Volumetric Model Genesis in Medical Domain for the Analysis of Multimodality 2-D/3-D Data Based on the Aggregation of Multilevel Features

Muhammad Owais, Se Woon Cho, and Kang Ryoung Park, *Member, IEEE*

***Abstract*—The automatic and accurate classification of medical imaging data has potential applications in computer-aided disease diagnosis, prognosis, and treatment. However, it remains a challenge to optimize recent deep learning algorithms in the medical domain for the accurate classification of large-scale three-dimensional (3-D) volumetric data. To address these challenges, we propose an efficient deep volumetric classification network based on the aggregation of multilevel deep features for the accurate classification of large-scale medical 2-D/3-D imaging data. To perform a detailed quantitative analysis of our method, 26 different datasets were fused to construct a single large-scale multimodal database that comprises a total of seventy different classes, including 151,095 data samples. Additionally, 15 different baseline methods were configured under the same experimental protocol for volumetric model genesis and extensive performance comparison with our method. The experimental results of our method exhibited promising performance as an area under the curve of 93.66% and outperformed various state-of-the-art methods.**

***Index Terms*—Computer-aided diagnosis (CAD), medical data analysis, three-dimensional (3-D) deep learning (DL), volumetric model genesis.**

Manuscript received 11 May 2022; revised 8 December 2022 and 19 January 2023; accepted 24 February 2023. Date of publication 6 March 2023; date of current version 18 October 2023. This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) through the Basic Science Research Program under Grant NRF-2021R1F1A1045587, in part by the NRF funded by the MSIT through the Basic Science Research Program under Grant NRF-2022R1F1A1064291, and in part by the MSIT, Korea, under the ITRC (Information Technology Research Center) support program under Grant IITP-2023-2020-0-01789 supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). Paper no. TII-22-2035. *(Corresponding author: Kang Ryoung Park.)*

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

Muhammad Owais is with the KUCARS and C2PS, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi 127788, UAE, and also with the Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea (e-mail: malikowais266@gmail.com).

Se Woon Cho and Kang Ryoung Park are with the Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea (e-mail: jsu319@dongguk.edu; parkgr@dgu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2023.3252541.

Digital Object Identifier 10.1109/TII.2023.3252541

I. Introduction

WITH the development of digital devices, the use of different types of imaging modalities [such as magnetic resonance imaging (MRI), X-rays, optical projection tomography (OPT), ultrasonography, computed tomography (CT), angiography, positron emission tomography (PET), and visible light cameras] has become commonplace in the medical diagnostic domain. These imaging modalities provide diagnostic assistance to medical experts by capturing the visual representation of different body organs as 2-D/3-D imaging data [1], [2]. Consequently, the production of multimodal 2-D/3-D imaging data has grown exponentially in recent years. In addition, the application of multimodal data in various medical diagnosis areas is also increasing rapidly. For example, multimodal images such as CT and MRI images are being fused to create a single mark image that can be more suitable for diagnostic evaluation than individual images [3], [4]. Recently, a variety of multimodal fusion-based algorithms are evolving for safe and secure telehealth applications [5]. Therefore, effective organization and analysis of existing multimodal data can offer various potential applications in the medical domain. For example, medical professionals can obtain a diagnostic clue for a complex medical condition by retrieving relevant cases from the existing database using efficient classification algorithms. Consequently, an accurate and timely diagnosis of acute medical conditions results in better treatment [1], [2].

However, subjective exploration, classification, and retrieval of intended content from a huge collection of visual data are challenging and time-consuming tasks. Recently, advancements in the artificial intelligence (AI) domain have provided various potential applications in general, as well as in the medical field [6], [7]. Efficient analysis of medical imaging data is also one of the key applications of AI algorithms. Consequently, various state-of-the-art computer-aided diagnosis (CAD) methods have been proposed in the literature that utilizes the power of AI in medical data analysis and enable effective diagnostic decisions [8], [9], [10], [11]. Among the different AI methods, a subset of deep learning (DL) algorithms has received special attention owing to its remarkable performance, particularly in the case of visual data analysis [12], [13], [14]. Such DL-driven CAD methods mimic the processing of the human brain and deliver accurate diagnostic results, similar to those of medical experts. With respect to image- and sequence-based CAD methods, convolutional neural networks (CNNs), a well-known variant of DL algorithms, have received special consideration. Various types of CNNs [15], [16], [17], [18] have been proposed in the literature for general and medical applications. The structure of a CNN model mainly comprises multiple convolutional layers and fully connected (FC) layers, including trainable parameters [9]. Initially, these parameters are trained using an independent training dataset. Consequently, a trained model classifies the testing data sample into its target class after analyzing it through multiple convolutional and FC layers.

A. Potential Research Gaps and Motivation

Most of the existing methods [18], [19], [20], [21] are disease and modality-specific, and optimized for a limited number of data samples. Moreover, these methods are designed to make diagnostic decisions based on 2-D imaging data employing image-based classification models [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], even in case of 3-D imaging data. There is very limited research related to the joint classification of multimodal 2-D/3-D imaging data considering a large number of classes. The main objective of this study is to encapsulate the computer-aided diagnostic capability of various kinds of diseases in a single DL model that can be scaled up in future work by including more data and classes. In addition, we aim to provide new grounds for developing an efficient jointly connected content-based medical image and sequence retrieval (CBMISR) framework by applying our proposed classification model. Various existing medical retrieval methods in the literature are image-based and consider limited classes and data samples to validate their proposed models. Therefore, we further highlight the application of our proposed model in developing a jointly connected 2-D/3-D imaging retrieval framework. An efficient CBMISR framework can assist medical professionals in validating their diagnostic decision for a complex medical condition by retrieving relevant cases from the existing database. Finally, we aim to encapsulate the diverse features of large-scale medical 2-D/3-D imaging data in a single model that can provide new grounds for future research related to medical domain-specific transfer learning (MDS-TL). Based on the proposed framework, the strengths of MDS-TL can be further explored and additional performance improvements can be achieved in various medical diagnostic applications.

B. Main Contributions

We mainly propose a novel jointly connected classification framework based on a multiscale dilated fused (DF) residual network (MDF-RN) and a spatiotemporal block classification network (STB-CN) for the classification of both medical 2-D/3-D imaging data. This is the first study to present a pretrained classification model in the medical domain, which is trained with a large-scale multimodal database that includes both 2-D/3-D imaging data. The main contributions of this study are as follows.

1) The main contribution is the development of a novel 2-D-CNN architecture (named MDF-RN) that leverages multiscale dilated convolution and a concept of multilevel feature fusion in a mutually beneficial manner to achieve state-of-the-art performance.
2) Three additional branches are created in the proposed MDF-RN model by including three DF-blocks that primarily exploit multiscale/multilevel features and enhance the overall performance.
3) Subsequently, the second-stage STB-CN model further utilizes the strength of recurrent neural networks (RNNs) and transfer learning in classifying 3-D imaging data without influencing the overall training parameters of the whole pipeline (MDF-RN+STB-CN).
4) The proposed STB-CN model works for both 2-D and 3-D imaging data and does not limit the processing of fixed-length sequences as restricted by 3-D-CNNs, but can classify variable-length sequences of successive slices/frames.
5) In addition, we evaluated the performance of fifteen state-of-the-art image-based and sequence-based classification models to provide standard benchmarks for this study. Finally, our proposed classification framework (including implementations of both MDF-RN and STB-CN models) is freely accessible to enable fair comparisons and future research.

The rest of this article is organized as follows: Section II presents a brief literature review related to DL-driven CAD methods. Section III provides a detailed explanation of the proposed methodology. In Section IV, we briefly describe the datasets, experimental protocols, and results. Finally, the results are discussed and conclusions are drawn in Sections V and VI, respectively.

II. Related Work

This section presents a brief overview of the existing state-of-the-art CAD methods related to the classification of medical imaging data. These methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] utilized the strength of transfer learning by employing existing pretrained CNN models [22], [23], [24], [25], [26], [27], [28], [29], [30], [31] in developing their CAD solutions. All these studies cover a vast scope of 1) disease-specific, 2) modality-specific, 3) multimodality-based, and 4) multidisease-based CAD solutions. In addition, the source codes of all these methods are also publicly available for a fair comparison. Therefore, we selected these surveyed papers in this section. In particular, these CAD methods classify 2-D and 3-D imaging data into different categories, including normal and diseased classes.

A. Image-Based Methods (2-D Models)

In the context of 2-D imaging data, Adnan et al. [7] proposed a classification-based medical image retrieval framework using a revised version of AlexNet [21] that can classify multimodal (CT, MRI, PET, OPT, and fundus camera) 2-D imaging data into one of 24 different classes. In another study, Falconi et al. [8] utilized the strength of transfer learning in breast lesion classification tasks using different pretrained CNN models. Three CNN architectures were modified and trained to classify mammogram images into one of six classes. Among the existing models, VGG19 [23] achieved superior results. Later, Owais et al. [9] addressed the limitations of [7] and proposed a new content-based medical image retrieval framework based on a modified version of ResNet50 [24], which was trained to classify multimodal 2-D imaging data into one of 50 different categories, including both disease and normal cases.

Apart from [7] and [9], most of the existing CNN-based CAD methods are domain-specific and perform binary classification (either diseased or normal). For example, Kaur et al. [10] proposed a CAD method using a pretrained VGG16 model [23] with the capability to categorize pathological brain images as normal or abnormal. However, a limited number of data samples (20 normal and 140 abnormal MRI images) were used to validate the proposed method. Subsequently, Ashraf et al. [11] used another pretrained CNN named GoogleNet [25] for medical image classification. A multimodal dataset (including 3600 images related to 12 different categories) was used to train and validate the method. Their method also includes a limited number of data samples (300 images per class). Similar to [10], Akpinar et al. [12] proposed a binary-classification CAD method for detecting chest abnormalities. An existing pretrained SqueezeNet [26] model was employed to categorize X-ray images into normal or abnormal groups. In total, 660 X-ray images were used to validate the method. Subsequently, Aloyayri et al. [14] utilized the strength of transfer learning in breast cancer classification using histopathological images. Three different CNN architectures were trained to classify data samples as either benign or malignant. Among the different baseline models, ResNet18 [24] achieved superior results.

Souid et al. [15] proposed a multiclass diagnostic framework for chest lesions. A lightweight deep CNN model, named MobileNetV2 [27], was trained to predict multiclass lung pathologies (considering 14 different classes) from chest X-ray images. A single-modality large-scale dataset, including a total of 64699 images, was used to calculate the performance of the CAD method. Similarly, Jasil et al. [16] and Çakmak et al. [17] utilized different CNNs for skin lesion classification tasks. In [16], a pretrained CNN, named DenseNet201 [28], was employed to classify dermoscopy images into one of seven different classes of skin lesions. A single-modality limited dataset, including a total of 3091 images, was used to validate the method. Later, Çakmak et al. [17] used a lightweight CNN, named NASNet-Mobile [29], for melanoma detection from dermoscopy images. They considered a larger dataset (including a total of 10015 images) than that of [16]. In the context of diabetic retinopathy (DR), Gambhir et al. [18] proposed a severity classification CAD method that was able to detect and distinguish DR into different severity levels. An existing ShuffleNet [30] model was trained to categorize the input DR image into one of five different classes (including one normal and four diseased cases).

B. Sequence-Based Methods (3-D Models)

There is very limited research related to the classification of large-scale multimodal 3-D imaging data for clinical decision support systems. For example, Shahzadi et al. [19], Srinivasu et al. [20], and Ebrahimi et al. [21] proposed sequence-based classification methods using existing CNNs and long short-term memory (LSTM) models [31]. Shahzadi et al. [19] proposed a binary-classification framework (comprising VGG16 [23] and LSTM [31] models) for the recognition of brain tumors from 3-D brain MRI scans. Subsequently, a skin lesion classification framework based on the MobileNetV2 [27] and LSTM [31] models was proposed by Srinivasu et al. [20]. Rather than using a single image, a sequence of images was used for disease classification. Similar to [19], another binary-classification framework for Alzheimer's disease detection was proposed in [21]. A cascade of ResNet18 [24] and LSTM models [31] was configured using 3-D brain MRI scans. Ebrahimi et al. [21] used a larger dataset (compared to [19]), including a total of 35550 MRI samples.

C. Limitations of the Existing Methods

The concept of joint multiscale and multilevel feature fusion has gained less attention in medical 2-D/3-D imaging data classification. Different fusion techniques, such as early fusion, late fusion, and ensemble learning [6], exist and have improved DL performance. However, they require additional pre- and postprocessing overhead and influence the overall computational cost of a DL model. In a recent study, Abdar et al. [6] explored the strength of multilevel feature fusion by employing the concept of conventional ensemble modeling and proposed an image-based classification model. However, their proposed feature extractor scheme consists of a total of four different pretrained models, containing a total of 162 million trainable parameters and requiring extensive computational overhead. In addition, most existing studies related to medical data analysis are disease-specific and consider a limited number of classes, as well as data samples, to develop and validate their proposed classification methods. Moreover, various methods employed image-based models that consider only spatial information in making diagnostic decisions in the case of 3-D imaging data such as CT or MRI scans. Consequently, the loss of 3-D anatomical information may result in false predictions and finally a decrease in the overall prediction probability of the testing data.

D. Singularity of Our Method

To address the limitations of existing studies, we propose a domain-specific pretrained model related to medical diagnostic applications using large-scale, multiclass, and multimodal 2-D/3-D imaging data. This is the first study to present a pretrained classification model in the medical domain including both 2-D/3-D imaging data. In total, 26 publicly available datasets (based on 11 different modalities) [9], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41] were fused to construct a single large-scale database comprising 70 different classes, including 151,095 data samples. The proposed CAD solution utilized the strength of multiscale/multilevel feature fusion and encapsulates the computer-aided diagnostic capability of various kinds of diseases in a single DL model. Our proposed model leverages transfer learning in classifying 3-D imaging data without influencing the overall training parameters and works for both 2-D and 3-D

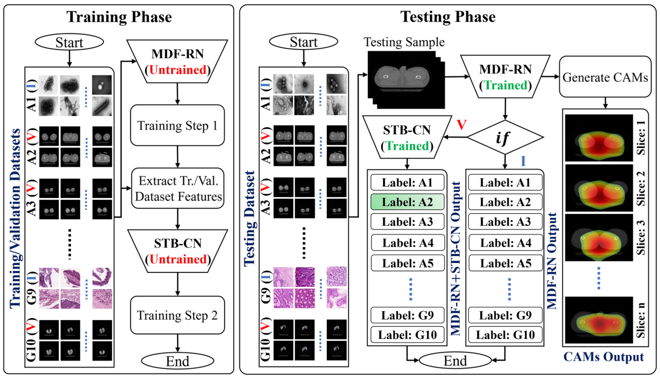


Fig. 1. Comprehensive workflow diagram of the proposed classification framework (MDF-RN+STB-CN), including both training and testing phases.

imaging data. It has the capability to classify variable-length sequences in case of 3-D imaging data. The experimental results reveal the superior performance of the proposed framework over various state-of-the-art methods.

III. Proposed Method

A. Workflow Overview

This study aims to develop a deep classification model with the capability to classify multiclass medical data, including both 2-D/3-D imaging data. In particular, the proposed method can classify a variable-length sequence of n successive slices/frames (i.e., $\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3, \ldots, \mathrm{I}_\mathrm{n}, \boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{F}^3, \ldots, \boldsymbol{F}^n$) with significant performance gain compared to image-based models. After selecting appropriate datasets, we developed a cascade of two classification networks, MDF-RN and STB-CN, for the accurate classification of multimodal 2-D/3-D imaging data. The overall procedure of the proposed model development mainly includes a training phase followed by a testing phase as shown in Fig. 1. Both networks were trained, validated, and tested using independent training, validation, and testing datasets. In the first step, an untrained MDF-RN model was trained to exploit and learn the spatial features from the training dataset that included a total of p data samples and corresponding class labels notated as $\langle[\boldsymbol{F}_T]_{i=1}^p, [l_T]_{i=1}^p\rangle$. In the next step, all training data samples $[\boldsymbol{F}_T]_{i=1}^p$ were converted into feature vectors $[\boldsymbol{f}_T]_{i=1}^p$ after processing each data sample through our trained MDF-RN model. Consequently, we obtained a new training dataset (denoted as $\langle[\boldsymbol{f}_T]_{i=1}^p, [l_T]_{i=1}^p\rangle$) in the feature domain. In the next step, the second untrained STB-CN model was trained to learn the 3-D anatomical dependencies (in the case of 3-D imaging data) from $\langle[\boldsymbol{f}_T]_{i=1}^p, [l_T]_{i=1}^p\rangle$. We divided the training data into 2-D and 3-D imaging data according to the given information in each class label. In detail, all the classes with 2-D imaging data are notated with "I" along with the name of their actual class labels as shown in Fig. 1. Similarly, 3-D imaging classes are differentiated with "V" along with the name of their actual class labels as mentioned in Fig. 1.

After training, the performance of the proposed classification framework (MDF-RN+STB-CN) was evaluated for an independent testing dataset, denoted as $\langle[\boldsymbol{F}_{Ts}]_{i=1}^r, [l_{Ts}]_{i=1}^r\rangle$. In the case of 2-D images, a trained MDF-RN model exploits the spatial features and performs class prediction. In the case of 3-D imaging data such as endoscopy videos, CT, and MRI scans, the second trained STB-CN further improves the overall performance by exploiting 3-D anatomical dependencies and results in an additional performance gain. Initially, MDF-RN sequentially transforms the sequence of n successive slices/frames (i.e., $\boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{F}^3, \ldots, \boldsymbol{F}^n$) into n feature vectors (i.e., $\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \ldots, \boldsymbol{f}^n$). Then, the second stage STB-CN model parallelly processes these feature vectors to exploit additional 3-D anatomical features and perform class prediction. To provide visual insight into the network decision, we also visualized the class activation map for each input 2-D or 3-D imaging data sample as an additional output (see Fig. 1).

B. MDF-RN Model Structure and Workflow

To achieve superior classification performance and fast execution speed, the proposed MDF-RN design utilizes the following strengths.

1) Residual blocks [labeled skip residual (SR)-block and projected residual (PR)-block in Fig. 2] of ResNet (RN) models [24].
2) Our newly included DF-block (as shown in Fig. 2) based on multiscale dilated convolution layers.
3) A concept of multilevel feature fusion in a mutually beneficial manner.

The complete structure of our MDF-RN model includes five SR-blocks, three PR-blocks, three DF-blocks, and some other layers, as shown in Fig. 2.

1) Residual Blocks: In general, residual blocks (SR- and PR-blocks) avoid the vanishing gradient problem in a training process and achieve the optimal convergence of the entire network. Therefore, we selected the residual blocks in our network design to exploit high-level semantic features. Both residual blocks consist of two 3×3 convolutional layers and a residual connection, as shown in the bottom-left corner of Fig. 2. The SR-block includes a SR connection and transforms the input tensor $\boldsymbol{F}_k \in \mathcal{R}^{w_k \times h_k \times d_k}$ into the final output tensor $\boldsymbol{F}_l \in \mathcal{R}^{w_k \times h_k \times d_k}$ without influencing the dimension. By contrast, the PR-block consists of a PR connection based on a 1×1 convolutional layer and maps the input tensor $\boldsymbol{F}_k \in \mathcal{R}^{w_k \times h_k \times d_k}$ into the final output tensor $\boldsymbol{F}_l \in \mathcal{R}^{w_k/2 \times h_k/2 \times 2d_k}$. Mathematically, the input tensor $\boldsymbol{F}_k \in \mathcal{R}^{w_k \times h_k \times d_k}$ undergoes the following transformations after passing through these residual blocks:

$$\Psi_{SR}\left(\boldsymbol{F_k}, \varphi\right) = h_{\varphi_l}\left(h_{\varphi_k}\left(\boldsymbol{F_k}\right)\right) + \boldsymbol{F_k} \tag{1}$$

$$\Psi_{PR}\left(\boldsymbol{F_k}, \varphi\right) = h_{\varphi_l}\left(h_{\varphi_k}\left(\boldsymbol{F_k}\right)\right) + h_{\varphi_m}\left(\boldsymbol{F_k}\right) \tag{2}$$

where $\Psi_{SR}(\cdot)$ and $\Psi_{PR}(\cdot)$ denote the SR- and PR-blocks as transfer functions, respectively. $h_{\varphi_k}(\cdot)$ $h_{\varphi_l}(\cdot)$, and $h_{\varphi_m}(\cdot)$ represent the convolutional layers with training parameters φ_k, φ_l, and φ_m, respectively.

2) DF Block: Additionally, we proposed a DF-block (as shown in the bottom-left corner of Fig. 2) followed by an average pooling layer to capture a multiscale representation of multilevel (i.e., low-, intermediate-, and high-level) features acquired from

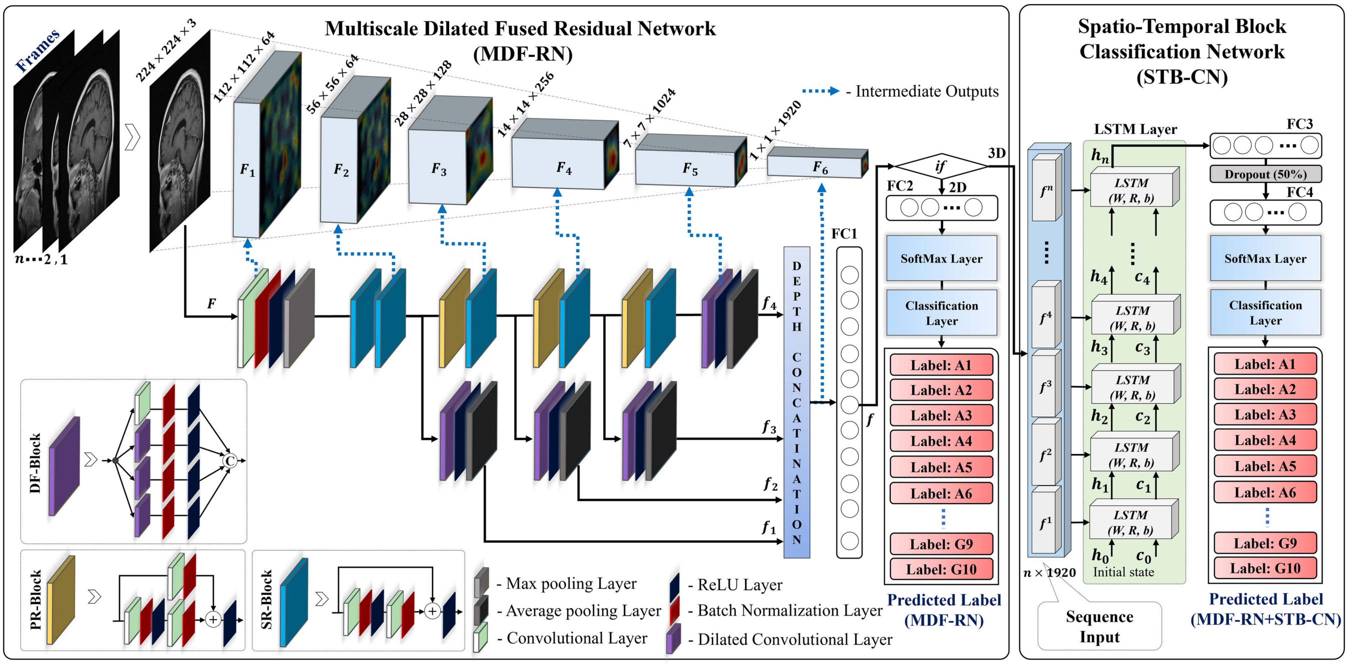


Fig. 2. Overall architecture of the proposed classification framework including both MDF-RN and STB-CN models.

different residual blocks (see Fig. 2). The key intuition behind the development of DF-block is to aggregate the multiscale representation of deep features at different resolutions. Quantitative results (in the result section) have shown the significant strength of our designed DF-block. The structure of our proposed DF-block includes a total of three parallelly connected dilated convolutional layers (with a filter size of 3×3 and dilation rates of 6, 12, and 18) and a PR-connection based on a 1×1 convolutional layer. The DF-block transforms the input tensor $\boldsymbol{F}_k \in \mathcal{R}^{w_k \times h_k \times d_k}$ into the output tensor $\boldsymbol{F}_l \in \mathcal{R}^{w_k \times h_k \times 2d_k}$ by exploiting additional multiscale features from the output of different residual blocks. Mathematically, the input tensor $\boldsymbol{F}_k \in \mathcal{R}^{w_k \times h_k \times d_k}$ undergoes the following transformations after passing through a DF-block:

$$\Psi_{DF}\left(\boldsymbol{F_k}, \varphi\right) = h^*_{\varphi_k^{18}}\left(\boldsymbol{F_k}\right) \odot h^*_{\varphi_k^{12}}\left(\boldsymbol{F_k}\right) \odot h^*_{\varphi_k^{6}}\left(\boldsymbol{F_k}\right) \odot h_{\varphi_k}\left(\boldsymbol{F_k}\right) \quad (3)$$

where $\Psi_{DF}(\cdot)$ denotes the DF-block as a transfer function. $h_{\varphi_k}(\cdot)$ and $h^*_{\varphi_k^x}(\cdot)$ represent simple and dilated convolutional layers with training parameters φ_k and φ_k^x, respectively. The symbol $\odot$ presents the depth-wise feature concatenation.

3) Multilevel Feature Fusion: A concept of multilevel feature fusion is introduced in our MDF-RN model by aggregating the joint contribution of the multiscale low-, intermediate-, and high-level semantic features (i.e., $\boldsymbol{f}_1$– $\boldsymbol{f}_4$) in the final classification decision. These multilevel features are obtained from different residual blocks using multiple DF-blocks (see Fig. 2) and provide a diverse representation of a particular class. A detailed ablation study (in a later section) shows the substantial contribution of multilevel feature fusion in achieving state-of-the-art performance.

4) Model Workflow: Initially, a 7×7 convolutional layer explores the input image $\boldsymbol{F}$ and generates an output tensor of size 112×112×64, which is further processed by a 3×3 max-pooling layer and downsampled into a new output tensor of size 56×56×64. Consequently, a stack of nine building blocks (including five SR-blocks, three PR-blocks, and one DF-block, as shown in Fig. 2) sequentially processes the output of the preceding layer/block and finally generates a multiscale high-level feature vector of size 1×1×1024 (labeled as $\boldsymbol{f}_4$ in Fig. 2).

Additionally, three DF-blocks were included to exploit multiscale low- and intermediate-level semantic features (i.e., $\boldsymbol{f}_1$–$\boldsymbol{f}_3$) from three different residual blocks. These residual blocks are selected based on the different spatial sizes of their output tensors (i.e., 56×56, 28×28, and 14×14) to obtain low- and intermediate-level semantic features. Moreover, each DF-block is followed by an average pooling layer that further transforms the 2-D output tensor of the DF-block into a 1-D vector space. A depth concatenation layer followed by the first FC layer (FC1; Fig. 2) fused all multilevel semantic features (i.e., $\boldsymbol{f}_1$– $\boldsymbol{f}_4$) and further exploited more discriminative patterns. Consequently, we obtained a multilevel semantic representation of input image $\boldsymbol{F}$ as an output feature vector $\boldsymbol{f}$ of size 1×1×256. In the case of a 2-D image, the MDF-RN model further performs the class prediction by processing the output feature vector $\boldsymbol{f}$ with a stack of three additional layers (FC2, SoftMax, and classification layers; Fig. 2).

C. STB-CN Model Structure and Workflow

In the case of 3-D imaging data consisting of n successive slices/frames (i.e., $\boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{F}^3, \ldots, \boldsymbol{F}^n$), the proposed MDF-RN model sequentially processes each input slice/frame and generates a set of n feature vectors (i.e., $\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \ldots, \boldsymbol{f}^n$) of size $1\times1\times256\times n$. All these feature vectors are extracted from the FC1 layer of our MDF-RN model. These feature vectors are further processed by the second-stage STB-CN model to exploit additional 3-D anatomical features and perform class prediction. The STB-CN includes a revised variant of RNNs called the LSTM model [19], [20], [21], which resolves the vanishing gradient problem in the training process and can leverage transfer learning in the case of volumetric data analysis without influencing the overall training parameters. Therefore, we utilized the strength of LSTM in designing our second-stage STB-CN model for the effective classification of volumetric data in the medical domain.

The overall structure and workflow of the proposed STB-CN are shown in Fig. 2. First, a sequence input layer passes a set of n feature vectors (i.e., $\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \ldots, \boldsymbol{f}^n$) to the LSTM layer, which exploits additional 3-D anatomical dependencies among these feature vectors after processing through a sequence of n LSTM cells (see Fig. 2) and finally generates a single feature vector $\boldsymbol{h}_n$ of size $1\times1\times1200$ (obtained from the last LSTM cell). The output feature vector $\boldsymbol{h}_n$ incorporates both 2-D spatial and 3-D anatomical information of the 3-D imaging data (i.e., $\boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{F}^3, \ldots, \boldsymbol{F}^n$) and further refined by a third FC layer (FC3; Fig. 2) to exploit more discriminative patterns. Finally, a stack of three additional layers (FC4, SoftMax, and classification layers; Fig. 2) predicts a single class label for the entire 3-D imaging data sample based on the highest probability score (similar to MDF-RN) using the final output feature vector $\boldsymbol{h}_n$.

D. Training Loss

A two-step training process of both the MDF-RN and STB-CN models was performed sequentially to attain optimal convergence of our proposed classification framework. In the first step, the MDF-RN was trained to exploit and learn the spatial features from the entire training dataset denoted as $[\boldsymbol{F}_T]_{i=1}^p, [l_T]_{i=1}^p$ using a cross-entropy (CE) loss function [9]. The initial weights of different residual blocks in MDF-RN were obtained from a pretrained RN [24] that was trained with a large-scale ImageNet dataset using the CE loss function. Therefore, a similar loss function was used to train our MDF-RN model. In the next step, the training and validation datasets were converted into training (denoted as $[\boldsymbol{f}_T]_{i=1}^p, [l_T]_{i=1}^p$) and validation (denoted as $[\boldsymbol{f}_V]_{i=1}^q, [l_V]_{i=1}^q$) feature vectors after processing each data sample through MDF-RN. Subsequently, the second STB-CN model was trained to learn the 3-D anatomical dependencies in the case of 3-D imaging data using the same CE loss function. The overall two-step loss function of the proposed models can be expressed as

$$\mathcal{L}oss = \begin{cases} \underset{w'_{MDF-RN}}{\arg\min}\ \mathcal{L}_1\left(\psi_1\left(w_{MDF-RN}, [F_T]_{i=1}^p\right), [l_T]_{i=1}^p\right), & \text{Step 1} \\ \underset{w'_{STB-CN}}{\arg\min}\ \mathcal{L}_2\left(\psi_2\left(w_{STB-CN}, [f_T]_{i=1}^p\right), [l_T]_{i=1}^p\right), & \text{Step 2} \end{cases} \tag{4}$$

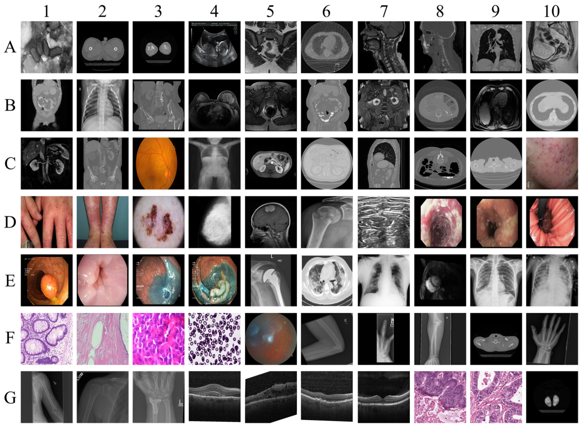


Fig. 3. Visualization of a few data samples for each class in our dataset.

where ψ_1 and ψ_2 represent the MDF-RN and STB-CN models as transfer functions, respectively. $\mathcal{L}_1(\cdot)$ and $\mathcal{L}_2(\cdot)$ are the CE loss functions.

IV. Results and Analysis

A. Dataset and Experimental Setup

To perform a quantitative analysis of our proposed classification framework, twenty-six different datasets as used in [9], [32], [33], [34], [35], [36], [37], [38], [39], [40], and [41] were fused to build a single large-scale database that included a total of 151,095 data samples. Consequently, the whole dataset was divided into 70 different classes according to the given ground-truth labels, which included various types of normal and disease categories. In this study, we tried our best to select various publicly available 2-D/3-D imaging datasets related to the medical diagnostic domain. Therefore, we explored numerous publicly available datasets and eventually, selected well-known data repositories from large publicly available collections based on their publication venue. To provide a visual representation of our final dataset, Fig. 3 shows a few example images of each class. In addition, Table I shows the details of each class in our selected datasets in notation L(X/Y/Z) that provides the following information:

1) actual ground-truth label (L),
2) type of imaging modality (X),
3) whether it includes 2-D or 3-D imaging data (Y), and
4) the total number of data samples in terms of the number of images/slices/frames (Z).

Most of the 3-D imaging data are related to CT and MRI imaging modalities that do not include time information. However, their length information (as a number of slices) is included in Table I. In detail, the number of slices for each class related to 3-D imaging data was determined by counting the total number of all the slices in each 3-D scan of a particular class. In addition, such meta-information (i.e., number of slices) was also provided in each dataset related to CT and MRI imaging modalities. A few classes (i.e., D8, D9, D10, E1, and E2) comprise endoscopy data encoded at a frame rate of 25 frames per second and having a

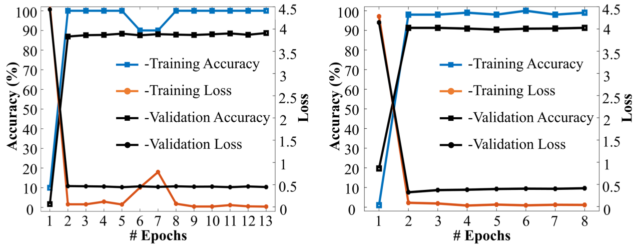
TABLE I

Brief Description of Each Class in Our Selected Datasets Is Provided in Notation L(X/Y/Z), Where L: Actual Label, X: Imaging Modality, Y: 2D Imaging Data (I) or 3-D Imaging Data (V), and Z: Total Number of Data Samples. ("CT: Computed Tomography," "MS: Microscope," "MRI: Magnetic Resonance Imaging," "XR: X-Rays," "VLC: Visible Light Camera," "PET: Positron-Emission Tomography," "US: Ultrasound," "ES: Endoscopy," "FC: Fundus Camera," "CNV: Choroidal Neovascularization," "OCT: Optical Coherence Tomography," "DME: Diabetic Macular Edema," "GI: Gastrointestinal," "MSI: Microsatellite Instability," "MSS: Microsatellite Stability"). Note: The Notation "A1, A2, …, G10" Presents "Class 1, Class 2, …, Class 70"

	1	2	3	4	5	6	7	8	9	10
A	Microscopic Virus (MS/I/1245)	Hip (CT/V/800)	Knee (CT/V/350)	Uterus Normal (US/V/1000)	Cervix Carcinoma (MRI/V/1000)	Lung Cancer (CT/V/1000)	Head & Neck Cancer (MRI/V/1000)	Head & Neck Cancer (CT/V/1000)	Lung Normal (CT/V/1000)	Bladder Cancer (MRI/V/1000)
B	Bladder Cancer (CT/V/1000)	Lung Normal (XR/I/1341)	Urography (CT/V/1000)	Breast Cancer (MRI/V/1000)	Prostate Cancer (MRI/V/1000)	Uterus Carcinoma (CT/V/1000)	Rectum Adenocarcinoma (MRI/V/1000)	Ovarian Cystadenocarcinoma (CT/V/1000)	Liver Carcinoma (MRI/V/1000)	Liver Carcinoma (CT/V/1000)
C	Kidney & Renal Cell Carcinoma (MRI/V/1000)	Kidney & Renal Cell Carcinoma (CT/V/1000)	Normal Fundus (FC/I/321)	Body Topogram (PET/I/1000)	Pancreas Adenocarcinoma (MRI/V/1000)	Pancreas Adenocarcinoma (CT/V/1000)	Stomach Adenocarcinoma (CT/V/1000)	Colon Cancer (CT/V/1000)	Esophageal Carcinoma (CT/V/1000)	Facial Acne/Allergies (VLC/I/974)
D	Hand & Foot Allergies (VLC/I/476)	Legs & Arms Allergies (VLC/I/144)	Other Skin Lesions (VLC/I/900)	Breast Cancer (XR/I/322)	Brain Tumor (MRI/V/1000)	Fractured Bones (XR/I/338)	Neck Nerve Structure (US/I/1000)	Esophageal Candidiasis (ES/V/838)	Esophageal Papillomatosis (ES/V/544)	Hiatal Hernia (ES/V/1296)
E	Polyps (ES/V/500)	Normal z-line (ES/V/500)	Dyed Resection Margins (ES/I/500)	Dyed Lifted Polyps (ES/I/500)	Shoulder Implants (XR/I/597)	Lung COVID-19 Infection (CT/V/3254)	Lung COVID-19 Infection (XR/I/3296)	Cardiac (MRI/V/1000)	Tuberculosis (XR/I/394)	Lung Viral Pneumonia (XR/I/1345)
F	Gland in Colon (MS/I/165)	Breast Tumor (Benign) (MS/I/2480)	Breast Tumor (Malignant) (MS/I/5429)	Malaria in Blood Smears (MS/I/1328)	Abnormal Fundus (FC/I/850)	Elbow (XR/I/5552)	Fingers (XR/I/5567)	Forearm (XR/I/2126)	Shoulder (CT/V/910)	Hand (XR/I/6003)
G	Humerus (XR/I/1560)	Shoulder (XR/I/8942)	Wrist (XR/I/10411)	Retinal Disease (CNV) (OCT/I/8000)	Retinal Disease (DME) (OCT/I/8000)	Retinal Drusen Disease (OCT/I/8616)	Retinal Normal (OCT/I/8000)	GI Cancer (MSI) (MS/I/7503)	GI Cancer (MSS) (MS/I/11727)	Ankle (CT/V/150)

variable-length in terms of the number of frames as mentioned in Table I. In the data preprocessing step, all the data samples were resized to a fixed spatial dimension of 224×224 (as the fixed input layer size of our proposed MDF-RN model). Additionally, online data augmentation was performed to resolve the class imbalance problem during the training process.

The MATLAB (R2019a) coding framework (including the DL toolbox) was used for model development and simulation using a desktop computer with an Intel Core i7 CPU, 16 GB RAM, NVIDIA GeForce graphics processing unit (GPU) (GTX 1070), and Windows 10 operating system. In our optimization scheme, a stochastic gradient descent optimizer [42] with a learning rate of 0.001 was used for training both networks. Various existing studies [43], [44], [45], [46] related to medical image analysis considered such a small learning rate value of 0.001 for the optimal training of their proposed models. Generally, in case of a small value of learning rate, a minimum can be reached eventually; however, it will require many epochs to get there [47]. Nevertheless, when the learning rate is relatively large, the training loss drops sharply at first, fluctuates above the minimum, and never decays to the minimum [47]. Therefore, we chose a small value of learning rate (as reported in various existing studies [43], [44], [45], [46]) to achieve optimal convergence of the proposed model. We selected mini-batch sizes of 10 and 100 for training the MDF-RN and STB-CN models, respectively. These optimal values for mini-batch sizes were experimentally determined based on the maximum convergence of training accuracies, as shown in Fig. 4. In addition, because of the memory size limitation of GPU, it was not possible to select further higher values (i.e., >10 and >100) of mini-batch sizes. For the other hyperparameters, we used the default parametric scheme provided by MATLAB (R2019a). In all experiments, two-fold cross-validation was performed using 40% (60432 data samples), 10% (15108 data samples), and 50% (75554 data samples) of the whole dataset for training, validation, and testing, respectively. Two-fold cross-validation includes a smaller number of training data and shows lower accuracy compared to ten-fold cross-validation as reported in [48]. In addition, various existing studies related to medical image analysis [49], [50], [51], [52] also considered two-fold cross-validation to validate their proposed methods. Therefore, we considered two-fold cross-validation in all the experiments to achieve higher accuracy using a smaller number of training data. In most classes, different patient datasets were chosen for training, validation, and testing. Fig. 4 shows the training/validation losses and accuracies of both networks according to the increment of epoch. The convergence of training curves validates



(a) (b)

Fig. 4. Training/validation accuracies and losses of the proposed (a) MDF-RN model and (b) STB-CN model.

TABLE II
QUANTITATIVE RESULTS OF OUR PROPOSED MDF-RN+STB-CN (BEST MODEL) ALONG WITH THE PERFORMANCE OF MDF-RN (OUR SECOND-BEST PROPOSED MODEL) AND RN (BASELINE MODEL)

Model	ACC	F1	PRE	REC	AUC
RN [24]	86.56	84.21	85.92	82.57	90.55
(Image-based Model)	(1.07)	(1.48)	(0.78)	(2.13)	(0.40)
MDF-RN	89.10	85.92	87.15	84.73	92.30
(Image-based Model)	(0.46)	(0.75)	(0.06)	(1.53)	(0.15)
MDF-RN+STB-CN	**89.83**	**88.10**	**89.46**	**86.78**	**93.66**
(Sequence-based Model)	**(0.22)**	**(0.14)**	**(0.54)**	**(0.23)**	**(0.48)**

The best results are presented in boldface.

the sufficient training of both networks with training data. In addition, validation curves further confirmed that our models were not overfitted with training data. Numerous medical image classification studies measure the effectiveness of their proposed model with the following top-5 performance evaluation metrics:

1) average accuracy (ACC),
2) F1-score (F1),
3) precision (PRE),
4) recall (REC), and
5) area under the curve (AUC).

Therefore, we measured the effectiveness of the proposed model compared to various baseline methods using these five performance evaluation metrics as key indicators.

B. Testing Results (Ablation Studies)

We proposed a cascade of two networks for the classification of both 2-D/3-D imaging data related to the medical domain. Table II presents the quantitative results of our proposed MDF-RN+STB-CN, along with the performance of MDF-RN (our second-best proposed model) and RN (baseline model) as an ablation study. The results in Table II primarily highlight the contribution of multilevel feature fusion using the proposed DF-blocks (MDF-RN versus RN) and second-stage STB-CN model (MDF-RN+STB-CN versus MDF-RN) in terms of performance gains. The regularity of this comparative analysis (see Table II) is defined as follows.

1) In our first comparison (MDF-RN versus RN), we disregarded the 3-D anatomical dependencies of 3-D imaging data by considering the whole data as 2-D imaging data. For the data conversion from 3-D volume to 2-D slices/images, the same class label of each 3-D data sample was considered for its corresponding slices. Consequently, the whole 3-D data samples were converted into 2-D imaging data.
2) In our second comparison (MDF-RN+STB-CN versus MDF-RN), we further highlighted the contribution of 3-D anatomical dependencies of 3-D imaging data by introducing our second-stage STB-CN for the additional feature extraction in case of 3-D volumetric data samples (as explained in Section III-C).

The first proposed MDF-RN model (comprising SR-, PR-, and DF-blocks) outperforms the RN model (comprising only SR- and PR-blocks) with average gains of 2.54%, 1.71%, 1.23%, 2.16%, and 1.75% in terms of ACC, F1, PRE, REC, and AUC, respectively. Subsequently, the addition of the STB-CN model further improved the performance of the MDF-RN model, with average gains of 0.73%, 2.18%, 2.31%, 2.05%, and 1.36% in terms of ACC, F1, PRE, REC, and AUC, respectively. Ultimately, our proposed MDF-RN+STB-CN model significantly outperformed the RN (baseline model), with average gains of 3.27%, 3.89%, 3.54%, 4.21%, and 3.11% in terms of ACC, F1, PRE, REC, and AUC, respectively. In a t-test analysis, our first proposed MDF-RN achieved an average p-value of 0.001 $(p < 0.01)$, and the final MDF-RN+STB-CN attained a p-value of 0.00003 $(p < 0.01)$ compared to RN (baseline model). These lower p-values $(p < 0.01)$ imply that both networks significantly outperformed the baseline model at a 99% confidence score.

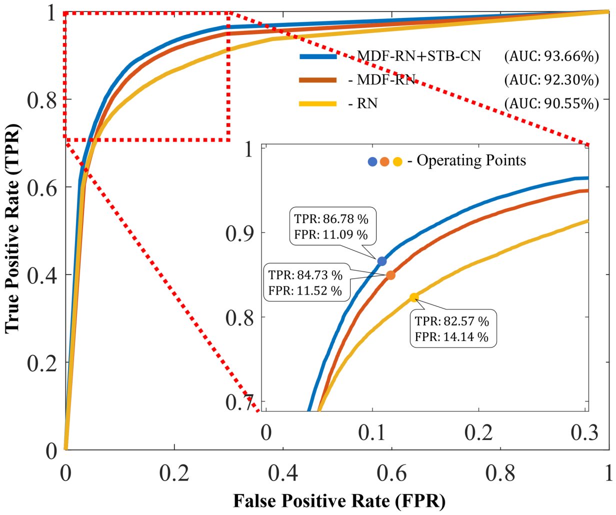


Fig. 5. Receiver operating characteristic curves of MDF-RN+STB-CN (our proposed best model), MDF-RN (our proposed second-best model), and RN (baseline model).

Moreover, Fig. 5 further highlights the performance gain of our proposed MDF-RN+STB-CN (best model) compared to MDF-RN (proposed second-best model) and RN (baseline model) as receiver operator characteristic (ROC) curves. In detail, in case of image-based models (i.e., MDF-RN and baseline RN), we evaluated slice/frame-level classification performance of 3-D imaging data by considering only spatial features. While, in case of our sequence-based model (i.e., MDF-RN+STB-CN), we evaluated block-level classification performance of 3-D imaging data by exploiting both spatial and 3-D anatomical features. Each ROC curve (see Fig. 5) indicates a tradeoff between the true positive rate (TPR) and false positive rate (FPR) at different classification thresholds, ranging from 0 to 1 in 0.01 increments. We attain the optimal validation performance of each model at a particular classification threshold (labeled as operating points in Fig. 5). The values of the optimal operating points for MDF-RN+STB-CN (our proposed best model), MDF-RN (our proposed second-best model), and RN (baseline model) were 0.41, 0.44, and 0.46, respectively.

TABLE III
PROGRESSIVE PERFORMANCE GAINS OF THE PROPOSED MDF-RN AND MDF-RN+STB-CN MODELS BASED ON MULTILEVEL FEATURE FUSION

Model	#Features	ACC (Std)	F1 (Std)	PRE (Std)	REC (Std)	AUC (Std)
MDF-RN (Image-based Model)	f_1	79.17 (2.35)	76.52 (0.27)	78.86 (0.31)	74.32 (0.23)	79.74 (0.18)
	$f_1 - f_2$	85.79 (0.47)	82.82 (1.15)	84.42 (1.19)	81.28 (1.12)	87.89 (0.34)
	$f_1 - f_3$	87.38 (1.38)	85.28 (0.00)	86.76 (0.56)	83.86 (0.52)	91.45 (0.92)
	$f_1 - f_4$	**89.10** **(0.46)**	**85.92** **(0.75)**	**87.15** **(0.06)**	**84.73** **(1.53)**	**92.30** **(0.15)**
MDF-RN+STB-CN (Sequence-based Model)	f_1	80.17 (2.31)	78.51 (0.73)	80.23 (1.24)	76.86 (0.24)	81.73 (1.65)
	$f_1 - f_2$	86.85 (0.67)	82.77 (1.98)	84.13 (1.72)	81.46 (2.24)	88.43 (2.27)
	$f_1 - f_3$	87.61 (0.92)	86.28 (0.94)	88.22 (0.06)	84.43 (1.86)	90.76 (1.45)
	$f_1 - f_4$	**89.83** **(0.22)**	**88.10** **(0.14)**	**89.46** **(0.54)**	**86.78** **(0.23)**	**93.66** **(0.48)**

The best results are presented in boldface.

Compared with RN (baseline model), our best model significantly reduced the FPR from 14.14% to 11.09% with an average gain of 3.05% [14.14%–11.09%] and increased the TPR from 82.57% to 86.78% with an average gain of 4.21% [82.57%–86.78%]. Subsequently, our second-best model also significantly reduced the FPR from 14.14% to 11.52% with an average gain of 2.62% [14.14%–11.52%] and increased the TPR from 82.57% to 84.73% with an average gain of 2.16% [82.57%–84.73%] in comparison with RN. Consequently, our proposed MDF-RN+STB-CN accurately classified a total of 3181 more data samples compared to RN (baseline model).

A detailed ablation study was further conducted to demonstrate the significance of multilevel feature fusion in the proposed image-based model (MDF-RN). Successive feature-level performance was calculated to highlight the contribution of multiscale low-, intermediate-, and high-level semantic features (i.e., $\boldsymbol{f}_1 - \boldsymbol{f}_4$). Subsequently, the same ablation study was conducted for our proposed sequence-based model (MDF-RN+STB-CN). Table III lists the successive feature-level performances of both networks. It can be observed (see Table III) that the fusion of multilevel features (i.e., $\boldsymbol{f}_1 - \boldsymbol{f}_4$) results in a progressive gain, and finally, a high-performance MDF-RN model was attained based on multilevel feature fusion. Similarly, the proposed MDF-RN+STB-CN also showed progressive results with the fusion of multilevel features in the MDF-RN model.

The initial weights of the different residual blocks in our MDF-RN model were obtained from a pretrained RN [24] through a transfer learning approach. Therefore, we also trained our models from scratch to highlight the importance of transfer learning in terms of quantitative performance. For the MDF-RN model, the results indicate that transfer learning compared with training from scratch exhibits superior results with average gains of 20.11%, 18.29%, 16.77%, 19.60%, and 23.02% in terms of ACC, F1, PRE, REC, and AUC, respectively. Similarly, we observed significant performance gains of 18.79%, 16.45%, 15.51%, 17.26%, and 18.86% in terms of ACC, F1, PRE, REC, and AUC, respectively, for MDF-RN+STB-CN. In addition to a large number of data samples in our training dataset, a significant impact of transfer learning can be observed (see Table IV) in developing our image-based and sequence-based classification models.

TABLE IV
COMPARATIVE RESULTS OF THE PROPOSED MDF-RN AND MDF-RN+STB-CN MODELS WITH AND WITHOUT PERFORMING TRANSFER LEARNING. ("T.L: TRANSFER LEARNING")

Model	T.L	ACC (Std)	F1 (Std)	PRE (Std)	REC (Std)	AUC (Std)
MDF-RN (Image-based Model)	×	68.99 (0.51)	67.63 (0.06)	70.38 (1.45)	65.13 (1.35)	69.28 (2.21)
	✓	**89.10** **(0.46)**	**85.92** **(0.75)**	**87.15** **(0.06)**	**84.73** **(1.53)**	**92.30** **(0.15)**
MDF-RN+STB-CN (Sequence-based Model)	×	71.04 (0.01)	71.65 (2.47)	73.95 (4.12)	69.52 (1.01)	74.80 (0.00)
	✓	**89.83** **(0.22)**	**88.10** **(0.14)**	**89.46** **(0.54)**	**86.78** **(0.23)**	**93.66** **(0.48)**

The best results are presented in boldface.

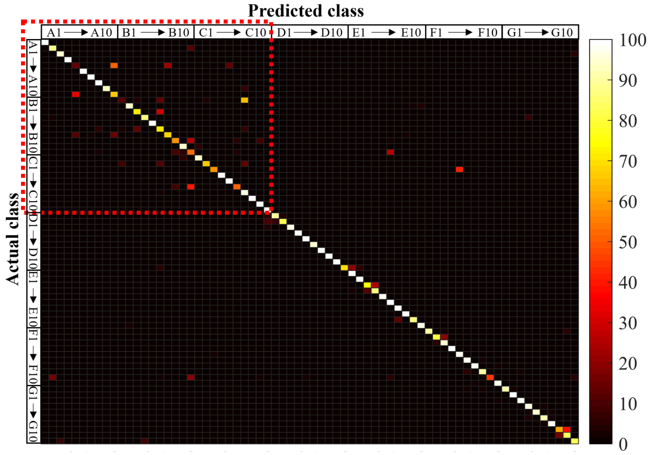


Fig. 6. Clustering of classification results of the final proposed model in terms of confusion matrix to highlight the performance degradation of each individual class as type I (false-positive) or type II (false-negative) errors.

In addition, Fig. 6 further presents the clustering of classification results of the final proposed model in terms of the confusion matrix. These results (see Fig. 6) show the individual performance of each class by visualizing the number of false predictions as type I (false-positive) or type II (false-negative) errors. It can be observed that the data samples of classes A1 to C10 show a higher number of false predictions (highlighted with red-box in Fig. 6) as compared to other classes. The high interclass similarities cause such performance degradation due to the following reasons: 1) overlapped body organs in different classes and 2) data samples of different imaging modalities (i.e., CT and MRI) presenting similar types of diseases in different classes. However, the overall performance of our proposed method is significantly improved as compared to other baseline methods (as shown in subsequent Section IV-C).

C. Comparison

This section provides a detailed comparison of our proposed MDF-RN+STB-CN (best model) and MDF-RN (second-best model) with several state-of-the-art image- and sequence-based

TABLE V
COMPARATIVE PERFORMANCE ANALYSIS OF OUR PROPOSED MDF-RN+STB-CN (BEST MODEL) AND MDF-RN (SECOND-BEST MODEL) WITH THE VARIOUS STATE-OF-THE-ART METHODS

Network Type	Study	ACC (Std)	F1 (Std)	PRE (Std)	REC (Std)	AUC (Std)
Image-based Models	Adnan *et al.* 2017 [7]	77.02 (0.36)	70.33 (0.06)	73.39 (0.49)	67.51 (0.30)	78.24 (0.37)
	Falconi *et al.* 2018 [8]	87.26 (0.47)	84.40 (1.30)	85.88 (1.10)	82.97 (1.48)	91.55 (1.42)
	Owais *et al.* 2019 [9]	86.29 (0.23)	82.13 (0.05)	83.93 (0.84)	80.41 (0.67)	89.43 (0.08)
	Kaur *et al.* 2019 [10]	86.94 (1.12)	83.72 (2.09)	85.33 (1.84)	82.17 (2.33)	91.36 (0.98)
	Ashraf *et al.* 2020 [11]	87.50 (0.42)	84.09 (0.67)	85.40 (0.23)	82.82 (1.08)	90.66 (0.79)
	Akpinar *et al.* 2020 [12]	83.51 (0.21)	78.27 (1.61)	82.20 (2.06)	74.70 (1.22)	84.86 (0.97)
	Igarashi *et al.* 2020 [13]	81.38 (0.25)	78.96 (2.14)	81.38 (0.25)	76.73 (3.80)	85.91 (3.65)
	Aloyayri *et al.* 2020 [14]	86.56 (1.07)	84.21 (1.48)	85.92 (0.78)	82.57 (2.13)	90.55 (0.40)
	Souid *et al.* 2021 [15]	87.29 (0.04)	84.06 (0.59)	84.62 (0.96)	83.51 (0.23)	90.87 (0.51)
	Jasil *et al.* 2021 [16]	88.46 (0.62)	85.17 (1.05)	86.30 (0.78)	84.07 (1.30)	91.91 (0.23)
	Çakmak *et al.* 2021 [17]	87.97 (0.37)	84.56 (0.74)	85.38 (0.70)	83.76 (0.77)	90.84 (0.02)
	Gambhir *et al.* 2021 [18]	86.25 (0.21)	83.28 (0.02)	84.18 (0.31)	82.39 (0.27)	90.07 (0.25)
	Proposed (MDF-RN)	**89.10 (0.46)**	**85.92 (0.75)**	**87.15 (0.06)**	**84.73 (1.53)**	**92.30 (0.15)**
Sequence-based Models	Shahzadi *et al.* 2018 [19]	87.33 (1.04)	84.85 (1.77)	86.80 (1.66)	83.01 (1.87)	90.91 (1.38)
	Srinivasu *et al.* 2021 [20]	87.85 (0.01)	85.25 (0.36)	85.77 (0.56)	84.74 (0.16)	92.04 (0.50)
	Ebrahimi *et al.* 2021 [21]	86.68 (0.71)	84.59 (1.17)	86.27 (0.45)	82.99 (1.85)	91.82 (0.11)
	Proposed (MDF-RN+STB-CN)	**89.83 (0.22)**	**88.10 (0.14)**	**89.46 (0.54)**	**86.78 (0.23)**	**93.66 (0.48)**

The best results are presented in boldface.

CAD methods. This is the first study related to the classification of large-scale 2-D/3-D imaging data and no standard benchmarks are given in the literature with the selected dataset. Therefore, we explored the existing literature related to medical image classification and selected fifteen different methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] closely related to our work for this comparison. In detail, all these competitor methods utilized the strength of transfer learning employing existing pretrained CNN models [22], [23], [24], [25], [26], [27], [28], [29], [30], [31] in developing their CAD solutions. All these studies cover a vast scope of 1) disease-specific, 2) modality-specific, 3) multimodality-based, and 4) multi-disease-based CAD solutions. In addition, the source codes of all these methods are also publicly available for a fair comparison. Therefore, we selected these methods for comparative analysis with our proposed solution. To make a fair comparison and provide standard benchmarks, we evaluated the performance of these existing methods with our selected dataset. Table V presents the comparative results of our proposed models in comparison with 15 different state-of-the-art methods.

The regularity of this comparative study is defined as follows: Initially, we compared the performance of our first image-based model with various image-based classification methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] using our selected dataset. All these comparative models are labeled as image-based models in Table V. In this comparison, we ignored the 3-D anatomical information of 3-D imaging data by considering the whole data as 2-D imaging data (as explained in Section IV-B). Subsequently, we further compared the performance of our final sequence-based model with three different sequence-based classification methods [19], [20], [21] under the same experimental setting. These comparative models are labeled as sequence-based models in Table V. In this comparison, we also considered the contribution of 3-D anatomical dependencies of 3-D imaging data (as explained in Section III-C).

Compared to various image-based classification methods, our proposed 2-D-CNN model (MDF-RN) shows better results. Jasil et al. [16] proposed a CAD method based on DenseNet201 [16], which showed comparable results and ranked it as the second-best method among other image-based models. However, our proposed MDF-RN model outperformed [16] in terms of quantitative as well as computational performance. In detail, our MDF-RN model outperformed DenseNet201 (used by Jasil et al. [16]) with average gains of 0.64%, 0.75%, 0.85%, 0.66%, and 0.39% in terms of ACC, F1, PRE, REC, and AUC, respectively. In a t-test analysis, our MDF-RN model outperformed DenseNet201 at a 99% confidence score by reaching an average p-value of 0.001 ($p < 0.01$). In addition, the average inference time (class prediction time of one data sample) of our MDF-RN model was approximately 50% lower than that of Jasil et al. [16]. To be specific, our MDF-RN took approximately 13.26 ms, whereas DesneNet201 (used by Jasil et al. [16]) required approximately 25.88 ms for one image. The average inference time was evaluated using the same experimental setup described in Section IV. Moreover, our final MDF-RN+STB-CN model gave better results than MDF-RN and further outperformed the second-best image-based method (Jasil et al. [16]) with average gains of 1.37%, 2.93%, 3.16%, 2.71%, and 1.75% in terms of ACC, F1, PRE, REC, and AUC, respectively. Subsequently, a t-test analysis also showed the superior performance of our final MDF-RN+STB-CN model compared to [16] at a 99% confidence score by reaching an average p-value of 0.001 ($p < 0.01$).

In the context of volumetric data classification, three different sequence-based models were proposed in the literature [19], [20], [21] using pretrained 2D-CNNs [10], [15], [22] as backbone networks. The performance of these methods [19], [20], [21] was also evaluated for comparison with our proposed MDF-RN+STB-CN (best model) under the same experimental setup. Srinivasu et al. [20] proposed a method based on MobileNetV2+LSTM [15], [20] ranked as second-best

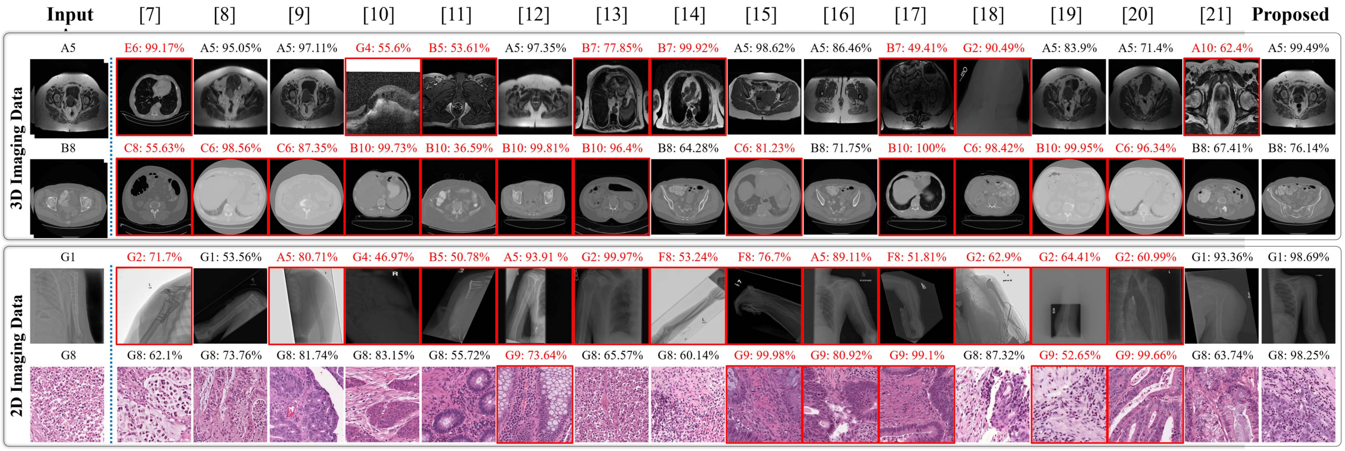


Fig. 7. Qualitative classification and content-based medical image and sequence retrieval (CBMISR) performance of our proposed and the various state-of-the-art methods (red box: False predictions).

among the other two methods [19], [21]. However, our final MDF-RN+STB-CN model outperformed the method of Srinivasu et al. [20], with average gains of 1.98%, 2.85%, 3.69%, 2.04%, and 1.62% in terms of ACC, F1, PRE, REC, and AUC, respectively. A t-test analysis also highlights the superiority of our final model over [20] at a 99% confidence score by reaching an average p-value of 0.001 $(p < 0.01)$. The proposed pipeline mainly includes a novel 2-D-CNN architecture (named MDF-RN) that leverages multiscale dilated convolution and a concept of multilevel feature fusion in a mutually beneficial manner to achieve state-of-the-art performance. Additionally, the second subnetwork (STB-CN) further aggregates the overall performance by exploiting 3-D anatomical dependencies in case of 3-D imaging data and results in an additional performance gain. Consequently, the proposed model offers better results compared to various existing methods (see Table V).

Fig. 7 further presents the qualitative classification and CBMISR results of our method compared with all the testing baseline methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. We provided the predicted class label, probability score, and best-matched data sample for each method. We retrieved the best-matched case from the testing database using the predicted class label as explained in [7]. The key objective of this qualitative analysis is to highlight the predictive confidence score and CBMISR performance of our method in comparison with all the baseline methods. It can be observed (see Fig. 7) that our method attains the highest confidence score, particularly in low-contrast data samples (class B8 and G1), and shows a significant resemblance of best-matched cases with the input query data samples.

V. Discussion

In the case of 3-D imaging data, 2-D-CNNs (image-based models) only extract the spatial features from each slice/frame for class prediction. By contrast, 3-D-CNNs extract both spatial and 3-D anatomical features from the entire 3-D sequence and make a class prediction. In the first scenario, 2-D-CNNs neglect 3-D anatomical information for the sequence of 2-D slices, which can result in performance degradation. In the second scenario, 3-D-CNNs include several trainable parameters and require high computation power for training. In addition, 3-D-CNNs are restricted to process a fixed-length volumetric data and may cause performance degradation in case of variable-length data in a real-world scenario. To address these issues, a sequence-based classification framework is proposed for the accurate classification of both 2-D/3-D imaging data. Initially, our first proposed image-based model (MDF-RN) extracts a set of n multilevel spatial feature vectors (i.e., $\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \ldots, \boldsymbol{f}^n$) from a given sequence of slices/frames (i.e., $\boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{F}^3, \ldots, \boldsymbol{F}^n$). Subsequently, the second-stage STB-CN model further exploits 3-D anatomical features from a set of spatial feature vectors and performs the final class prediction. In the case of 3-D imaging data, the use of LSTM models with 2-D-CNN makes it more expedient than 3-D-CNNs in terms of computational complexity. In addition, our proposed sequence-based model leverages transfer learning in volumetric data analysis without influencing the overall training parameters. It can also classify variable-length sequences. Our comprehensive ablation study proves the significance of multilevel feature fusion (see Table III) and transfer learning (see Table IV) in developing the proposed sequence-based classification framework for the efficient classification of multimodal 2-D/3-D imaging data.

In the context of image classification, our first MDF-RN model outperformed various state-of-the-art 2-D-CNNs (image-based models) (see Table V). In addition to quantitative performance gains, the average inference time of our MDF-RN model was approximately 50% lower than that of the second-best image-based method (Jasil et al. [16]). We also evaluated the performance of various disease and modality-specific models with our selected datasets and observed that our proposed model significantly (t-test:$p < 0.01$) outperformed these models in case of large cohort (see Table V). For example, our final MDF-RN+STB-CN model outperforms the second-best modality-specific method of Srinivasu et al. [20] with average gains of 1.98%, 2.85%, 3.69%, 2.04%, and 1.62% in terms of ACC,

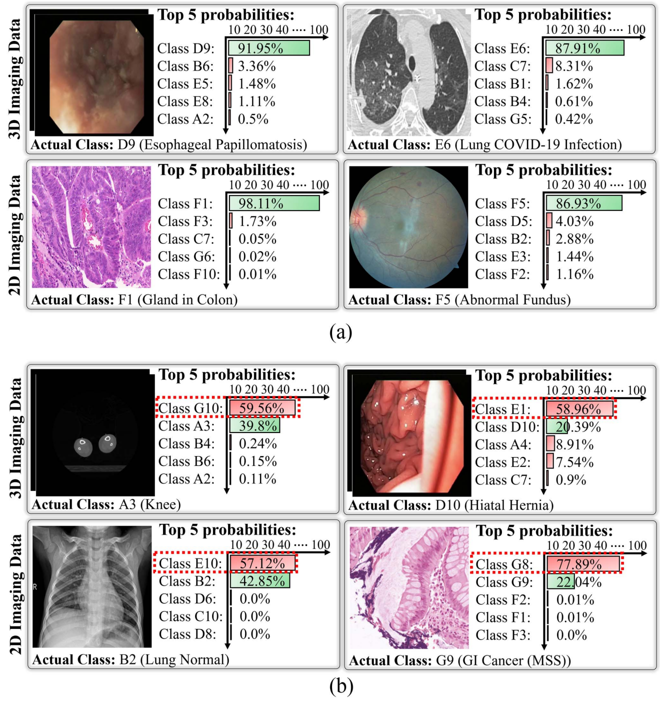


Fig. 8. Illustration of (a) correctly classified and (b) incorrectly classified data samples including top five predicted probabilities.

F1, PRE, REC, and AUC, respectively. Most of the existing methods [18], [19], [20], [21] exploit only high-level semantic features to make a classification decision. Our network design mainly leverages multiscale dilated convolutions (DF-blocks) and multilevel feature fusion in a mutually beneficial manner and finally achieves superior classification results.

Fig. 8 shows a few examples of correctly and incorrectly classified data samples (including both 2-D/3-D imaging data) using our proposed classification framework. To visualize the discriminative capability of our model, we additionally included the top five predicted probabilities along with each data sample. In most of the correctly classified data samples [see Fig. 8(a)], the best probability score was significantly higher (>85%) than the other classes, which highlights the distinctive characteristics of the proposed method. However, a few incorrect predictions [see Fig. 8(b)] may also turn out because of the existence of analogous shapes and texture patterns in different classes. For example, the data samples of class B2 (lung normal) and E10 (lung viral pneumonia) show high inter-class similarities in terms of shapes and texture patterns. Therefore, the input sample of class B2 (lung normal) was incorrectly classified as E10 (lung viral pneumonia) by achieving the best probability score of 57.12%. Fig. 8(b) visualizes a few more examples of such incorrectly classified data samples. Regardless of high inter-class similarities, the poor annotation of data samples can also lead a deep classification model toward false predictions. However, visual assessment can assist medical experts in performing cross-validation of all predicted results.

Despite the considerable gains of the proposed classification framework, there are a few limitations that may influence the overall performance of our classification framework in a real-world setting. The main concern is the issue of generalizability, particularly with respect to those classes with a limited number of data samples. Second, real-world data can show high intra-class variance due to various types of imaging modalities and may influence the prediction results. However, these limitations can be resolved by including a large collection of diversified and well-annotated datasets.

VI. Conclusion

This article aims to develop a deep classification model with the capability to classify multimodal and multiclass medical data, including both 2-D and 3-D imaging data. In particular, a sequence-based deep classification framework (named MDF-RN+STB-CN) is proposed, which mainly leverages transfer learning in the case of volumetric data analysis without influencing the overall training parameters. This is the first study to offer a pretrained classification model in the medical domain based on a large-scale multimodal dataset (including a total of 151 095 data samples related to 70 different classes). Finally, the experimental results exhibited promising performance values of 89.83%, 88.10%, 89.46%, 86.78%, and 93.66% in terms of the average accuracy, F1-score, precision, recall, and area under the curve, respectively, and outperformed various state-of-the-art methods. In a future study, we will explore more heterogeneous datasets and intend to resolve generality issues thoroughly. In addition, the proposed model provides new grounds for future research related to MDS-TL. The strengths of MDS-TL can be further investigated and additional performance improvements can be achieved in numerous medical diagnostic applications.

References

[1] A. Masood et al., “Automated decision support system for lung cancer detection and classification via enhanced RFCN with multilayer fusion RPN,” *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7791–7801, Dec. 2020.

[2] A. Kumar, V. Purohit, V. Bharti, R. Singh, and S. K. Singh, “MediSecFed: Private and secure medical image classification in the presence of malicious clients,” *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5648–5657, Aug. 2022.

[3] O. P. Singh, A. K. Singh, and H. Zhou, “Multimodal fusion-based image hiding algorithm for secure healthcare system,” *IEEE Intell. Syst.*, Sep. 2022.

[4] K. N. Singh, O. P. Singh, A. K. Singh, and A. K. Agrawal, “Watmif: Multimodal medical image fusion-based watermarking for telehealth applications,” *Cogn. Comput.*, pp. 1–17, 2022.

[5] O. P. Singh, A. K. Singh, A. K. Agrawal, and H. Zhou, “SecDH: Security of COVID-19 images based on data hiding with PCA,” *Comput. Commun.*, vol. 191, pp. 368–377, 2022.

[6] M. Abdar et al., “Hercules: Deep hierarchical attentive multi-level fusion model with uncertainty quantification for medical image classification,” *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 274–285, Jan. 2023.

[7] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, “Medical image retrieval using deep convolutional neural network,” *Neurocomputing*, vol. 266, pp. 8–20, 2017.

[8] L. Falconí, M. Pérez, W. Aguilar, and A. Conci, “Transfer learning and fine tuning in mammogram BI-RADS classification,” in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst.*, 2020, pp. 475–480.

[9] M. Owais, M. Arsalan, J. Choi, and K. R. Park, “Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence,” *J. Clin. Med.*, vol. 8, no. 4, 2019, Art. no. 462.

[10] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," in *Proc. IEEE Int. Conf. Inf. Technol.*, 2019, pp. 94–98.

[11] R. Ashraf et al., "Deep convolution neural network for big data medical image classification," *IEEE Access*, vol. 8, pp. 105659–105670, 2020.

[12] K. N. Akpinar, S. Genc, and S. Karagol, "Chest X-ray abnormality detection based on SqueezeNet," in *Proc. IEEE Int. Conf. Elect., Commun., Comput. Eng.*, 2020, pp. 1–5.

[13] S. Igarashi, Y. Sasaki, T. Mikami, H. Sakuraba, and S. Fukuda, "Anatomical classification of upper gastrointestinal organs under various image capture conditions using AlexNet," *Comput. Biol. Med.*, vol. 124, 2020, Art. no. 103950.

[14] A. Aloyayri and A. Krzyżak, "Breast cancer classification from histopathological images using transfer learning and deep neural networks," in *Proc. Int. Conf. Artif. Intell. Soft Comput.*, 2020, pp. 491–502.

[15] A. Souid, N. Sakli, and H. Sakli, "Classification and predictions of lung diseases from chest x-rays using mobilenet v2," *Appl. Sci.-Basel*, vol. 11, no. 6, 2021, Art. no. 2751.

[16] S. G. Jasil and V. Ulagamuthalvi, "Skin lesion classification using pretrained DenseNet201 deep neural network," in *Proc. IEEE 3rd Int. Conf. Signal Process. Commun.*, 2021, pp. 393–396.

[17] M. Çakmak and M. E. Tenekecı, "Melanoma detection from dermoscopy images using Nasnet mobile with transfer learning," in *Proc. IEEE 29th Signal Process. Commun. Appl. Conf.*, 2021, pp. 1–4.

[18] R. Gambhir, S. Bhardwaj, A. Kumar, and R. Agarwal, "Severity classification of diabetic retinopathy using ShuffleNet," in *Proc. Int. Conf. Intell. Technol.*, 2021, pp. 1–5.

[19] I. Shahzadi, T. B. Tang, F. Meriadeau, and A. Quyyum, "CNN-LSTM: Cascaded framework for brain tumour classification," in *Proc. IEEE-EMBS Conf. Biomed. Eng. Sci.*, 2018, pp. 633–637.

[20] P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM," *Sensors*, vol. 21, no. 8, 2021, Art. no. 2852.

[21] A. Ebrahimi, S. Luo, and R. Chiong, "Deep sequence modelling for Alzheimer's disease detection using MRI," *Comput. Biol. Med.*, vol. 134, 2021, Art. no. 104537.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representation*, 2015, pp. 1–14.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2016, pp. 770–778.

[25] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2015, pp. 1–9.

[26] F. N. Iandola et al., "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 MB model size," 2016, *arXiv:1602.07360*.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2018, pp. 4510–4520.

[28] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2017, pp. 4700–4708.

[29] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2018, pp. 8697–8710.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2018, pp. 6848–6856.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[32] M. Owais, H. S. Yoon, T. Mahmood, A. Haider, H. Sultan, and K. R. Park, "Light-weighted ensemble network with multilevel activation visualization for robust diagnosis of COVID19 pneumonia from large-scale chest radiographic database," *Appl. Soft Comput.*, vol. 108, 2021, Art. no. 107490.

[33] K. Pogorelov et al., "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 164–169.

[34] G. Urban, S. Porhemmat, M. Stark, B. Feeley, K. Okada, and P. Baldi, "Classifying shoulder implants in X-ray images using deep learning," *Comp. Struct. Biotechnol. J.*, vol. 18, pp. 967–972, 2020.

[35] K. Sirinukunwattana et al., "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, 2017.

[36] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2366–2378, Nov. 2015.

[37] A. Aksac, D. J. Demetrick, T. Ozyer, and R. Alhajj, "BreCaHAD: A dataset for breast cancer histopathological annotation and diagnosis," *BMC Res. Notes*, vol. 12, 2019, Art. no. 82.

[38] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nature Methods*, vol. 9, no. 7, pp. 637–637, 2012.

[39] P. Rajpurkar et al., "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," 2017, *arXiv:1712.06957*.

[40] J. N. Kather, "Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples," *ZENODO*, 2019. Accessed: Feb. 01, 2019. [Online]. Available: http://doi.org/10.5281/zenodo.2530835

[41] D. J. Matuszewski and I. M. Sintorn, "TEM virus images: Benchmark dataset and deep learning classification," *Comput. Methods Prog. Biomed.*, vol. 209, 2021, Art. no. 106318.

[42] S. I. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, 1993, pp. 185–196.

[43] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT Exp.*, vol. 6, pp. 312–315, 2020.

[44] T. Mahmood et al., "Accurate segmentation of nuclear regions with multi-organ histopathology images using artificial intelligence for cancer diagnosis in personalized medicine," *J. Pers. Med.*, vol. 11, 2021, Art. no. 515.

[45] Y. J. Suh, J. Jung, and B. J. Cho, "Automated breast cancer detection in digital mammograms of various densities via deep learning," *J. Pers. Med.*, vol. 10, 2020, Art. no. 211.

[46] Y. Qiu, Y. Liu, S. Li, and J. Xu, "Miniseg: An extremely minimum network for efficient COVID-19 segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, pp. 4846–4854, 2021.

[47] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Adv. Neural. Inf. Process. Syst.*, vol. 26, pp. 315–323, 2013.

[48] S. Nanglia, M. Ahmad, F. A. Khan, and N. Z. Jhanjhi, "An enhanced predictive heterogeneous ensemble model for breast cancer prediction," *Biomed. Signal Process. Control*, vol. 72, 2022, Art. no. 103279.

[49] M. S. Bispo et al., "Computer tomographic differential diagnosis of ameloblastoma and odontogenic keratocyst: Classification using a convolutional neural network," *Dentomaxillofacial Radiol.*, vol. 50, 2021, Art. no. 20210002.

[50] Y. Zhang, P. Yang, and V. Lanfranchi, "Tensor multi-task learning for predicting alzheimer's disease progression using MRI data with spatio-temporal similarity measurement," in *Proc. IEEE 19th Int. Conf. Ind. Informat.*, 2021, pp. 1–8.

[51] J. Yang et al., "Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning," *Comp. Struct. Biotechnol. J.*, vol. 20, pp. 333–342, 2022.

[52] S. Rani, S. Kumar, D. Ghai, and K. M. V. V. Prasad, "Automatic detection of brain tumor from CT and MRI images using wireframe model and 3D alex-net," in *Proc. IEEE Int. Conf. Decis. Aid Sci. Appl.*, 2022, pp. 1132–1138.

Muhammad Owais received the B.S. and M.S. degrees in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2014 and 2016, respectively, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2022.

In 2022, he has worked as an assistant professor in the division of electronics and electrical engineering, Dongguk University, Seoul, South Korea. Since 2023, he has been with the Department of Electrical Engineering and Computer Science, Khalifa University, United Arab Emirates, where he is currently working as a Postdoctoral Fellow. His research interests include deep learning-based medical image analysis, biomedical 2-D/3-D imaging data processing, pattern recognition, and image recognition.

Se Woon Cho received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2017. He received the degree of combined course of M.S. and Ph.D. in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2022.

Then, he worked in Electronics and Telecommunications Research Institute, in 2022. His research interests include faster Region-based Convolutional Neural Network-based face detection in nighttime images, and Cycle-consistent Generative Adversarial Network-based image restoration and deep learning-based semantic segmentation with the frontal-viewing camera images of low light in vehicle. In addition, his research interests include multistage segmentation with ultrasound images for breast cancer diagnosis.

Kang Ryoung Park (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in electrical and computer engineering from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively.

He has been a Senior Researcher with LG Electronics Co., Ltd. from 2000 to 2003. From 2003 to 2008, he has been a Full-Time Lecturer and assistant professor with the Division of Digital Media Technology, Sangmyung University, Seoul, South Korea. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University since March 2013. He serves as a Section Editor in sensors and an Associate Editor in expert systems with applications. His research interests include deep learning-based image processing, image recognition, and biometrics.