第 48 卷 第 12 期 湖南大学学报（自然科学版） Vol. 48 No. 12
2021 年 12 月 Journal of Hunan University (Natural Sciences) December 2021

Open Access Article

Medical Image Concept Detection Using Full Scale VGG-like Shallow and Transfer Learning Networks

Farhat Ullah Khan*, Izzatdin Aziz, Nordin Zakaria

Computer and Information Sciences Department, University Teknologi Petronas, Seri Iskander, Malaysia

Abstract: Over the last two decades, medical imaging examinations, and technologies together have been exponentially increased. With the increased demand for medical examinations, the demand for medical imaging experts is also increased. Manual identification and annotation of biomedical concepts tend to be rigorous and error-prone due to the varied knowledge of imaging experts. There is a critical need for automated Medical Concept Detection methods. Finding the relevant biomedical concepts present in a medical image holds the key to solve many automated clinical diagnosis problems, a machine learning pipeline for medical information retrieval, and other related issues, like creating and managing legacy or cloud-based descriptive digital repository. Appropriate mapping from biomedical image concepts into precise textual summary highly depends on the efficiency of Medical Concept Detection techniques. A novel clustering technique is presented as a complementary data preconditioning step to reach high concept detection results. The authors grouped 8767 Concept unique Identifiers (CUIs) into 970 clusters (label size decreased by 26% approximately using 97.7% images from the dataset). The main objective of this research is to examine the state-of-the-art convolution-based deep learning pre-trained and full-scale training models for the task of multi-label classification of medical concepts using medical image input. The research work evaluates the performance of transfer learning networks: InceptionV3, Xception, Dense Convolution Network (DenseNet) 121, VGG-16, and MobileNet. This work also presents one full-scale learning CNN architecture for the identification of relevant biomedical concepts that exist in medical images. Transfer learning technique using Xception model has achieved the highest F1 score of 36.29. The shallow VGG-like full-scale training architecture also has shown a promising result with an F1 score of 20.018. The obtained results reflect the significant improvement from previous experiments, offering state-of-the-art performance, with new data preconditioning precedence for highly variable and complex datasets.

Keywords: concept detection, concept annotation, deep learning, medical image processing, neural networks, machine learning.

使用全尺寸 VGG 类浅层和迁移学习网络进行医学图像概念检测

摘要：在过去的二十年里，医学影像检查和技术一起呈指数级增长。随着医学检查需求的增加，对医学影像专家的需求也随之增加。由于成像专家的知识多种多样，对生物医学概念的手动识别和注释往往是严格且容易出错的。迫切需要自动化的医学概念检测方法。找到医学图像中存在的相关生物医学概念是解决许多自动化临床诊断问题、用于医学信息检索的机器学习管道以及其他相关问题的关键，例如创建和管理遗留或基于云的描述性数字存储库。从生物医学图像概念到精确的文本摘要的适当映射高度依赖于医学概念检测技术的效率。提出了一种新颖的聚类技术作为补充数据预处理步骤，以达到高概念检测结果。作者将 8767 个概念唯一标识符 (CUI) 分组为 970 个集群（使用数据集中 97.7% 的图像，标签大小减少了大约 26%）。本研究的主要目的是检查最先进的基于卷积的深度学习预训练和全面训练模型，用于使用医学图像输入进行医学概念的多标签分类任务。研究工作评估了迁移学习网络的性能：盗梦空间五 3、异常、密集卷积网络 (致密) 121、VGG-16 和移动网络。

Received: August 17, 2021 / Revised: October 16, 2021 / Accepted: November 6, 2021 / Published: December 30, 2021

About the authors: Farhat Ullah Khan, Izzatdin Aziz, Nordin Zakaria, Computer and Information Sciences Department, University Teknologi Petronas, Seri Iskander, Malaysia

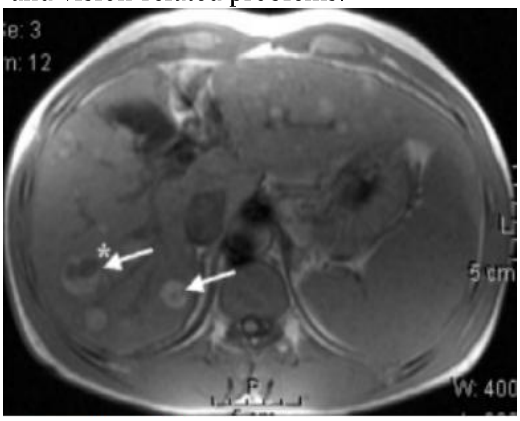
Corresponding author Farhat Ullah Khan, farhat_17000870@utp.edu.my

这项工作还提出了一种全面的学习美国有线电视新闻网架构，用于识别医学图像中存在的相关生物医学概念。使用异常模型的迁移学习技术取得了 36.29 的最高 F1 分数。类似 VGG 的浅层全尺寸训练架构也显示出令人鼓舞的结果，F1 得分为 20.018。获得的结果反映了先前实验的显着改进，提供了最先进的性能，并为高度可变和复杂的数据集提供了新的数据预处理优先级。

关键词：概念检测、概念注释、深度学习、医学图像处理、神经网络、机器学习.

1. Introduction

Analyzing medical images for extraction of semantic concepts and interpreting the unique medical image content in a natural language is a special long unsolved problem of Image annotation. Systems to automatically find the relevant, meaningful concepts from meagerly an image input (Figure 1) may help implement machine learning pipelines to solve many image and vision-related problems.



Concepts (CUIs):
– C0016911: Gadolinium
– C0021485: Injection or therapeutic agent
– C0024485: Magnetic Resonance Imaging
– C0577559: Mass of body structure
– C1533685: Injection Procedure

Fig. 1 A medical image with relevant concept unique identifiers (CUIs)

For example, detecting semantic concepts present in the medical image and further combining this comprehension to generate a descriptive summary in a natural language serve as a basis to solve the problem of automatic clinical diagnosis. This will be very useful since diagnosing is a prolonging task even for highly skilled professionals. In another perspective, designing a robust computerized concept detection framework and implementing it on an appropriated legacy or cloud-based environment may extend and overhaul the functions to create, host, and manage modality-based descriptive digital repositories. This research proposes a novel deep learning-based concept detection framework based on the above premises. The proposed concept detection framework also utilizes a novel clustering technique [1] as a complementary data preprocessing, discussed in detail in section 3.2 (Data Cleansing). The key contributions can be summarized as follows:

- This research proposes a *convolutional neural network*-based concept detection framework for mapping biomedical image concepts.
- This research also utilizes the novel clustering technique as a preconditioning step to enhance the classification results.
- Finally, this research presents an experimental comparison between deep learning *transfer learning* and *full-scale* training on medical image concept detection tasks.

The rest of the paper is organized as follows: Section 2 describes the relevant research. Section 3 discusses the comprehensive data and experimental preparations. Section 4 explains the proposed concept detection method, which covers the methodology for concept detection using transfer learning and full-scale training from scratch. Finally, in section 5, results are discussed. Section 6, in the end, draws the conclusion and recommends some future scope of the proposed research.

2. Literature Review

Since the evolution of deep learning methods and their huge success over image data, automatic medical image concept detection, and annotation were studied intensively. Various campaigns and challenges [2]-[6] were also organized to attract researchers worldwide to solve the related problems. The automatic Medical Concept Detection problem was poised to provide medical image interpretation by extracting medical semantics. Once the semantic concept vocabulary is detected during the concept detection step, other participating systems can operate together for different system-specific needs. Medical Concept Detection may serve as a preliminary step for all those systems. Several approaches were used for the medical image concept detection task, covering traditional retrieval systems and modern deep learning techniques. The discussion will be based on research groups that implemented their models using various deep learning techniques. The research publications [7]-[15] included RNNs, deep CNNs, and GANs to represent visual information.

AUEB NLP Group [16] has secured the top 1st, 2nd, 3rd, and 5th positions for their four different nominations with the best run F1 score of 28.2%, at ImageCLEFmed 2019 Concept Detection Task [2]. Their first ranked system is a recreation work of CheXNet [17] with an extension of handling larger label sets (5528 labels instead of 14) of the data. They used DenseNet-121 [18] for encoding images and added an FFNN to assign one or more of the 5528 output labels (classes) to each image using sigmoid activations to produce a probability per label. Their second-best system was a combination of the CheXNet-based and k-NN image retrieval-based techniques. In the k-NN retrieval phase, k most similar training images are retrieved with their gold concepts. Then cosine similarity between test and k-retrieved images are obtained to assign a score.

DAMO [19], ranked second top according to the best-run entries, has used residual network ResNet-101 for the multilabel classification approach. To handle the problem of data imbalance, they have applied several data filtering methods to get a balanced reduced dataset. They have achieved a 27% F1 score for their best entry. Their filtering methods emphasize the necessity of handling data imbalance problems to achieve higher accuracy.

ImageSem [7], [20], a second-time participant in the challenge, has designed the concept detection pipeline in two stages. In the first stage, which they called the pre-classification stage, they have divided images into four clusters according to the body parts and fine-tuned their multilabel classifier for the highest frequency concepts subset. They have achieved 22% of F1 score, secured the third position, and ranked the 8th best run. Their previous research for the same task had relied on heavy data preconditioning, and they have applied image retrieval and transfer learning. Their approach is based on the Lucene Image Retrieval Engine (LIRE) used in combination with Latent Dirichlet Allocation (LDA) for grouping concepts of similar images. A fine-tuned CNN, pre-trained with ImageNet weights, was also utilized to predict a selected subset of concepts by ImageSem.

UA.PT Bioinformatics [14], [21], which participated in 2018 ImageCLEF, has secured the fourth-best team position that year and was ranked 16th with their best F1-score of 21%. From their eight-run submissions, the best score resulted in SimpleNet configuration. In the past year's challenges, they have achieved remarkable results by using an adversarial auto-encoder and performing unsupervised feature learning. Experiments by UA.PT also included the BoVW (bag of visual words) algorithm, using OFAST and rotated BRIEF (ORB) keypoint descriptors. Two classification algorithms, namely a logistic regression and a k-nearest neighbor (k-NN) variant, have also been used for concept detection over the learned feature spaces. Using the adversarial auto-encoder technique features, they achieved the best results of a mean F1 score of 11% for a generalized linear classifier.

The CS MS group [13] and the AILAB used multimodal Recurrent Neural Networks (RNNs) as an encoder-decoder model. AILAB used a partial dataset with only 4000 images for visual feature extraction by a pre-trained CNN, and using word embedding, they obtained the text features. They used LSTM to merge two modalities and processed at dense layer to generate concept prediction. CS MS group [13] have encoded captions and are used as input to the RNN, whereas image features extracted from the deep network were encoded using a pre-trained CNN like AILAB. Combined encoded inputs were used to generate concepts finally.

NLM [22] has used Convolution Neural Networks (CNNs) and Binary Relevance using Decision Trees (BR-DT) for concept detection. PRNA [23] used an encoder-decoder-based framework that utilized an attention-based mechanism in CNN-based architecture to map the visual feature representation into relevant captions. BMET [24] group extended the NICv2 model [25], which consists of two varieties of neural networks combined to form an encoder-decoder for the image to language mapping.

AAI [26], MAMI [27], and MUPB [28] used a very deep neural network, but they were not effective as compared to shallower CNNs. Traditional bag-of-visual-words representations [29]-[30] or a mix of both have also been used in the challenge. However, deep convolution models are likely to deliver more robust results, whereas some of the best results are also based on the traditional features. Few researchers used retrieval-based mechanisms to identify highly visually related images on the ImageClef dataset [29]-[30]. Such related image captions are then searched for biomedical concepts assigned to the candidate image. This unique approach is proven to be good as it also has shown very promising results for the concept detection problem.

With four consecutive annual evaluation challenges, the trend is clear that the deep learning techniques will dominate sooner. Conclusively, where CNN-based models seem to deliver robust results on average, the traditional feature-based mechanisms were so far good. However, the deep learning methods may surpass traditional representations in terms of descriptive power. Other techniques are improving every year and also got satisfactory results, but still, even the highest score is far from the strong baseline to compare with.

Our research focuses on dealing with two related underlying problems. The deep neural network architectures are successful under the basic constraint that data should be largely homogeneous. These limitations restrict the efficient usage of complex deep neural architectures for data with greater complexity and variance. Hence the proposed novel clustering method helps reduce the complexity and variance of the task data. Second, this research exposes the limitations of such architectures, and eventually, a line of rigorous experiments is performed to test and recommend the suitable CNN architecture using transfer learning and full-scale training methods for data with such complexity and variance.

The research work is organized in a way that it should present a comprehensive comparative analysis from SOTA deep learning models and techniques. This paper also presents the customized shallow convolutional network implemented for medical image concept detection using a full-scale network training method. Section 2 explains the data characteristics and related preparations using the novel clustering method. Section 3 describes the methodology, while in section 4, experimental and comparison results are discussed. Finally, in section 5, the research conclusions and future direction of the work are presented.

3. Data and Experimental Setup

3.1. Data Overview

The dataset includes 222,305 training and 10,000 biomedical testing images. The images are released from scholarly articles in PubMed Central (PMC) (http://www.ncbi.nlm.nih.ov/pmc/). Each training image is provided with a set of UMLS CUIs for concept learning. A total of 111,156 unique UMLS concepts are extracted from the training set with the help of the QuickUMLS library [31]. Biomedical Concepts referred to a set of clinical concepts relevant to the medical image and provided by the US National Library of Medicine (NLM) and are called Concept Unique Identifiers (CUIs). In the Unified Medical Language System (UMLS) of the National Library of Medicine (NLM), the Concept Unique Identifier (CUI) is an 8-character identifier beginning with the letter C and followed by seven digits. Each medical concept is assigned such a CUI. An example medical image with its relevant concepts (CUIs) is shown in Figure 1. Similarly, different types of (multimodal) medical images may lie under the same concept label, as shown in Figure 2, making the classification very challenging, and data preprocessing for good homogeneous training input becomes evidently necessary.

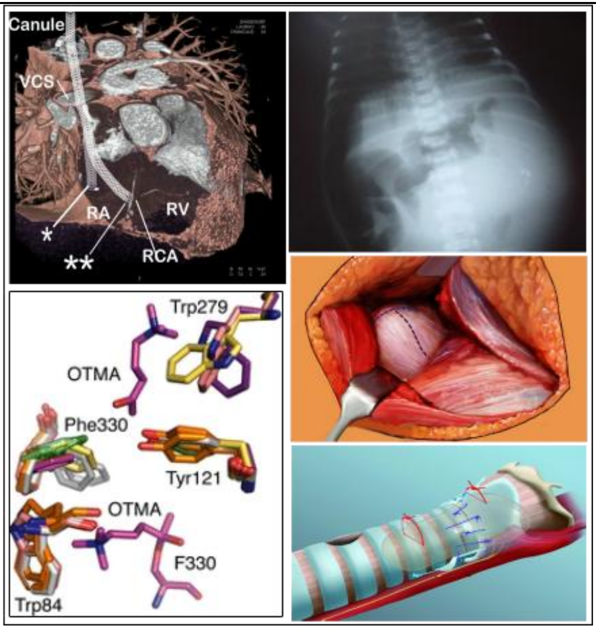


Fig. 2 Challenge: Multiple types of (multimodal) medical images under the concept (CUI) label, e.g., C0150305 ()

3.2. Data Cleansing

Data analysis was performed to understand and get better insight from the data [20]. Data preconditioning was helpful and necessary for this kind of high data diversity to achieve high performance from the concept detection models. Additional data processing was performed for better understanding of the data nature as a complementary step before inputting our model. We also considered the data analysis performed by ImageSem [20]. ImageSem analyzed the annotated concept frequency distribution for multilabel training object selection and similar image measurement. Their data analysis reflected that the most frequently used concepts are very few in numbers, including the redundant ones.

Table 1 Frequency distribution of medical concept labels (CUIs) in medical image training data

Frequency	Number	Ratio
0-100	102480	92.19%
100+	8676	7.81%
Total	111156	100%

Concepts with a frequency less than 100 are very high (92.19%). This fact shows the uneven distribution of CUIs over training data. To increase the concept coverage, we performed some additional processing. We considered the concepts in the list, which have a frequency of more than 1000 (7.81% CUIs) as suggested by ImageSem [20]. Further, we modified the *similarity score* calculation method (equation 1) and increased the concept coverage using the modified similarity score renamed *membership _score* (equation 2). ImageSem [20] selected the grouping candidate CUIs from only the concepts having a frequency of more than 1000. We did the same in our experiments and found that 1312 CUIs have above 1000 frequencies. These 1312 CUIs are selected as the label representing CUIs for multilabel classification. ImageSem [20] also reflected one important insight: the co-occurrence of CUIs in the same cluster of images. They devised a formula for clustering the images based on co-occurrence of CUI, which they called similarity score.

$$SIMILAR\_SCORE(A,B) = \frac{images\_A \cap images\_B}{images\_A \cup images\_B} \quad (1)$$

The groups with a similarity score above 80% are combined in one group. As a result, 1312 CUIs with co-occurrence were reduced into 459 class representations containing 208595 medical images.

During the statistical analysis of the data, it was observed that many CUI labels have a large number of similar images. However, the similarity score is low due to the big difference in numbers. For example, CUI 'C0000726' has 3783 images, whereas the matching CUI 'C0153662' has 1566 images, out of which 1564 images match the CUI 'C0000726', but those groups were ignored as their similarity score was below 80% due to the formula by ImageSem. Table 3 presents more similar matching statistics. It can be observed from table 3 that due to high variation in image count causes the poor similarity score, eventually disallowing many good candidate CUIs to be a part of the cluster representation group. To increase and improve the coverage of the concepts and make richer CUIs group representation to be part of the training, the similarity score was reformulated to reduce the chances of ignoring good candidate CUIs. The reformulation was given a new name called Membership score. The reformulated equation is as follows:

$$M\_SCORE(A,B) = \frac{images\_A \cap images\_B}{Min(images\_A | images\_B)} \quad (2)$$

where images_A and images_B are the images from the groups CUI-A and CUI-B, respectively. M_SCORE(A, B) calculates and obtains the degree of each group membership in addition to the similarity score so that even if the similarity score is below 80%, and the membership score is above 85%, the CUI will be selected to be combined to the bigger cluster. This work is already published as independent research [32]. Table 2 presents the S_Score and M_Score to the present reformulated score gap and impact.

Table 2 High number of matches but similarity score is below threshold (%)

CUI	Images (No)	Matching CUI	Images (No)	Match	S_Score(%)
C0000726	3783	C0153662	1566	1564	41.32
C0001613	1043	C0007776	1596	989	59.94
C0003893	1758	C0177601	2075	1662	76.55
C0004763	4753	C1704653	17527	4712	26.82

Now, for creating the cluster group, either of the conditions is checked, which means that if the similarity score is at par, the group will be considered merged to be a part of a larger cluster group. Alternatively, if M_Score is above 85%, the group will be merged in the representing cluster. The new data analysis has resulted in 217209 images, mapped with 970 cluster groups to be used finally as input to the model designed for multilabel classification (Table 3).

Table 3 A high number of matches but similarity score is not at par (%)

CUI	Images (No)	Matching CUI	Images (No)	Match	S_Score(%)	M_Score
C0000726	3783	C0153662	1566	1564	41.32	99.87%
C0001613	1043	C0007776	1596	989	59.94	94.82%
C0003893	1758	C0177601	2075	1662	76.55	94.54%
C0004763	4753	C1704653	17527	4712	26.82	99.14%

Table 4 Overall frequency dispersion of medical concept labels (CUIs) in medical image training data

CUI Code	Concept Name	No of Images
C1550557	Relationship Conjunction-and	77,003
C1706368	And - dosing instruction fragment	77,003
C1704254	Medical Image	20,165
…	...	...
C0202823	Chest CT	7,917
…	...	...
C0400569	Closed fracture of neck of femur	1

Table 4 shows that there are CUIs that have occurred in the highest number of images (C1550557 occurred in 77,003 images). There are chances that some of the concept groups may have many images in the set, and some may have very little. To standardize the concept distribution over training images and avoid over-fitting later, we will under-sample or over-sample the clusters after preprocessing step concerning the average number of images per cluster.

3.3. Experimental Setup

For the implementation of our deep learning model, we used Keras (version 2.1.5) library in a Python 3 (version 3.6.9) SciPy environment on top of the TensorFlow (version 1.12.0) backend. Keras facilitates a clean and convenient way to create various deep learning models TensorFlow stack, executed on GeForce GTX 1070 GPU given the underlying frameworks. Other necessary libraries, such as SciKit-Learn, OpenCV, Pandas, NumPy, Pickle, and Matplotlib, are also installed and used to support visual and text feature extraction for model inputs, generation, saving, and plotting. We planned our research work in two different ways to achieve higher chances of correct predictions. We have fine-tuned our training for transfer learning with DenseNet121, ResNet50, InceptionV3, Xception, and MobileNet CNN models. We have designed a VGG-like shallow convolutional neural network model for full training.

4. Concept Detection Method

Medical image concept detection is a special image classification problem where multimodal medical images are provided with their CUI (Concept Unique Identifiers) labels. Systems are being developed to identify complex image features to learn and further identify the concepts (labels) (Figure 3).

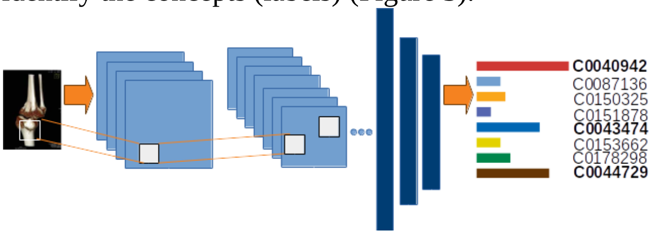


Fig. 3 Concept detection using deep learning

There was a great shift in the methods being used for image classification, and deep learning has taken a big leap in this. Our research focuses on the deep learning methods used for medical image conception detection tasks. The concept detection experiments were designed to use five representations from transfer learning and one full-scale training technique. One shallow full-scale training network method is also utilized (Figure 4).

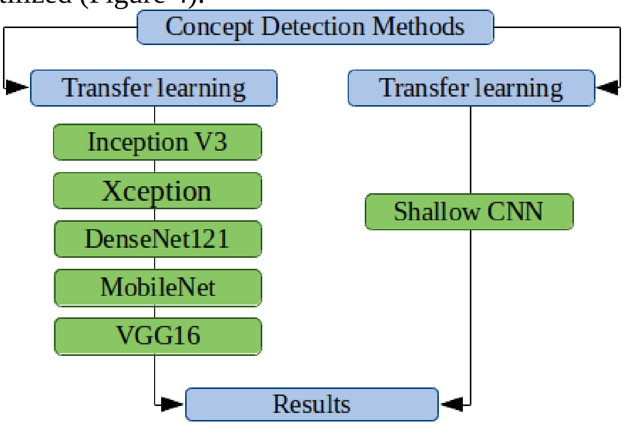


Fig. 4 Concept detection methods

4.1. Transfer Learning

Lately, deep neural structures, especially convolutional neural networks (CNNs), demonstrated excellent results and surpassed the human-level performance on object identification and image classification problems. Convolution networks can discriminate the complex visual indicators for image, vision, and object recognition tasks, achieving comparatively superior performance from the classical techniques by convolving with tens of convolution filters and training multiple depths of layers. Since deep neural networks such as Convolution Neural Networks (CNNs) were introduced for image-related solutions, their successful applications have shown wide applicability in other specialized domains like medical image concept detection and annotation. Moreover, using pre-trained networks even from a different application domain could be a better starting point. Pre-trained networks are already designed for a different task, but the layers and their training weights can be utilized as a starting point for some other new tasks. The medical image concept detection is treated as a multilabel classification task. In the multilabel classification method, Convolution Neural Network (CNN) is applied to assign one or more CUIs from the predefined CUIs label set. Here, we have used the ImageNet pre-trained model for transfer learning, and then our preprocessed biomedical image dataset is trained for domain adaptation.

This was a multilabel classification problem, and we limited our experiment to the top 20 frequent labels. Input images were resized according to the allowed minimum pixels for each CNN with no cropping. We used a randomly shuffled batch with the size of 32 and 0.0001 initial learning rate. The binary cross-entropy loss function is used with Adam optimizer with default beta values. Rescaling, zooming, rotation, shearing, and horizontal augmentation techniques are performed using Keras Image Data Generator. Data is split into 85% training set and 15% validation set.

Table 5 summarizes the hyper-parameters of the experiments, and Table 6 summarizes the utilized network characteristics. The performance result of each CNN is being discussed in the result section. Figure 5 presents the transfer learning network pre-trained with ImageNet weights.

Table 5 Hyper-parameters setting for the experiments

Hyper-Parameters	Value
Optimization technique	Adam
Initial learning rate	0.0001
Epochs	75
Batch size	32

Table 6 Summary of the utilized transfer learning networks

Model	Parameters (Millions)	Accuracy (%) on Imagenet	Depth	Input Size
Inception V3	23.9	78.2	48	299x299
Xception	22.8	79	72	299x299
DenseNet121	0.8	74.98	121	224x224
MobileNet	4.2	70.6	28	224x224
VGG16	134	71.5	16	224x224

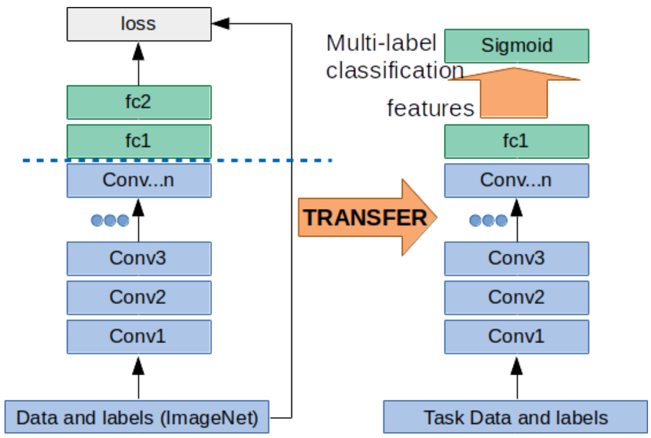


Fig. 5 Domain adaptation for concept detection

We have studied several transfer learning techniques [18], [33-38] proven to be very successful in ImageNet and other task datasets and deemed fit for medical image concept detection tasks.

4.2. Full Training – Proposed Shallow CNN for Training from Scratch

Context-specific descriptors were deemed unsuccessful because the training and validation dataset included a wide variety. There were images with large numbers covering radiology X-ray, tissue cell structures, and even the charts and graphs are in the same group (Figure 2).

Training and predicting a neural network with large enough depth seems to be a more suitable choice for managing this variety and complexity. However, considering the complexity of the data, it was realized that the performance of the state-of-the-art deep neural structures in this context is limited because the transfer learning networks are trained in a different context. The task data is used only for model adaptation. Hence, it has emerged as a need to testify the performance of a deep neural network fully trained on task data. In our model, we have proposed a VGG-like shallow [39] and a compact variant of VGGNet network architecture to perform training from scratch. VGGNet [36], developed and trained by Oxford renowned Visual Geometry Group (VGG), refers to a 16-layer deep CNN for object recognition, which achieved a position amongst the top performances on the ImageNet dataset. During the complimentary data preprocessing step (Figure 6), the standardized clusters are fed as training data to our custom convolution neural network.

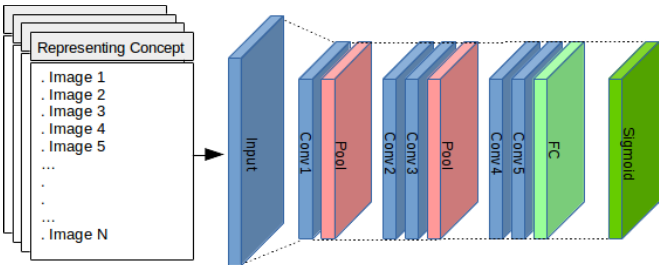


Fig. 6 Proposed Mini VGGNet like CNN for multi-label classification using full-scale training technique

The proposed custom CNN architecture characteristics could be summarized as follows (Figure 7):

An arrangement of 3x3 convolution layers stacked in increasing depth with a 25 % dropout rate and max-pooling is used to reduce volume. Dense layers are used at the end of the network, just before the sigmoid classifier. The activation function RELU is followed by batch normalization. Dropout is utilized by randomly deactivating neural units in respective layers. During training, this random disconnection process helps control overfitting in the model and introduces redundancy. Eventually, no single node in the layer is responsible for predicting a certain class, shape, or object. This combined arrangement of multiple *CONV+RELU* layers before pooling helps the model learn a rich abstraction of features that suits inputs like medical images.

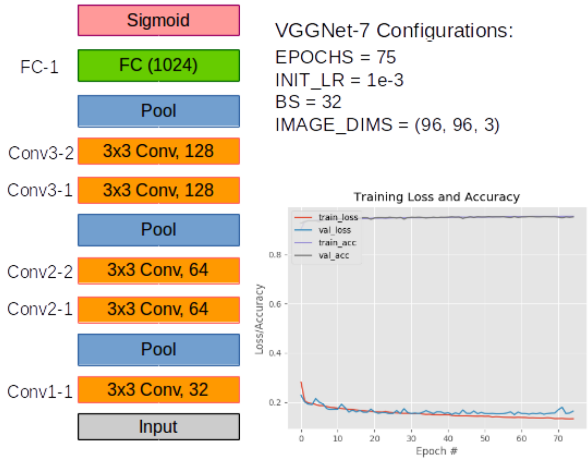


Fig. 7 Proposed convolution neural network layer stack

4.3. Cross-Entropy Loss

The Binary Cross-Entropy or log loss is used for loss minimization and backpropagation. The CE Loss is defined as:

$$CE = -\sum_{i}^{C} x_i log(Y_i) \tag{3}$$

where x_i is the ground truth and Y_i is the prediction score for every class I in class C. Activation function sigmoid is applied to the scores to compute the Cross-Entropy Loss; $f(Y_i)$ refers to the activations.

4.3.1. Binary Cross-Entropy Loss for Multilabel Classification

Binary Cross-Entropy Loss or Sigmoid Cross-Entropy loss is a Sigmoid activation with a Cross-Entropy loss. It differs from Softmax loss in that the calculated loss is not affected by the other neural network output vector component because each vector output component is treated independently.

By using the sigmoid activation function with binary cross-entropy loss function, each label will be treated as an individual binary label. The relatedness degree of each component belonging to a certain class will not affect the degree of belongingness of the other class. That is how it is suitable for multilabel classification. It functions as a binary classification problem among two classes and for each class in class C. It is called Binary Cross-Entropy Loss, as discussed above. The following formulation of Cross-Entropy Loss for binary problems is often used:

$$CE = -\sum_{i=1}^{C'=20} x_i log(f(Y_i))$$

$$= -\sum_{i=1}^{C'=20} [x_i log(f(Y_i)) + (1 - x_i) log(1 - f(Y_i)) \tag{4}$$

Setting up C disjoint binary classification problems (C=20 top frequent cluster labels.) and then adding up the loss over the different independent binary problems is the way to obtain gradients of every binary problem, which adds to backward propagation, and the losses to analyze the overall loss.

5. Results

5.1. Transfer Learning

The main idea behind using transfer learning is to take a model which is already successfully used and repurpose it for domain adaptation. We have used several successful CNN models for verification and validation purposes and fine-tuned them for our task dataset. We have also presented the activation maps in five intermediate layers of convolutional outputs for each transfer learning model used.

The model includes InceptionV3, Xception, VGG16, MobileNet, and DenseNet121 network architectures. The purpose of the experiment with these models is to investigate a detailed hyperparameter setting and its impact, particularly on the biomedical dataset used.

5.1.1. Inception V3

With 42 layers, introducing the idea of factorizing convolutions has helped reduce the number of parameters without compromising the network efficiency. The lower error rate is obtained to make InceptionV3 [35] the Runner Up in ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2015. When using InceptionV3 for transfer learning or domain adaptation of the medical image dataset, we obtained the best F1 score of 16.75 (Figure 8) with training accuracy of 92.98, validation accuracy of 92.92 training loss 20.21, and validation loss of 20.80 in max pooling.

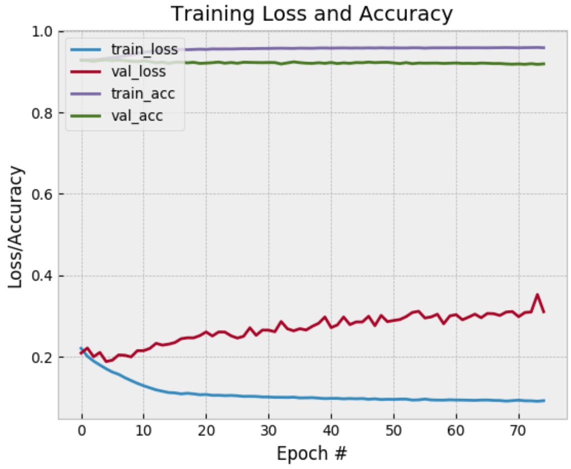


Fig. 8 (a) Performance graph of InceptionV3 for concept detection with average pooling

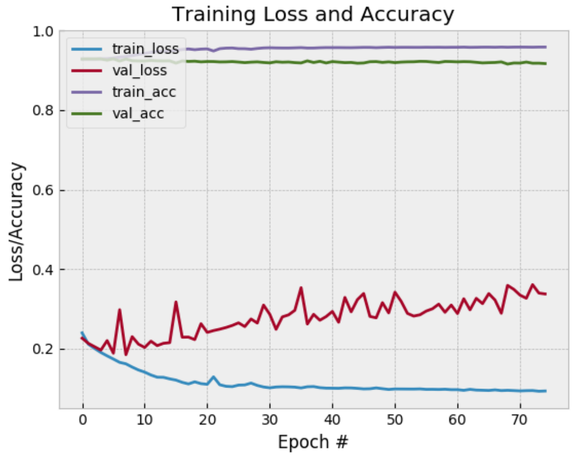


Fig. 8 (b) Performance graph of InceptionV3 for concept detection with max pooling

While, with the average pooling method, the F1 score is obtained at 18.96 with a training accuracy of 93.79, validation accuracy of 92.74, training loss 15.55, and validation loss of 20.19. ImageNet weights were used for transfer learning. Figure 9 presents the qualitative result of the same.

Fig. 9 Feature maps for InceptionV3 learning

5.1.2. Xception

Xception [37] by Google, stands for Extreme version of Inception, it is even better than InceptionV3 (also by Google, 1st Runner Up in ILSVRC 2015) for both ImageNet ILSVRC and JFT datasets. The power comes in Xception because there is no intermediate activation. Eventually, the highest accuracy was obtained compared to those models which used ReLU or ELU. When it was tested for biomedical image dataset and used for transfer learning or domain adaptation of medical image dataset, has obtained the best F1 score 35.21 (Figure 10) with training accuracy of 95.87, validation accuracy of 91.72, training loss of 09.29, and validation loss of 33.37 in max pooling. While, with the average pooling method, the F1 score obtained is 36.26 with a training accuracy of 95.86, validation accuracy of 91.90, training loss of 09.21, and validation loss of 31.06. ImageNet weights were used for transfer learning.

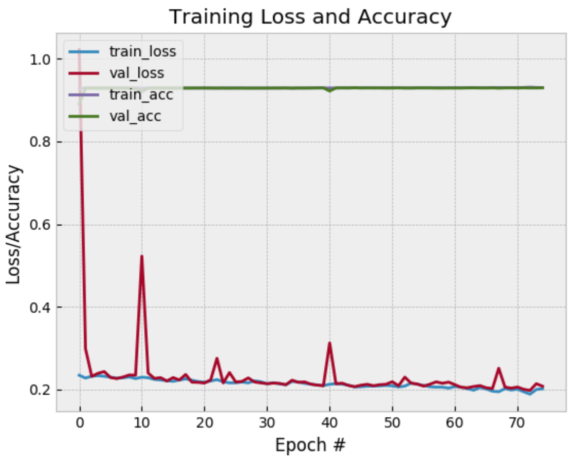


Fig. 10 Performance graph of Xception for concept detection with max pooling

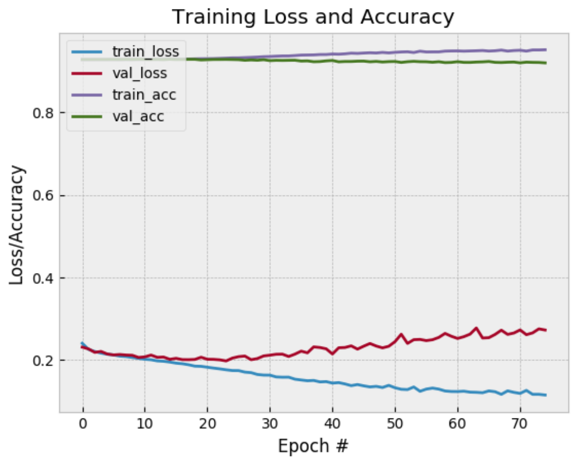


Fig. 12 (b) Performance graph of VGG16 for concept detection with max pooling

Figure 11 presents the qualitative output from the Xception transfer learning.

Fig. 11 Feature maps for InceptionV3 learning

5.1.3. VGG 16

VGG-16 has 21 layers altogether, but only 16 layers are weight layers, and that is why it is named VGG-16. It was successful because it uses a small filter (3x3) in the first and second convolution layers instead of 11x11 or 5x5, enabling local features to be captured.

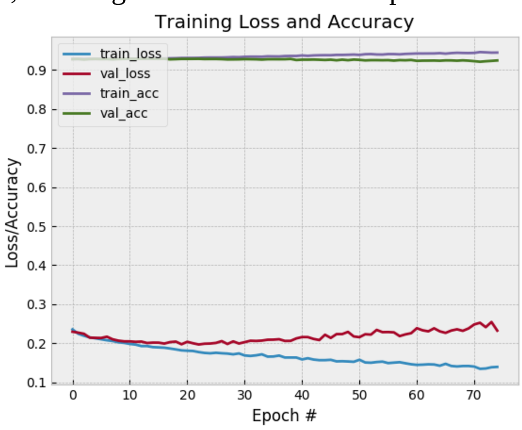


Fig. 12 (a) Performance graph of VGG16 for concept detection with average pooling

Small filter size also results in a few parameters to be learned, which is eventually good for faster convergence and reduced overfitting problems. When it was tested for biomedical image dataset and used for transfer learning or domain adaptation of medical image dataset, the best F1 score 21.28 was obtained (Figure 12) with training accuracy of 95.21, validation accuracy of 92.03, training loss of 11.50, and validation loss of 27.25 in average pooling. While, with the max-pooling method, the F1 score obtained is 19.33 with a training accuracy of 94.49, validation accuracy of 92.46, training loss of 13.91, and validation loss of 23.22. Imagenet weights were used for transfer learning.

5.1.4. DenseNet121

The CNNs with a deeper path like ResNet (100 to 1000 layers deep) have issues that information from the input layer and the output layer may vanish before reaching the other side (and gradient in the direction of output to input).

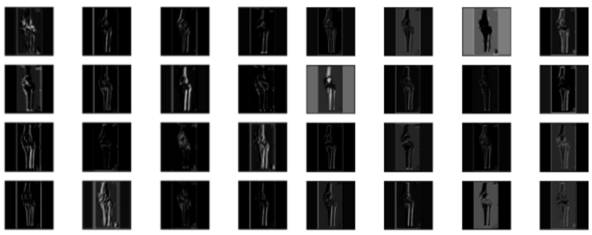


Fig. 13 Feature maps for VGG16 learning

DenseNet uses the network potential through feature reuse instead of exploiting the representation power of wider and deeper networks. Eventually, DenseNet requires fewer parameters and drops redundant feature maps.

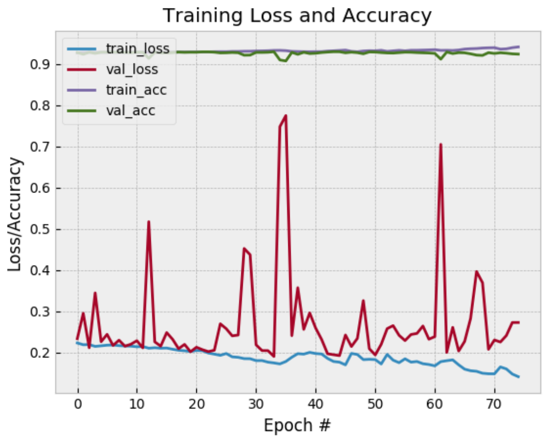


Fig. 14 (a) Performance graph of DenseNet121 for concept detection with average pooling

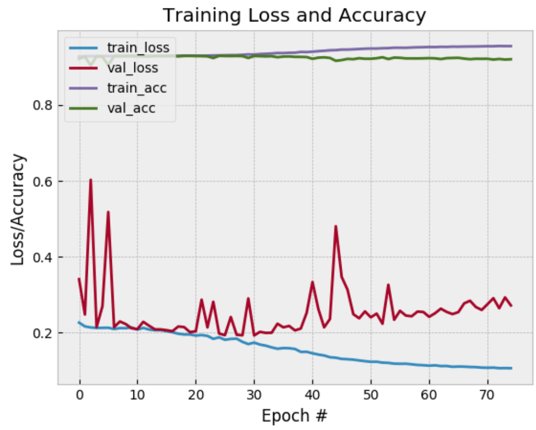


Fig. 14 (b) Performance graph of DenseNet121 for concept detection with max pooling

When it was tested for biomedical image dataset and used for transfer learning or domain adaptation of medical image dataset, the best F1 score 21.13 was obtained (Figure 14) with training accuracy of 94.14, validation accuracy of 92.39, training loss of 14.19, and validation loss of 27.32 in average pooling.

While, with the max-pooling method, the F1 score obtained is 23.76 with a training accuracy of 95.49, validation accuracy of 92.03, training loss of 10.66, and validation loss of 27.20. ImageNet weights were used for transfer learning.

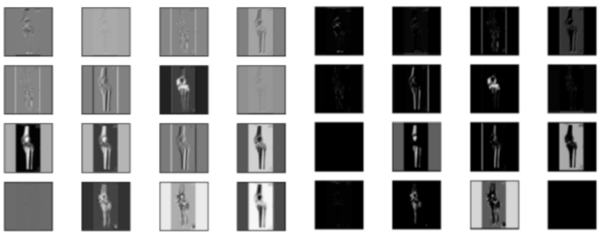


Fig. 15 Feature maps for DenseNet121 learning

5.1.5. MobileNet

In pursuit of finding the better network architecture for the special problem, we also looked for architecture out of the general trend of deeper and complicated CNN architectures. MobileNet is mainly introduced in view of solving embedded vision and mobile application problems. It has reduced the complexity and size of the model to make use of deep neural networks in mobile devices.

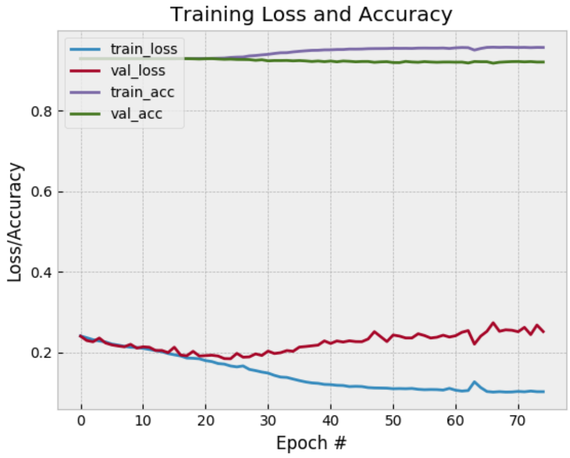


Fig. 16 (a) Performance graph of MobileNet for concept detection with average pooling

When it was tested for biomedical image dataset and used for transfer learning or domain adaptation of medical image dataset, the best F1 score 24.99 was obtained (Figure 16) with training accuracy of 95.67, validation accuracy of 92.25, training loss of 10.23, and validation loss of 25.71 in average pooling. While, with the max-pooling method, the F1 score obtained is 24.65 with a training accuracy of 95.64, validation accuracy of 921.98, training loss of 10.52, and validation loss of 26.89. ImageNet weights were used for transfer learning.

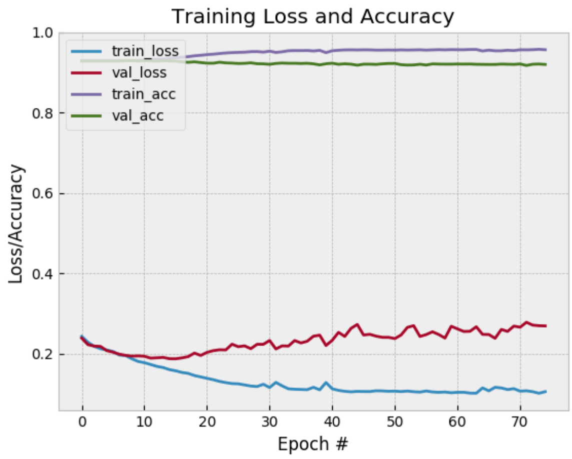


Fig. 16 (b) Performance graph of MobileNet for concept detection with max pooling

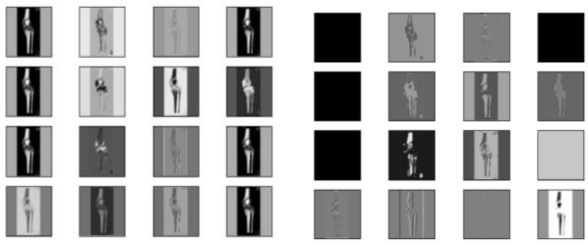


Fig. 17 Feature maps for MobileNet learning

5.2. Full Training - Proposed Shallow CNN for Training from Scratch

Several tests were performed with full-scale training of deep neural networks to present the impact of training from scratch on task data for the experimental support extension. Although training a deep neural network for full-scale training is not encouraged unless a large amount of labeled training data is available. In our case, our preprocessing methods have allowed performing the experiments because the training data was resampled for sufficient training samples.

When tested for target biomedical image dataset and used for full-scale training, the best F1 score of 20.018 was obtained (Figures 18 and 19) with training accuracy of 95.33, validation accuracy of 92.16, training loss of 11.20, and validation loss of 28.25.

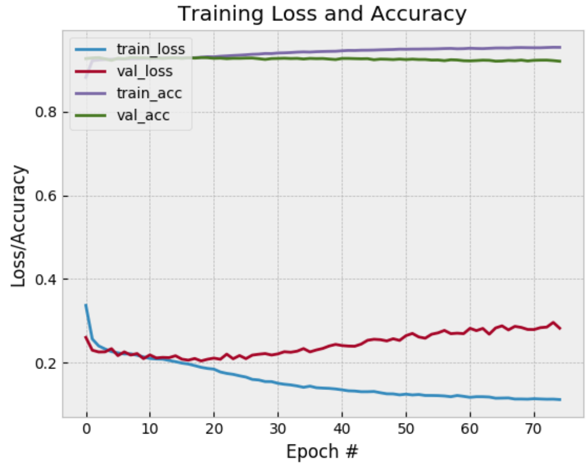


Fig. 18 Performance graph of full-scale training using shallow-CNN

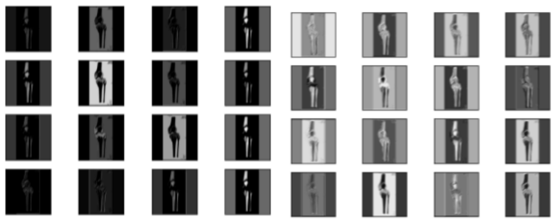


Fig. 19 Feature maps for full-scale shallow network learning

6. Discussion

To better understand the nature of data and achieve higher performances, various statistical tests are performed, and medical images are organized in clusters based on a proposed statistical arrangement called *Membership_score* of the concept groups. These clusters are labeled by the representing concept. The images in the representation concept group are given as input to the multilabel classification model.

Deep learning Networks, particularly convolution networks (CNNs) for generic image and object classification, is a recent success in image and vision technologies. Our proposed model intends to build a simple yet efficient concept detection design using CNN, which works on concept selection strategy using transfer learning or full-scale training techniques. Several traditional CNN transfer learning methods and techniques from a different set of deep learning network groups were tested. Full-scale training using Shallower VGGNet convolutional neural network variant is also used for multilabel classification of concepts.

This research aims to assess the transfer learning and full-scale training models for the multilabel classification of medical concepts (CUIs) using medical image input. The performance of both techniques is significantly improved due to the rigorous data preprocessing. Data preprocessing seems to be a crucial and necessary step as the medical image dataset was immensely diverse. The dataset was divided into 75% training and 25% testing. Therefore, five pre-trained (on ImageNet) CNN models, namely InceptionV3, Xception, DenseNet-121, VGG16, and MobileNet, were trained by the transfer learning technique. In addition, a full-scale custom VGG-like shallow network also was developed and trained to understand the performance under the concept of training from scratch (Table 7).

Table 7 Comparison of top 6 previous results with the proposed research result

Previous Results (F1 Score)		**Our Results (F1 Score)**	
Name	**Score**	**Method Name**	**Score**
DAMO [19]	26.55	Xception	36.26
ImageSem [7, 20]	22.35	MobileNet	24.99
UAPT [14,21]	20.58	DenseNet121	23.76
Richard	19.52	VGG16	21.28
Sam	17.49	Inception V3	20.80
MS-CSIRO	14.35	MiniVGG	20.01

These training models were evaluated by accuracy, loss, and F-score metrics using the same hyper-parameter settings. Table 7, complemented with a performance graph (Figure 20), presents a general performance. Xception under the concept of transfer learning technique achieved the highest result than other CNN architectures. Figure 21 maps the table 7 performance and presents a visual comparison of baselines considered with our current results.

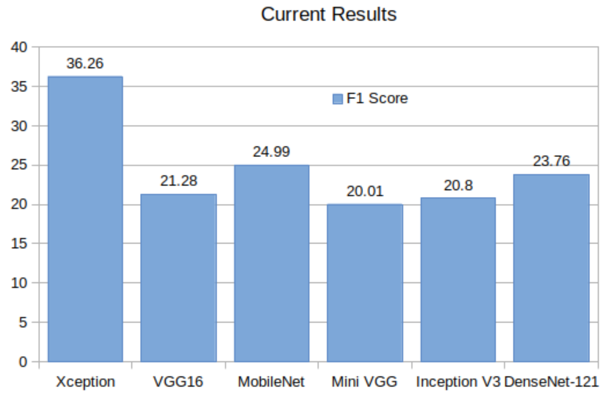


Fig. 20 Performance graph for CNN methods

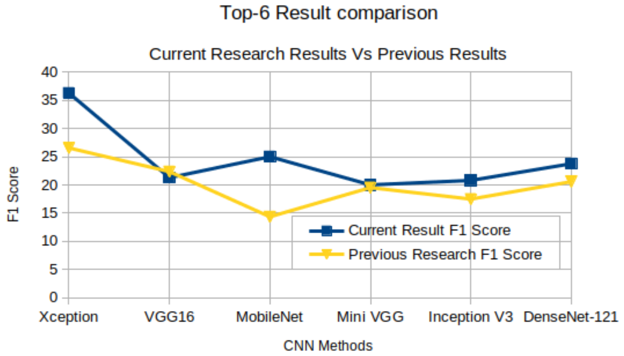


Fig. 21 Performance comparison of the proposed research with top SIX CNN-based state-of-the-art results

Figure 10 shows the training loss and accuracy in max and average pooling conditions for Xception, achieving a 36.06 best F1 scores with training accuracy of 95.86, validation accuracy of 91.90, training loss 09.21, and validation loss of 31.06. Figures 22 and 23 represent the multi-labeled ROC curve and ROC curve for each label, respectively.

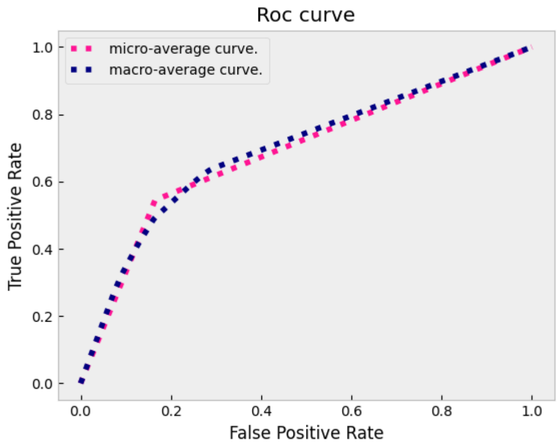


Fig. 22 Multilabel ROC curve

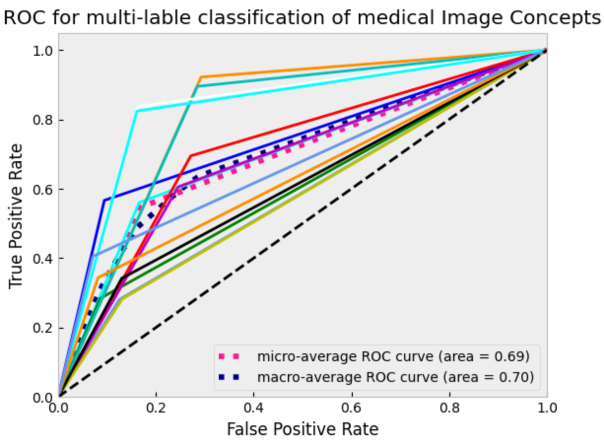


Fig. 23 ROC curve for each label

The results implicate that the success of the models is improved but limited. Still, they are not at a satisfactory level to be claimed as a benchmark performance for the task. The recent release of the ROCO multimodal image dataset [40] has taken care of the wide variety and reduced it to only Radiology containing over 81k radiology images. All compound and non-radiology images are now removed. Data quality is improved and homogenized to a certain extent but still has complexity and variance even within limited classes. There is still a need for developing methods to amalgamate techniques further to learn the nature of the data at first and flexible deep neural networks to automatically adjust the learning network and settings according to the complexity of the data. New dynamic deep learning methods are developed that learn automatically from the complexity of the data while training incrementally [41]-[42]. However, their application to concept detection is still a concern.

7. Conclusion and Future Scope

This research thoroughly investigated the end-to-end reason and impact-based potential CNN models both in full-scale training and fine-tuning methods. Our analysis is based on the dataset prepared during the preprocessing step, which helped to make a homogeneous dataset. We kept our dataset size identical for both methods, and results clearly indicate that transfer learning is always at the lead. We have compared the performance of CNNs with different depths, and Xception turns out to be the best for preconditioned medical image data.

Several other research findings were observed. The data is highly diverse and complex in nature, and our research establishes a premise that data has to be preconditioned to be homogenized for higher training performance and accuracy. We also observed that the full-scale training or transfer learning methods have limitations in learning the complex and diverse variety of data. We need to look beyond the classical methods of deep neural networks. We need to seek an end-to-end amalgamation of methods for observing the complexity of data to adapt and learn complex features and flexible neural architectures to fit accordingly better.

Acknowledgment

We would like to acknowledge the tremendous support from the Centre for Research in Data Science (CerDaS) and the Department of Computer and Information Sciences, UTP, Malaysia.

References

[1] KHAN FU & AZIZ IB. Reducing high variability in medical image collection by a novel cluster-based synthetic oversampling technique. *Proceedings of the 2019 IEEE Conference on Big Data and Analytics (ICBDA),* 2019: 45–50. doi:10.1109/ICBDA47563.2019.8987171.

[2] PELKA O, FRIEDRICH C, GARCÍA SECO DE HERRERA A, & MÜLLER H. *Overview of the ImageCLEFmed 2019 concept detection task*. CLEF working notes, CEUR 2019.

[3] IONESCU B, MÜLLER H, VILLEGAS M, et al. Overview of ImageCLEF 2017: Information extraction from images. *Proceedings of the International Conference of the*

Cross-Language Evaluation Forum for European Languages. Springer, 2017: 315–337.
[3] DE HERRERA AGS, EICKHOFF C, ANDREARCZYK V, & MÜLLER H. *Overview of the ImageCLEF 2018 Caption Prediction Tasks.* CLEF (Working Notes), 2018.
[5] EICKHOFF C, SCHWALL I, DE HERRERA AGS, MÜLLER H. *Overview of ImageCLEFcaption – Image Caption Prediction and Concept Detection for Biomedical Images*, 2017.
CLEF (Working Notes), 2017.
[6] PALOTTI JR, ZUCCON G, GOEURIOT L, et al. *CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information About Medical Symptoms*. CLEF (Working Notes), 2015: 1–22.
[7] GUO Z, WANG X, ZHANG Y, & LI J. ImageSem at ImageCLEFmed caption 2019 task: a two-stage medical concept detection strategy. *CLEF2019 Working Notes. CEUR Workshop Proceedings*, (2019: 1613–0073.
[8] SINHA P, PURKAYASTHA S, & GICHOYA J. Full training versus fine tuning for radiology images concept detection task for the ImageClef 2019 challenge. *CLEF2019 Working Notes. CEUR Workshop Proceedings*, 2019: 1613–0073.
[9] LIANG S, LI X, ZHU Y, et al. *ISIA at the ImageCLEF 2017 Image Caption Task.* CLEF (Working Notes), 2017.
[10] LITJENS G, KOOI T, BEJNORDI B.E, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017, 42: 60–88.
[11] KOHLI M, PREVEDELLO LM, FILICE RW, & GEIS JR. Implementing machine learning in radiology practice and research. *American Journal of Roentgenology* 2017, 208: 754–760.
[12] WANG X, GUO Z, ZHANG Y, LI J. Medical Image Labelling and Semantic Understanding for Clinical Applications. *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2019: 260–270.
[13] RAHMAN M. *A Cross Modal Deep Learning Based Approach for Caption Prediction and Concept Detection by CS Morgan State*. CLEF (Working Notes), 2018.
[14] PINHO E, & COSTA C. *Feature Learning with Adversarial Networks for Concept Detection in Medical Images:* UA. PT Bioinformatics at ImageCLEF 2018. CLEF (Working Notes), 2018.
[15] DIMITRIS K, & ERGINA K. *Concept detection on medical images using Deep Residual Learning Network.* Working Notes CLEF 2017.
[16] KOUGIA V, PAVLOPOULOS J, & ANDROUTSOPOULOS I. AUEB NLP group at ImageCLEFmed Caption 2019. CLEF2019 Working Notes. *CEUR Workshop Proceedings*, 2019: 09-12.
[17] RAJPURKAR P, IRVIN J, ZHU K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv,* 2017, 1711.05225.
[18] HUANG G, LIU Z, VAN DER MAATEN L, & WEINBERGER KQ. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 4700–4708.
[19] XU J, LIU W, LIU C, et al. Concept detection based on multilabel classification and image captioning approach-damo at ImageClef 2019. CLEF2019 Working Notes. *CEUR Workshop Proceedings*, 2019: 1613–0073.
[20] ZHANG Y, WANG X, GUO Z, & LI J. *ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning*. CLEF CEUR Workshop, Avignon, France, 2018.
[21] GONALVES A, PINHO E, & COSTA C. Informative and intriguing visual features: Ua.pt bioinformatics in ImageClef caption 2019. CLEF2019 Working Notes. *CEUR Workshop Proceedings*, 2019: 1613–0073.
[22] ABACHA AB, DE HERRERA AGS, GAYEN S, et al. *NLM at ImageCLEF 2017 Caption Task*. CLEF (Working Notes), 2017.
[23] HASAN SA, LING Y, LIU J, et al. *PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks*. CLEF (Working Notes), 2017.
[24] LYNDON D, KUMAR A, & KIM J. *Neural Captioning for the ImageCLEF 2017 Medical Image Challenges.* CLEF (Working Notes), 2017.
[25] VINYALS O, TOSHEV A, BENGIO S, & ERHAN D. Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 3156–3164.
[26] KATSIOS D, & KAVALLIERATOU E. *Concept Detection on Medical Images using Deep Residual Learning Network*. CLEF (Working Notes), 2017.
[27] NY HOAVY N, MOTHE J, & RANDRIANARIVONY M.I. IRIT & MISA at Image CLEF 2017-Multi label classification. *Proceedings of the International Conference of the CLEF Association*, CLEF 2017 Labs Working Notes (CLEF 2017), 2017 (Dublin, Ireland): 1-7.
[28] STEFAN LD, IONESCU B, & MÜLLER H. Generating captions for medical images with a deep learning multi-hypothesis approach: MedGIFT–UPB Participation in the ImageCLEF 2017 Caption Task. *CEUR Workshop Proceedings*, 2017, 1856: article 1523.
[29] JOLY A, GOËAU H, GLOTIN H, et al. Lifeclef 2017 lab overview: multimedia species identification challenges. *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2017: 255-274.
[30] VILLEGAS M, MÜLLER H, GILBERT A, et al. General overview of ImageCLEF at the CLEF 2015 labs. International conference of the cross-language evaluation forum for European languages. In: MOTHE J. et al. (eds) *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2015*. *Lecture Notes in Computer Science*, 9283: 444–461, https://doi.org/10.1007/978-3-319-24027-5_45
[31] SOLDAINI L, & GOHARIAN N. *Quickumls: a fast, unsupervised approach for medical concept extraction*. MedIR workshop, SIGIR, 2016. Available from http://medir2016.imag.fr/data/MEDIR_2016_paper_16.pdf
[32] KHAN FU, & AZIZ IB. Reducing high variability in medical image collection by a novel cluster based synthetic oversampling technique. *Proceedings of the IEEE Conference on Big Data and Analytics (ICBDA)*, 2019: 45-50.
[33] KUBANY A, ISHAY SB, OHAYON RS, et al. Comparison of state-of-the-art deep learning APIs for image multilabel classification using semantic metrics. *Expert Systems with Applications*, 2020, 161: 113656.
[34] MAEDA-GUTIÉRREZ V, GALVÁN-TEJADA CE, ZANELLA-CALZADA LA, et al. Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases. *Applied Sciences*, 2020, 10: 1245.

[35] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2818-2826.
[36] SIMONYAN K, & ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*: 2014, 1409.1556.
[37] CHOLLET F. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 1251–1258.
[38] HOWARD A.G, ZHU M, CHEN B, et al Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv preprint arXiv*: 2017, 1704.04861.
[39] ROSEBROCK, A. Multi-label classification with Keras. 2018. Available from https://www.pyimagesearch.com/2018/05/07/multi-label-classification-with-keras/2018 (accessed May 3, 2020).
[40] PELKA O, KOITKA S, RÜCKERT J, et al. Radiology objects in COntext (ROCO): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*; Springer, 2018; pp. 180–189.
[41] KHAN FU, AZIZ IB, AKHIR EAP. Pluggable Micronetwork for Layer Configuration Relay in a Dynamic Deep Neural Surface. *IEEE Access*, 2021, 9: 124831–124846. doi:10.1109/ACCESS.2021.3110709.
[42] KHAN FU, ABDUL AZIZ I, AHIR EAP. PrimeNet: Adaptive Multi-Layer Deep Neural Structure for Enhanced Feature Selection in Early Convolution Stage. *Preprints*, 2021, 2021070699 (doi: 10.20944/preprints202107.0699.v1).

参考文:

[1] KHAN FU & AZIZ IB. 通过一种新的基于集群的合成过采样技术减少医学图像收集的高可变性。2019 年 IEEE 大数据与分析会议 (ICBDA) 论文集，2019：45-50。 doi:10.1109/ICBDA47563.2019.8987171。
[2] PELKA O、FRIEDRICH C、GARCÍA SECO DE HERRERA A 和 MÜLLER H. 图像谱号医学 2019 概念检测任务概述。谱号 工作笔记，欧元 2019。
[3] IONESCU B、MÜLLER H、VILLEGAS M 等人。图像谱号 2017 概述：从图像中提取信息。欧洲语言跨语言评估论坛国际会议论文集。施普林格，2017：315–337。
[3] DE HERRERA AGS 、 EICKHOFF C 、ANDREARCZYK V 和 MÜLLER H. 图像谱号 2018 字幕预测任务概述。谱号（工作笔记），2018。
[5] EICKHOFF C、SCHWALL I、DE HERRERA AGS、和 MÜLLER H. 图像谱号字幕概述 – 生物医学图像的图像标题预测和概念检测，2017 年。谱号（工作笔记），2017。
[6] PALOTTI JR，ZUCCON G，GOEURIOT L，等人。谱号电子健康评估实验室 2015，任务 2：检索有关医疗症状的信息。谱号（工作笔记），2015：1-22。
[7] GUO Z, WANG X, ZHANG Y, 和 LI J. 图像谱号医学标题处的图像扫描仪 2019 任务：两阶段医学概念检测策略。谱号 2019 工作笔记。欧元研讨会论文集，（2019：1613-0073。
[8] SINHA P、PURKAYASTHA S 和 GICHOYA J. 图像谱号 2019 挑战的放射学图像概念检测任务的全面训练与微调。谱号 2019 工作笔记。欧元研讨会论文集，2019：1613-0073。
[9] LIANG S, LI X, ZHU Y，等。国际工业协会在图像谱号 2017 图片说明任务上。谱号（工作笔记），2017。
[10] LITJENS G、KOOI T、BEJNORDI BE 等人。医学图像分析中的深度学习调查。医学影像分析，2017，42：60-88。
[11] KOHLI M、PREVEDELLO LM、FILICE RW 和 GEIS JR。在放射学实践和研究中实施机器学习。美国放射学杂志 2017，208：754–760。
[12] WANG X, GUO Z, 和 ZHANG Y, LI J。临床应用中的医学图像标注与语义理解。欧洲语言跨语言评估论坛国际会议论文集，施普林格，2019：260-270。
[13] RAHMAN M. CS 摩根州立大学的基于跨模式深度学习的字幕预测和概念检测方法。谱号（工作笔记），2018。
[14] PINHO E 和 COSTA C. 使用对抗网络进行特征学习，用于医学图像中的概念检测：UA。PT 图像谱号的生物信息学 2018。谱号（工作笔记），2018。
[15] DIMITRIS K 和 ERGINA K. 使用深度残差学习网络对医学图像进行概念检测。工作笔记谱号 2017。
[16] KOUGIA V 、 PAVLOPOULOS J 和 ANDROUTSOPOULOS I. AUEB NLP 小组在图像谱号医学标题 2019 上。谱号 2019 工作笔记。欧元研讨会论文集，2019：09-12。
[17] RAJPURKAR P, IRVIN J, ZHU K, 等人。互联网：通过深度学习对胸部 X 光片进行放射科医生级别的肺炎检测。arXiv 预印本 arXiv, 2017, 1711.05225。
[18] HUANG G, LIU Z, VAN DER MAATEN L 和 WEINBERGER KQ。密集连接的卷积网络。IEEE 计算机视觉和模式识别会议论文集，2017：4700–4708。
[19] 徐杰, 刘伟, 刘 C, 等。图像谱号 2019 上基于多标签分类和图像字幕方法的概念检测。谱号 2019 工作说明。欧元研讨会论文集，2019：1613-0073。
[20] ZHANG Y, WANG X, GUO Z, 和 LI J. 图像谱号上的图像扫描仪 2018 标题任务：图像检索和迁移学习。谱号欧元研讨会，法国阿维尼翁，2018 年。
[21] GONALVES A、PINHO E 和 COSTA C。信息丰富且有趣的视觉特征：图像谱号标题 2019 中的 Ua.pt 生物信息学。谱号 2019 工作笔记。欧元研讨会论文集，2019：1613-0073。
[22] ABACHA AB、DE HERRERA AGS、GAYEN S 等人。图像谱号 2017 字幕任务中的 NLM。谱号（工作笔记），2017。
[23] HASAN SA, LING Y, LIU J, 等。图像谱号 2017 字幕预测和概念检测任务中的核糖核酸。谱号（工作笔记），2017。
[24] LYNDON D、KUMAR A 和 KIM J. 图像谱号 2017 医学影像挑战赛的神经字幕。谱号（工作笔记），2017。
[25] VINYALS O、TOSHEV A、BENGIO S 和 ERHAN D. 展示和讲述：神经图像字幕生成器。IEEE 计算机视

觉和模式识别会议论文集，2015：3156-3164。

[26] KATSIOS D 和 KAVALLIERATOU E. 使用深度残差学习网络对医学图像进行概念检测。谱号（工作笔记），2017。

[27] NY HOAVY N 、 MOTHE J 和 RANDRIANARIVONY M.I. IRIT 和米萨在图像谱号 2017-多标签分类。谱号协会国际会议论文集，谱号 2017 实验室工作说明（谱号 2017），2017（爱尔兰都柏林）: 1-7。

[28] STEFAN LD、IONESCU B 和 MÜLLER H。使用深度学习多假设方法为医学图像生成字幕：医疗礼品-UPB 参与图像谱号 2017 字幕任务。欧元研讨会论文集，2017 年，1856 年：第 1523 条。

[29] JOLY A、GOËAU H、GLOTIN H 等人。救生圈 2017 实验室概述：多媒体物种识别挑战。欧洲语言跨语言评估论坛国际会议论文集，施普林格，2017：255-274。

[30] VILLEGAS M, MÜLLER H, GILBERT A, 等人。谱号 2015 实验室中图像谱号的一般概述。欧洲语言跨语言评估论坛国际会议。在：MOTHE J. 等人。(编辑) 实验红外满足多语言、多模态和交互。谱号 2015。计算机科学讲义，9283：444-461，https://doi.org/10.1007/978-3-319-24027-5_45

[31] SOLDAINI L 和 GOHARIAN N. 快捷方式：一种快速、无监督的医学概念提取方法。医学红外研讨会，SIGIR，2016 年。可从 http://medir2016.imag.fr/data/MEDIR_2016_paper_16.pdf 获取

[32] KHAN FU, 和 AZIZ IB. 通过一种新的基于集群的合成过采样技术减少医学图像采集中的高可变性。IEEE 大数据与分析会议 (ICBDA) 论文集，2019：45-50。

[33] KUBANY A、ISHAY SB、OHAYON RS 等人。使用语义指标比较用于图像多标签分类的最先进的深度学习 API。具有应用程序的专家系统，2020，161：113656。

[34] MAEDA-GUTIÉRREZ V、GALVÁN-TEJADA CE、ZANELLA-CALZADA LA 等人。用于番茄植物病害分类的卷积神经网络架构比较。应用科学，2020 年，10：1245。

[35] SZEGEDY C、VANHOUCKE V、IOFFE S 等人。重新思考计算机视觉的初始架构。IEEE 计算机视觉和模式识别会议论文集，2016：2818-2826。

[36] SIMONYAN K, 和 ZISSERMAN A. 用于大规模图像识别的非常深的卷积网络。arXiv 预印本 arXiv: 2014, 1409.1556。

[37] CHOLLET F. 异常：深度可分离卷积的深度学习。IEEE 计算机视觉和模式识别会议论文集，2017：1251-1258。

[38] HOWARD A.G, ZHU M, CHEN B, 等人。移动网络：用于移动视觉应用的高效卷积神经网络。ArXiv 预印本 arXiv：2017，1704.04861。

[39] ROSEBROCK, A. 喀拉斯的多标签分类。2018 。可从 https://www.pyimagesearch.com/2018/05/07/multi-label-classification-with-keras/2018 获得（2020 年 5 月 3 日访问）。

[40] PELKA O、KOITKA S、RÜCKERT J 等人。上下文中的放射学对象 (洛可可)：多模态图像数据集。在血管内成像和计算机辅助支架以及生物医学数据的大规模注释和专家标签合成方面；斯普林格，2018: 180-189。

[41] KHAN FU，AZIZ IB，和 AKHIR EAP。用于动态深度神经表面中层配置中继的可插拔微网络。IEEE 访问，2021，9：124831–124846。doi:10.1109/ACCESS.2021.3110709。

[42] KHAN FU，AZIZ IB，和 AKHIR EAP。总理网：在早期卷积阶段增强特征选择的自适应多层深度神经结构。预印本，2021、2021070699（土井：10.20944/preprints202107.0699.v1）。