Zhao, X. Luo, and X. Dou. Using deep learning for classification of lung nodules on computed tomography images. *Journal of healthcare engineering*, 2017, 2017.

[46] S. H. Mohammed and A. Çinar. Lung cancer classification with convolutional neural network architectures. *Qubahan Academic Journal*, 1(1):33–39, 2021.

[47] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha. Deep learning algorithm for autonomous driving using googlenet. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 89–96. IEEE, 2017.

[48] W. Nawaz, S. Ahmed, A. Tahir, and H. A. Khan. Classification of breast cancer histology images using alexnet. In *International conference image analysis and recognition*, pages 869–876. Springer, 2018.

[49] T. Akiba, S. Suzuki, and K. Fukuda. Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*, 2017.

[50] S. Ayyachamy, V. Alex, M. Khened, and G. Krishnamurthi. Medical image retrieval using resnet-18. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, page 1095410. International Society for Optics and Photonics, 2019.

[51] W. Sun, B. Zheng, and W. Qian. Computer aided lung cancer diagnosis with deep learning algorithms. In *Medical imaging 2016: computer-aided diagnosis*, volume 9785, page 97850Z. International Society for Optics and Photonics, 2016.

[52] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[53] G. E. Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[54] D. Zhao, D. Zhu, J. Lu, Y. Luo, and G. Zhang. Synthetic medical images using f&bgan for improved lung nodules classification by multi-scale vgg16. *symmetry*, 10(10):519, 2018.

[55] S. Suresh and S. Mohan. Roi-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis. *Neural Computing and Applications*, pages 1–21, 2020.

[56] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[57] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, 8, 2015.

[58] Y. Chen, X. Zhao, and X. Jia. Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2381–2392, 2015.

[59] M. Zhang, H. Li, S. Pan, J. Lyu, S. Ling, and S. Su. Convolutional neural networks based lung nodule classification: A surrogate-assisted evolutionary algorithm for hyperparameter optimization. *IEEE Transactions on Evolutionary Computation*, 2021.

[60] J. Sang, M. S. Alam, H. Xiang, et al. Automated detection and classification for early stage lung cancer on ct images using deep learning. In *Pattern Recognition and Tracking XXX*, volume 10995, page 109950S. International Society for Optics and Photonics, 2019.

[61] C. Zhang, X. Sun, K. Dang, K. Li, X.-w. Guo, et al. Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. *The oncologist*, pages theoncologist–2018, 2019.

[62] H. Polat and H. Danaei Mehr. Classification of pulmonary ct images by using hybrid 3d-deep convolutional neural network architecture. *Applied Sciences*, 9(5):940, 2019.

[63] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham, and W. Cai. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. *IEEE transactions on medical imaging*, 38(4):991–1004, 2018.

[64] Y. Xie, J. Zhang, Y. Xia, M. Fulham, and Y. Zhang. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Information Fusion*, 42:102–110, 2018.

[65] Y. Ren, M.-Y. Tsai, L. Chen, J. Wang, S. Li, Y. Liu, X. Jia, and C. Shen. A manifold learning regularization approach to enhance 3d ct image-based lung nodule classification. *International Journal of Computer Assisted Radiology and Surgery*, 15(2):287–295, 2020.

[66] S. B. Shrey, L. Hakim, M. Kavitha, H. W. Kim, and T. Kurita. Transfer learning by cascaded network to identify and classify lung nodules for cancer detection. In *International Workshop on Frontiers of Computer Vision*, pages 262–273. Springer, 2020.

[67] B. Veasey, M. M. Farhangi, H. Frigui, J. Broadhead, M. Dahle, A. Pezeshk, A. Seow, and A. A. Amini. Lung nodule malignancy classification based on nlstx data. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1870–1874. IEEE, 2020.

[68] N. Kalaivani, N. Manimaran, S. Sophia, and D. Devi. Deep learning based lung cancer detection and classification. In *IOP Conference Series: Materials Science and Engineering*, volume 994, page 012026. IOP Publishing, 2020.

[69] H. Jiang, F. Gao, X. Xu, F. Huang, and S. Zhu. Attentive and ensemble 3d dual path networks for pulmonary nodules classification. *Neurocomputing*, 398:422–430, 2020.

[70] S. Zhang, F. Sun, N. Wang, C. Zhang, Q. Yu, M. Zhang, P. Babyn, and H. Zhong. Computer-aided diagnosis (cad) of pulmonary nodule of thoracic ct image using transfer learning. *Journal of digital imaging*, pages 1–13, 2019.

[71] S. Zheng, Z. Shen, C. Peia, W. Ding, H. Lin, J. Zheng, L. Pan, B. Zheng, and L. Huang. Interpretative computer-aided lung cancer diagnosis: from radiology analysis to malignancy evaluation. *arXiv preprint arXiv:2102.10919*, 2021.

[72] Y. Liu, P. Hao, P. Zhang, X. Xu, J. Wu, and W. Chen. Dense convolutional binary-tree networks for lung nodule classification. *IEEE Access*, 6:49080–49088, 2018.

[73] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.

[74] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

[75] M.-A. Carbonneau, E. Granger, A. J. Raymond, and G. Gagnon. Robust multiple-instance learning ensembles using random subspace instance selection. *Pattern recognition*, 58:83–99, 2016.

[76] Y. Xiao, B. Liu, and Z. Hao. A sphere-description-based approach for multiple-instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):242–257, 2017.

[77] Z.-H. Zhou. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2, 2004.

[78] J. Yang. Review of multi-instance learning and its applications. *Technical report, School of Computer Science Carnegie Mellon University*, 2005.

[79] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems*, pages 1073–1080, 2001.

[80] W. T. Vetterling, W. H. Press, S. A. Teukolsky, and B. P. Flannery. *Numerical recipes: example book C (The Art of Scientific Computing)*. Press Syndicate of the University of Cambridge, 1992.

[81] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009.

[82] M. Neuhaus and H. Bunke. A quadratic programming approach to the graph edit distance problem. In *International Workshop on Graph-Based Representations in Pattern Recognition*, pages 92–102. Springer, 2007.

[83] T. Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1):49–58, 2003.

[84] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable multi-instance learning. In *2014 IEEE International Conference on Data Mining*, pages 1037–1042. IEEE, 2014.

[85] J. Sánchez, . F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[86] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable algorithms for multi-instance learning. *IEEE transactions on neural networks and learning systems*, 28(4):975–987, 2016.

[87] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[88] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[89] S. K. Dilger, J. Uthoff, A. Judisch, E. Hammond, S. L. Mott, B. J. Smith, J. D. Newell, E. A. Hoffman, and J. C. Sieren. Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. *Journal of Medical Imaging*, 2(4):041004, 2015.

[90] J. Uthoff, M. J. Stephens, J. D. Newell Jr, E. A. Hoffman, J. Larson, et al. Machine learning approach for distinguishing malignant and benign lung nodules utilizing standardized perinodular parenchymal features from ct. *Medical physics*, 46(7):3207–3216, 2019.

[91] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[92] W. Safta and H. Frigui. Multiple instance learning for benign vs. malignant classification of lung nodules in ct scans. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 490–494. IEEE, 2018.

[93] W. Safta, M. M. Farhangi, B. Veasey, A. Amini, and H. Frigui. Multiple instance learning for malignant vs. benign classification of lung nodules in thoracic screening ct data. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1220–1224. IEEE, 2019.

[94] I. H. Sarker. Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. *SN Computer Science*, 2(3):1–16, 2021.

[95] M. Sarkar and A. De Bruyn. Lstm response models for direct marketing analytics: Replacing feature engineering with deep learning. *Journal of Interactive Marketing*, 53:80–95, 2021.

[96] J. Lee, N. R. Brooks, F. Tajwar, M. Burke, S. Ermon, D. B. Lobell, D. Biswas, and S. P. Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17), 2021.

[97] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[98] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[99] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015.

[100] S. McKinley and M. Levine. Cubic spline interpolation. *College of the Redwoods*, 45(1):1049–1060, 1998.

[101] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[102] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[103] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1):41–51, 2018.

[104] T. H. Dou, T. P. Coroller, J. J. van Griethuysen, R. H. Mak, and H. J. Aerts. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. *PloS one*, 13(11):e0206108, 2018.

[105] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013.

[106] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 2019.

[107] G. Zhang, Z. Yang, L. Gong, S. Jiang, and L. Wang. Classification of benign and malignant lung nodules from ct images based on hybrid features. *Physics in Medicine & Biology*, 64(12):125011, 2019.

[108] A. Pezeshk, S. Hamidian, N. Petrick, and B. Sahiner. 3-d convolutional neural networks for automatic detection of pulmonary nodules in chest ct. *IEEE journal of biomedical and health informatics*, 23(5):2080–2090, 2018.

CURRICULUM VITAE

Wiem Safta

Multimedia and Research Laboratory

Department of Computer Science and Engineering

University of Louisville, Louisville, KY, USA

E-mail: *wiem.safta@louisville.edu*,

safta.wiem@gmail.com

Tel: (502) 644-8780

Education

2015-2021 Ph.D. Student, Department of Computer Science and Engineering, University of Louisville, Louisville, KY 40292, USA.

Ph.D. candidate since Spring 2018, GPA = $3.84.0/4.0$.

2014 B.E., Telecommunications engineer, Higher School of Communication of Tunis, Tunis, Tunisia.

Work Experience

05/2021 – 08/2021	Data science Intern, Bristol Myers Squibb, NJ. - Development of prognostic models for overall survival in glioblastoma randomized control trials.
01/2015 - 12/2021	Graduate Teacher Assistant, Department of Computer Science and Engineering, University of Louisville, Louisville, Ky. - Research project: Results increased for lung nodules diagnosis task over existent methods by 5 % by developing traditional and deep learning techniques and integrating them in new approaches with Multiple Instance Learning algorithms. - Research project: Results improvement by integrating Breath data with traditional Lung cancer CT images screening and estimation of which compounds in each person's breath can indicate cancer presence using machine learning technics. - Research project: Integration of Multiple instance learning (MIL) algorithms in image segmentation as a new technique to improve results.
02/2014 - 07/2014	Intern, Research Institute of Chemistry Of Paris (IRCP): Paris, France. - Design new algorithms that resulted in the acceleration of electron paramagnetic resonance images (IEPR). - Integrate Chebfun toolbox developed at Oxford University.
05/2013 - 09/2013	Intern, Higher School of Communication of Tunis (SUP'COM): Tunis, Tunisia. - Establish transmission channel of "Rice type" for Radio Over-Fiber system at 60 GHz. - Estimate range of cells in different environments (indoor or outdoor) when planning RoF network.

07/2012 - 08/2012 Intern, Tunisie Telecom Company: Monastir, Tunisia.

- Openness to the workplace.

- Acquisition of new knowledge about the structure and inner workings of the business.

- Development of knowledge on switching transmission techniques and routing of telephone calls.

Certifications and Awards

- ***Grace Hopper celebration scholarship, 2018.***
- ***ISBI 2018 - Lung Nodule Malignancy Prediction Challenge, Rank 12 over 100 participants, 2018.***
- ***Graduate teaching academy certificate, 2019.***
- ***Tri-State Celebration of Women in Computing (TRI-WIC) scholarship, 2019.***
- ***Travel award for the IEEE International Symposium on Biomedical Imaging conference, 2019.***
- ***AAAI conference on Artificial intelligence scholarship, 2020.***
- ***Neural Networks and Deep Learning, Coursera certificate, 2020.***
- ***SQL for Data Science, Coursera certificate, 2020.***
- ***CECS Arthur M. Riehl Award university of louisville, 2020.***
- ***AI for medical Prognosis, Coursera certificate, 2021.***

Publications

1. **W. Safta** and H. Frigui "Multiple Instance Learning for Benign vs. Malignant Classification of Lung Nodules in CT Scans" *Medicine, Computer Science 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),*

2. **W. Safta**, M. M. Farhangi, Benjamin Veasey, A. Amini, and H. Frigui , "Multiple Instance Learning for Malignant vs. Benign Classification of Lung Nodules in Thoracic Screening Ct Data," *Computer Science 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).*
3. Y. Balagurunathan, Andrew Beers, M. McNitt-Gray, L. Hadjiiski, S. Napel, D. Goldgof, G. Pérez, P. Arbeláez, Alireza Mehrtash, T. Kapur, Ehwa Yang, J. Moon, G. Bernardino, R. Delgado-Gonzalo, M. Mehdi Farhangi, A. Amini, Renkun Ni, Xue Feng, Aditya Bagari, Kiran Vaidhya, Benjamin Veasey, **W. Safta**, H. Frigui, Joseph Enguehard, A. Gholipour, L. S. Castillo, L. Daza, P. Pinsky, J. Kalpathy-Cramer, K. Farahani, "Lung Nodule Malignancy Prediction in Sequential CT Scans: Summary of ISBI 2018 Challenge," *Medicine IEEE transactions on medical imaging.*

UNIVERSITY SERVICES

August 2016-April 2017 Representative of the Computer engineering and Computer Science department for the Graduate Student Council at the University of Louisville.