IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 19, 2022, accepted March 26, 2022, date of publication April 5, 2022, date of current version April 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165077

DeepAIA: An Automatic Image Annotation Model Based on Generative Adversarial Networks and Transfer Learning

ABEER ALSHEHRI, MOUNIRA TAILEB, AND REEM ALOTAIBI
Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
Corresponding author: Abeer Alshehri (aalshehri1408@stu.kau.edu.sa)

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant FP-223-43.

ABSTRACT Automatic image annotation (AIA) has been adopted in different applications such as image retrieval and classification. Deep Learning is used in AIA to extract image features and then convert these features into text descriptions and labels. However, conventional AIA models that employ deep learning methods suffer from various shortcomings, such as poor annotation performance. This work proposes an AIA model based on convolutional neural networks (CNNs), generative adversarial networks (GANs), and transfer learning. GANs have attracted a lot of interest because of its ability to generate data without explicitly using probability density. Thus, it has proven its usefulness in image annotation and image augmentation. In this work, an Auxiliary classifier-GAN (ACGAN) has been used, where the discriminator predicts the class of an image rather than taking it as a given input; therefore, the stabilization of the training stage is ensured, and the generation of high-quality images is provided. Transfer learning is also used to enhance the performance of the classification. The proposed model outperforms the best state-of-the-art models in terms of MiAP, F-measure and error rate using ImageClef, ESPGame and IAPR-TC12 datasets.

INDEX TERMS Automatic image annotation, convolutional neural network, generative adversarial network, transfer learning.

I. INTRODUCTION

Along with the internet growth, the proliferation and the easy access to image capture devices such as smartphones, cameras, and drones, raised the number of images on the web tremendously. Moreover, various kinds of social media become popular, where it allows users to freely share image-based content such as Snapchat, Twitter, and Instagram. For instance, approximately, about 100 million images are uploaded daily to Instagram [1]. Most of the images contain valuable information for businesses and organizations from the consumer interest in a product or fashion events to patients' x-ray in the medical field. When a good image retrieval technique is used, it can facilitate a lot of real-life aspects and have a huge impact on many levels [2]. To facilitate retrieving images that satisfy the demand of users, it is important to label images correctly and precisely. To accomplish that, one of the best methods to manage large-scale image datasets is the Automatic Image Annotation (AIA) [3].

AIA aims at assigning annotations/labels to images that describe its visual content. AIA is divided into three approaches, which are text-based annotation, content-based annotation, and multimodal-based annotation. The first approach is the text-based image annotation [4], which indicates that the images are annotated based on the text assigned by users or the text that surrounds the images on the web pages. However, this process now is impractical due to the massive number of images on the web and inconsistent annotation could happen between two individuals on the same image. The second approach is the content-based image annotation, which concentrates on the low-level visual features such as color, shape, and texture in the process of annotating images. This approach suffers from a well-known issue called the semantic gap [5]. The third approach is the multimodal-based image annotation. This approach leverages

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

both of the earlier two approaches to solve the problem of the semantic gap. Multimodal-based models showed better results compared to the other approaches in image annotation.

Deep learning-based (DL) models have recently shown significant development in the AIA task, particularly with large-scale data. It can efficiently work with large datasets and learn feature representation automatically; thus, hand-crafted features became unnecessary. In the field of computer vision, one of the prominent methods is the convolutional neural network (CNN) [6], [7], where its structure is primarily based on multiple neural layers. The key to their success lies in the complex neural architecture that is capable of taking into account the global and the local characteristics of the input. Basically, CNN extracts rich hierarchical features from the image and produces probabilities of different possible labels.

Recently, another deep learning method that has had huge success in the field of computer vision, especially in the image/video generation field is called generative adversarial network (GAN) [8]. GANs architecture was inspired by a game called a two-player zero-sum game, where two players' cumulative is zero and the gain or loss of utility of each player is balanced. GAN basically consists of two neural networks namely the generative network and the discriminative network, where these two networks compete with each other, as shown in Figure 1. GANs succeed in different challenging tasks like generating animation, video frames, image generation, etc [9].

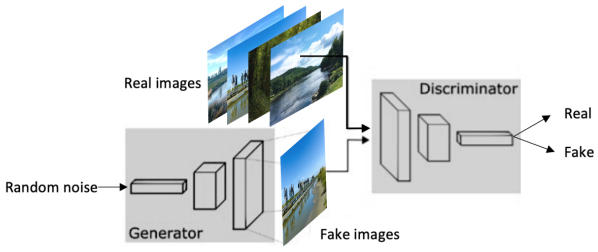

FIGURE 1. GAN architecture [10].

In this paper, an improved AIA model called DeepAIA is proposed. It relies on image augmentation using GAN and CNN. The CNN network extracts the visual features from images through a pre-trained deep CNN architecture. The CNN architecture is trained via a transfer learning technique to save up time and overcome the overfitting problem. Also, the GAN network is another main component of the DeepAIA model, which acts as a powerful technique of data augmentation to enhance the training of the CNN of the proposed model in an unsupervised manner. The GAN network tackles the problems resulting from overfitting and training with small-sized datasets in CNN architectures efficiently. In specific, this work adopts a pure data-augmentation method based on ACGAN to artificially synthesize new annotated images' samples and can solve the image augmentation problem of small-scale datasets. The proposed DeepAIA model has been evaluated using 4 large datasets. To the best of our knowledge, two of the datasets (ImageClef 2011 and ImageClef 2012) have never been used in the evaluation of deep learning-based models.

The core contributions of this paper are summarized in the followings:

- A comprehensive review of existing AIA models and highlighting the strengths and weaknesses of each.
- Proposing an AIA model named DeepAIA to automatically annotate the images with multiple labels. This model maintains the functionalities of both the pre-trained model using different architectures and image generation using ACGAN; therefore, the problems resulting from overfitting or training with small-sized datasets are addressed.
- Testing the proposed DeepAIA model on ImageClef 2011, ImageClef 2012 datasets, which is the first time that they are used to test AIA models with CNN architecture.
- Effectiveness verification of the proposed ACGAN-CNN model by comparing its performance with state-of-arts models on four datasets: ImageClef 2011, ImageClef 2012, ESPGame, and laprtc12.

The rest of this paper is organized as follows: Section II discusses the existing works conducted in the AIA field. Section III introduces the proposed DeepAIA model by explaining the details of its main stages. The implementation and experiment details are presented in Section IV. Section V discusses the experimental results. And Section VI concludes this work and presents some possible future directions.

II. RELATED WORK

With the assistance of the training set, AIA models are capable of learning the relationship between the visual content of the image and high-level image semantics. According to the training approach, deep learning annotation methods can be categorized into two categories, namely: training from scratch-based annotation, and transfer learning-based annotation.

A. TRAINING FROM SCRATCH-BASED ANNOTATION

Most of the annotation methods train their model according to their datasets from scratch, thus allowing the configurations of the model to be under control. The proposed AIA model in [11], relies on CNN and neighbor groups to annotate images using k-nearest neighbor (KNN) model that clusters similar visuals into groups. In the testing phase, the features extracted from a new image are compared to the KNN model to find its similar features. Then, the self-defined Bayesian model is used to assign the tags related to the neighbor set to a new image. However, the model is influenced by the size of the training set.

The last decade has brought significant development in the field of deep learning techniques that sufficiently tackle AIA tasks. Firstly, training the model in a semi-supervised way, where the training images are not fully labeled. Wu *et al.*

presented a model based on deep CNN to annotate images in a semi-supervised manner [12]. Images are sampled from the training image set that contains labeled and unlabelled images. These images are fed into three CNNs that share the same architecture and the same weights. Then, the learned feature representations are considered as activation in the ranking layer, while the Weighted Pairwise Ranking Loss (W2PR) loss layer takes the output of the ranking layer and classify.

Secondly, training in a supervised manner, where all the training images are fully labeled. Kiyokawa *et al.* proposed a fully automated annotation model based on CNN [13]. It uses a single visual marker along with a noise-masking to hide the marker to label images collected manually in automated factories. The labels for each object are identified based on using the IDs of the detected marker. However, in cases where the products are close to each other, the single marker method will fail to detect the product.

There are some hybrid-based deep learning methods that leverage different deep learning-based architectures. Feng *et al.* proposed a hybrid-based model to automatically annotate images [14]. They combined a CNN architecture to model the images and long short-term memory (LSTM) to model the user's tags, which will be concatenated using a multi-layer perceptron (MLP). In the end, a class distribution is produced in the SoftMax layer to predict the labels. In spite of these results, this model has considered the image annotation problem as an image classification problem.

With the intention of improving the training process and tackling the overfitting issue, the most common method is to increase the size of the dataset. Wang *et al.* presented a multi-task voting model based on data augmentation that improves the accuracy of annotation [15]. The proposed model adopts CNN architecture along with an adaptive label to achieve the best number of labels using SoftMax. The authors proved that traditional data augmentation methods are not practical, were sometimes important parts of the image got lost. Thus, another deep learning data augmentation method presented recently called GAN, which is a powerful technique to generate new images in an unsupervised manner.

Adar *et al.* proposed a Deep learning annotation model using GAN [16], where collecting a lot of data is a challenge in the medical field. The model is based on deep CNN as a classification approach for liver lesions datasets. Adopting GAN in the model significantly improves the accuracy of CNN annotation, where it achieved an improvement of 7% over traditional data augmentation. Also, medical images sometimes are unlabelled, and the others are annotated at the image-level.

Ke *et al.* proposed an end-to-end AIA model based on deep CNN (E2E-DCNN) [17], that deals with the feature learning and annotation in an end-to-end manner through the CNN method. They adopt the GAN method to enhance the annotation performance. The earlier mentioned attempts train the model from scratch based on the available datasets. It is known that training from scratch is time-consuming compared to a transfer learning approach that could save a lot of time and effort.

B. TRANSFER LEARNING-BASED ANNOTATION

Transfer learning has received an interest in the field of deep learning models and has proven to save a lot of training time. Also, it assists in AIA tasks such as multi-class and multi-label classification.

Raghu *et al.* proposed a multi-class deep learning-based model to classify types of seizures with a non-seizure electroencephalogram (EEG) by adopting deep CNN and following the transfer learning approach in training the CNN architecture [18]. Recognition of seizure types is crucial for the neurosurgeon to understand the cortical connectivity of the brain. Baltruschat *et al.* proposed a multi-label deep learning-based model to label diseases on chest X-ray images [19]. The proposed model adopted CNN architecture under the transfer learning approach. Also, they considered non-image data information such as (gender, and age) along with the image information to train the model. They proved that integrating patient information is a useful process and enhances classification.

III. DEEP AUTOMATIC IMAGE ANNOTATION MODEL

The proposed DeepAIA model is an end-to-end AIA based on integrating data augmentation method (i.e. an auxiliary classifier GAN (ACGAN)) and CNN classifier, as shown in Figure 2. The DeepAIA model consists of three main stages, namely: (i) data preparation, (ii) training, and (iii) testing. The following sections discuss the main stages of the proposed DeepAIA model in detail:

A. DATASET PREPARATION

In any machine learning method, the process of data preparation and transformation can have a significant impact on the success of the used method and can facilitate the process of learning. Since the proposed model is based on deep learning methods; deep GAN and deep CNN, the input of these methods should be in a specific format [20]. The output of this stage is a pre-processed dataset, which is used as input for the next stage (i.e. training stage).

B. TRAINING

The goal of the training stage is to train the classifier efficiently. It consists of three main phases, namely: (i) synthetic image augmentation, (ii) transfer learning, and (iii) training validation and parameters fine-tuning, which are discussed in detail in the following subsections.

1) SYNTHETIC IMAGE AUGMENTATION

In this research, synthetic image augmentation is used to augment the training set of the selected datasets to further strengthen the learning process of deep CNN architecture and mitigate the problems associated with training using small-sized datasets. A deep ACGAN image augmentation network is adopted in this research because it proves to have more

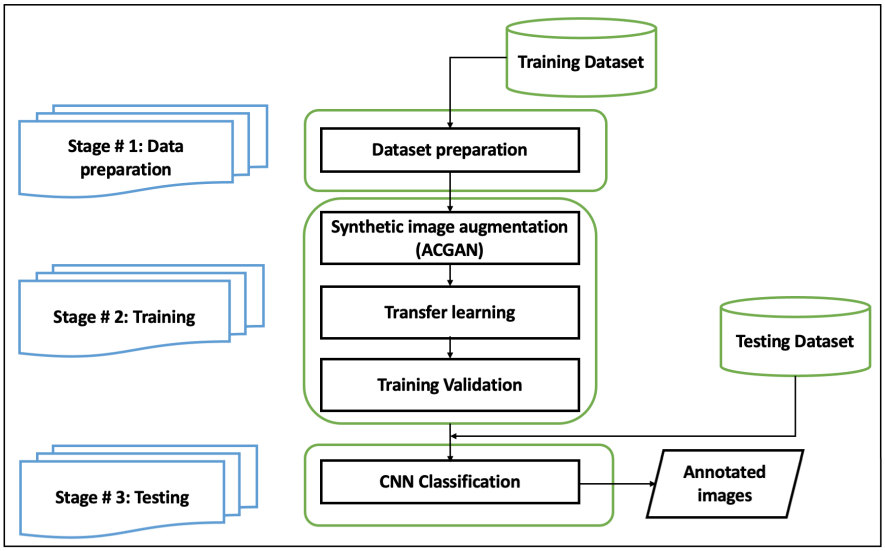

FIGURE 2. The general methodology.

stabilization and higher quality synthesized images compared to the previous GAN architectures.

Basically, the GAN framework comprises the generator network G and the discriminator network D [21]. The G aims at generating fake images $X_{fake} = G(z)$ from a noise vector z, this noise vector is of a fixed length and is drawn randomly from a Gaussian distribution in the latent space. On the other hand, the goal of the D is to differentiate the real image from the synthetic image through probability $P(O|X) = D(X)$, where the O denotes the origin of the image.

Figure 3 illustrates the architecture of DeepAIA. First, images with N labels from the dataset are considered real images, which are going to be as an input to the ACGAN part of the model.

The generator is basically a deconvolution network that will take a random noise along with the class label as an input to generate fake images. Then a mini-batch of the real images along with a mini-batch of fake generated images will be taken as an input to the discriminator, where it will determine the authenticity of the image as if it is synthesized or not, besides reconstructing the class label of the image by using an auxiliary decoder [22] to stabilize the network. Basically, the discriminator consists of a set of second convolutional layers backed up with a Leaky ReLU non-linearity [23]. If the input image is classified as fake the discriminator will calculate feedback to update the network to get better at discriminating in the next round, the same is for the generator, where it will be updated based on how well the synthesized samples fooled the discriminator and generate more realistic images in the next round, in which the ACGAN will loop through E epochs. Finally, the generator will output synthesized images classified as real by the discriminator, which will be combined with the original images in the training set of the dataset to train CNN through transfer learning in the next step of the model.

2) TRANSFER LEARNING

Transfer learning is a method of machine learning in which a model trained on a specific task is reused for a second task. The transfer learning technique has the advantage of reducing the training time and can result in less generalization error [18], [19]. Basically, when removing the classifier layer of a pre-trained CNN architecture, the model will take an image as input and output feature maps. Then, a new classifier layer will take the feature maps as input and learn the new task of annotating images with new labels.

As illustrated in Figure 3, the input to the CNN classifier will be the training set from the dataset along with synthesized image augmentation from ACGAN. Thus, the training of the new classifier of CNN will be empowered to a great extent, where the training set carries data points and noise. By leveraging the merits of transfer learning and ACGAN augmentation, the CNN classification is enhanced.

The main goal of synthetic image augmentation is to synthetically increase the number of samples used in training to ameliorate the performance of CNN [24]. As the GAN mainly grasps the inherent distribution of the data from a set of examples, to produce synthetic images from the learned distribution. Then, once the distribution of each label is learned, the synthesized images are produced using a normally distributed noise as an input vector [24].

3) TRAINING, VALIDATION AND PARAMETERS FINE-TUNING

CNN network has a number of parameters to be set up to avoid getting prone to configuration errors when using a manual tuning of the parameters. Besides, deep learning networks may face an overfitting problem, which means the network starts memorizing the training dataset. Therefore overfitting degrades the generalized performance of the network [25].

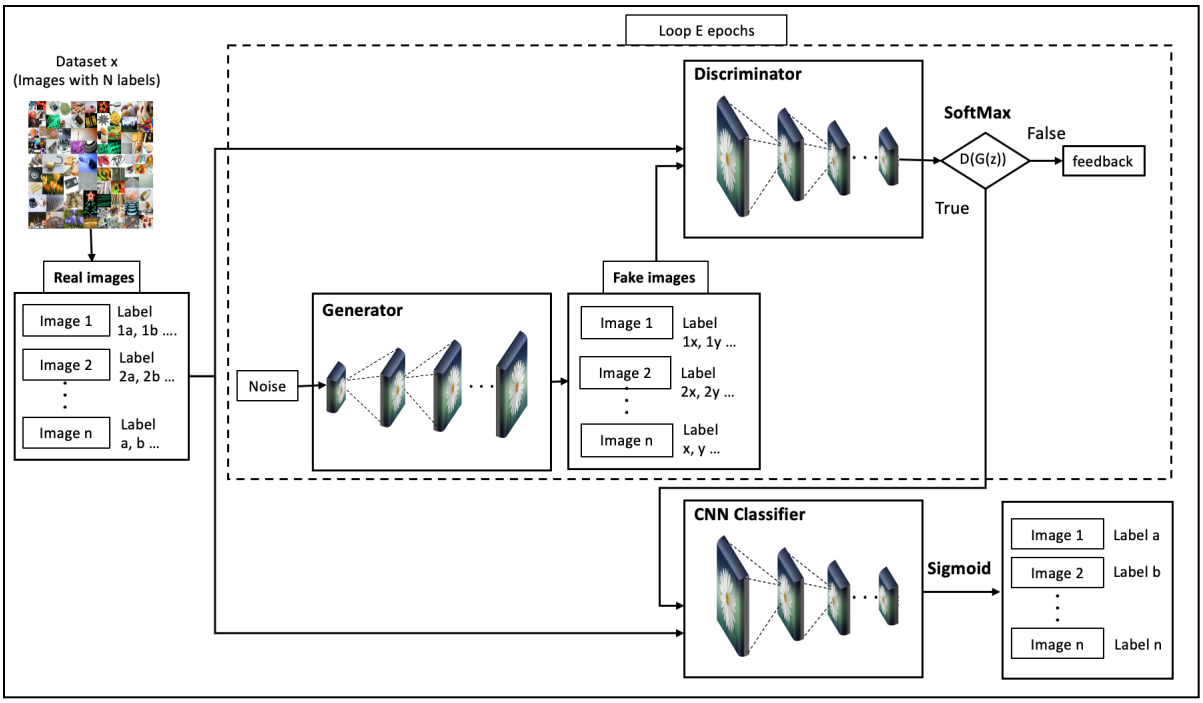

FIGURE 3. Detailed architecture of DeepAIA.

One of the well-known techniques to mitigate the risk of overfitting in CNN architectures is K-fold cross-validation [26]. In addition, K-fold cross-validation is used to evaluate the deep learning model while the model is in the process of parameter adjustment. Moreover, it includes splitting the training data into K number of partitions (usually K=4) [27].

In the Test stage, it aims to test the trained classifier using unseen images. The output of ACGAN is the generated image dataset that is compatible with the shape and size of the original dataset. All images (original dataset and generated dataset) are used to train CNN classifier. After the fully connected layer in the CNN classifier, the classification (i.e. annotation) is executed using the activation function based on the ground-truth set. Last, each image included in the testing set is labeled with one or more labels (i.e. class) based on class score, as shown in Figure 4.

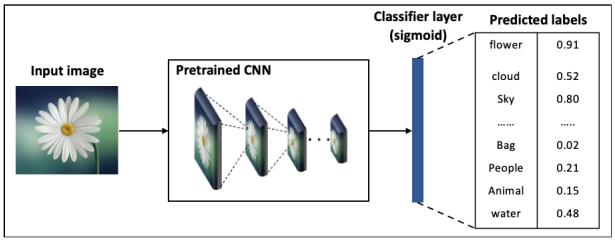

FIGURE 4. CNN classifier.

As mentioned earlier, the proposed DeepAIA aims to annotate images with multiple labels. Thus, the densely connected layer is adopted with the sigmoid activation function, which is an activation function for multi-label classification problems similar to the problem targeted in this research. Sigmoid function has a faster variance rate [28], also it has been adopted in this study since it belongs to [0 to 1].

IV. EXPERIMENTS

Throughout this section, we describe details of experiments with the proposed DeepAIA on four public annotation datasets. Further, we describe the experimental settings that include datasets, environment, parameters, and evaluation metrics. Also, we show the experimental results and analyses of the DeepAIA model. Then, compare the DeepAIA model with previous models in the field of AIA.

A. DATASETS

The proposed model is evaluated on four common, public, and large datasets from the field of image annotation: ImageClef 2011 [29], ImageClef 2012 [30], ESPGame [31], and Iaprtc12 [32]. The images in these datasets are of different categories such as color, weather, vehicles, natural, etc., making the annotation a difficult task. The Datasets with their ground truth are available at (https://zenodo.org/record/5570889#.YWoC3EZBw1I). The summary information of the four datasets are shown in Table 1.

The selected datasets are prepared in order to be used as an input for DeepAIA in the required format. Since images are of different sizes, all images in the datasets have been uniformly resized, where MobileNet and ResNet-101 require

TABLE 1. Statistics of the datasets.

Dataset	Training images	Testing images	Number of labels
ImageClef 2011	8000	10000	99
ImageClef 2012	15000	10000	94
ESPGame	18689	2081	268
Iapr-tc12	17665	1962	291

a 224 × 224 image input shape, while Inception requires a 299 × 299 image input shape.

B. ENVIRONMENT AND PARAMETERS

The environment employed in the experiments was as follows: windows 10 64-bit operating system, x64-based processor, intel(R) Core i7, CPU @ 2.80 GHz, 16 GB memory. The primary programming language used to build and execute the experiments was python.

All deep learning-based models contain a number of parameters not trained by the training set called hyperparameters, which influences the accuracy of the model. The ACGAN framework is associated with a set of hyperparameters that can affect the accuracy of the resulting augmentation. There are works that experiment with tuning these hyperparameters on a problem similar to ours. Thus, the batch size and latent size values were considered from [22]. In addition, Adam learning rate, and Adam beta values were considered from [33] along with LeakyReLU non-linearity on the discriminator. CNN architecture is also associated with a set of hyperparameters which can affect the accuracy of the resulting annotation. The learning rate value is set to 0.001 and batch size is set to 128 [34], on all selected datasets.

C. EVALUATION METRICS

To determine the effectiveness of the proposed model, it is evaluated with the most commonly used metric in image annotation including:

1) MEAN INTERPOLATED AVERAGE PRECISION (MiAP)

MiAP is a metric used to evaluate the performance per label [35]. MiAP is computed based on equation 1.

$$MiAP = \frac{1}{n}\sum_{i=1}^{n} AP_{Interp} \tag{1}$$

where the MiAP is obtained by interpolating the precision only at the 11 levels r taking the maximum precision whose recall value is greater than r, as illustrated in equation 2 and 3.

$$AP_{Interp} = \frac{1}{11}\sum_{r\in\{0.0,0.1,....,1.0\}} P_{interp}(r) \tag{2}$$

$$P_{interp}(r) = \max p(r'), r' \geq r \tag{3}$$

where n denotes the total number of labels, AP_{interp} denotes average interpolated precision, P_{interp} denotes interpolated precision, $p(r')$ is the precision at recall r'.

2) PRECISION (P)

Precision is a metric measuring the number of positive predictions that is made correctly [36], as in the following equation:

$$Precision(p) = \frac{TP}{TP + FP} \tag{4}$$

TP indicates the true positive, FP indicates the false positive.

3) RECALL (R)

Recall is a metric that quantifies the number of correct positive predictions from all positive predictions that could have been made [36], as in the following equation:

$$Recall(R) = \frac{TP}{TP + FN} \tag{5}$$

FN indicates a false negative.

4) F1

It is a weighted average of the precision and recall. It is used to evaluate the performance per image as shown in equation 6.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

where P denotes precision as in equation 4, and R denotes recall as in equation 5.

5) AREA UNDER THE CURVE (AUC)

AUC is used in the ImageClef annotation task to evaluate the performance of the annotation model per label. The AUC is computed using the height of the recall values by the false positive rate [37].

6) EQUAL ERROR RATE (EER)

It is also used in the ImageClef annotation task to evaluate the performance of the annotation model per label. EER is computed where the false positive rate (FPR) and false negative rate (FNR) intersect at a certain point [38]. FPR and FNR are calculated using equation on 7 and 8 respectively.

$$FPR = \frac{FP}{v} \tag{7}$$

$$FNR = \frac{FN}{v} \tag{8}$$

where v denotes the total number of negative elements, FP denotes the total number of misannotated positive elements, and FN denotes the total number of misannotated negative elements.

V. RESULTS AND DISCUSSION

The proposed DeepAIA model was implemented using transfer learning of three deep CNN architectures (MobileNet, ResNet-101, Inception), as mentioned earlier. After training and testing DeepAIA on the four selected datasets the results came as follows: the best-scored values were achieved with Inception CNN architecture for both ImageClef 2011 and Image-Clef 2012, as illustrated in Table 2 with the selected

TABLE 2. Experimental results of the proposed DeepAIA model using different transfer learning models on ImageClef datasets.

Model	ImageClef 2011						ImageClef 2012					
	P	R	F1	MiAP	EER	AUC	P	R	F1	MiAP	EER	AUC
MobileNet	0.903	0.898	0.898	0.640	**0.040**	0.802	0.946	0.954	0.949	0.518	**0.025**	0.791
ResNet-101	0.902	0.894	0.897	0.660	0.045	0.777	0.947	0.942	0.944	0.453	0.028	0.792
Inception	0.899	0.961	**0.927**	**0.714**	0.061	**0.806**	0.945	0.984	**0.963**	**0.567**	0.037	**0.800**

TABLE 3. Experimental results of the proposed DeepAIA model using different transfer learning models on ESPGame and Iapr-tc12 datasets.

Model	ESPGame						iapr-tc12					
	P	R	F1	MiAP	EER	AUC	P	R	F1	MiAP	EER	AUC
MobileNet	0.991	0.990	**0.991**	0.032	0.007	0.672	0.981	0.973	0.976	0.298	0.026	0.692
ResNet-101	0.992	0.968	0.980	0.042	0.025	0.667	0.981	0.968	0.974	0.242	0.030	0.713
Inception	0.991	0.989	0.990	0.045	0.007	0.624	0.981	0.994	**0.986**	0.218	0.017	0.669

evaluation metrics by strictly following the testing guidelines in ImageClef datasets.

while the best-scored values were achieved with MobileNet and Inception CNN architectures for ESPGame and IAPR-TC 12 datasets respectively, as illustrated in Table 3.

A. COMPARISON AGAINST THE STATE-OF-THE-ART MODELS

To demonstrate the improvement that DeepAIA achieved, a comparison against the state-of-the-art models was conducted. For ImageClef 2011, the proposed DeepAIA model achieved results that outperform all the state-of-the-art models considered in Table 4 in terms of MiAP, F-measure, and AUC. The proposed DeepAIA model outperforms even the best-scored results achieved by the multimodal model proposed in [39].

For ImageClef 2012, the proposed DeepAIA model achieved results that outperform all the state-of-the-art models considered in Table 4 in terms of MiAP. Where MiAP was the only metric publicly available for all models considered in the comparison. The proposed DeepAIA model outperforms the best-scored result achieved by the multimodal model proposed in [40].

For ESPGame and IAPR-TC 12 datasets, the evaluation metrics illustrated in the experimentation were precision, recall, and f-measure. The results achieved with the proposed model are compared to the results of the state-of-the-art models in the annotation task. The proposed DeepAIA model achieved results that outperform all the state-of-the-art models considered in Table 5. The proposed DeepAIA model outperforms the best-scored results achieved by the E2E-DCNN model proposed in [17] when considering both ESPGame and IAPR-TC 12 datasets.

B. DISCUSSION

The experimental results illustrate DeepAIA's capabilities of performing multi-label annotation for given images. The model has effectively succeeded in outperforming in all the datasets when compared against other studies, as shown in Table 4 and 5.

In ImageClef 2011 dataset, the DeepAIA model compared to the best-scored model [39] achieved a gain of 5.61% in terms of MiAP, a gain of 16.15% in terms of F-measure, and

TABLE 4. Comparison of DeepAIA and the state-of-the-art models on ImageClef datasets.

Model	ImageClef 2011				ImageClef 2012
	MiAP	F-measure	ERR	AUC	MiAP
IDMT [41]	0.371	0.551	0.303	0.752	-
ISIS [42]	0.433	0.622	0.246	0.821	-
BPACAD [43]	0.436	0.593	0.242	0.828	-
LIRIS [44]	0.437	0.567	0.233	**0.837**	0.436
MLKD [45]	0.402	0.559	0.253	0.817	0.318
TUBFI [46]	0.443	0.566	0.234	0.836	-
Multimodal in [39]	0.677	0.799	0.202	0.723	0.321
Multimodal in [40]	0.448	-	-	-	0.431
Multimodal in [47]	0.453	-	-	-	-
Multimodal in [48]	0.436	-	-	-	-
CEA LIST [49]	-	-	-	-	0.408
DMS-SZTAKI [50]	-	-	-	-	0.425
ISI [51]	-	-	-	-	0.413
KIDS NUTN [52]	-	-	-	-	0.171
The proposed model	**0.715**	**0.928**	**0.061**	0.807	**0.567**

TABLE 5. Comparison of DeepAIA and the state-of-the-art models on ESPGame and Iaprtc12 datasets.

Model	ESPGame			IAPR-TC 12		
	Precision	Recall	F-measure	Precision	Recall	F-measure
MBRM [53]	0.18	0.19	0.19	0.24	0.23	0.24
GS [54]	-	-	-	0.32	0.29	0.3
JEC [55]	0.22	0.25	0.23	0.28	0.29	0.29
LM3L [56]	0.4	0.26	0.32	0.44	0.28	0.34
$X^2kernel$ [57]	0.38	0.21	0.27	0.42	0.24	0.31
ANNOR-G [58]	0.36	0.29	0.32	0.38	0.31	0.34
FFSS [59]	0.21	0.23	0.22	0.29	0.29	0.29
MLRank [60]	-	-	-	0.38	0.32	0.35
TagProp [61]	0.39	0.27	0.32	0.46	0.35	0.4
NL-ADA [62]	0.36	0.21	0.27	0.42	0.3	0.35
DIA [63]	0.35	0.41	0.37	0.33	0.41	0.37
D^2GAN [64]	0.35	0.42	0.38	0.33	0.45	0.4
SEM [11]	0.38	0.42	0.4	0.41	0.39	0.4
E2E-DCNN [17]	0.48	0.39	0.43	0.48	0.43	0.45
Proposed Model	**0.991**	**0.990**	**0.991**	**0.981**	**0.994**	**0.986**

achieved a 69.8% reduction in terms of EER. However, for AUC the best-scored result was achieved by the "LIRIS" model [44], in which the DeepAIA achieved a result near the best-scored result and higher than most of the other models. While in ImageClef 2012 dataset, the proposed DeepAIA model achieved a gain of 30.05% in terms of MiAP compared to the best-scored model [39].

For the ESPGame dataset, compared to the best-scored model, the proposed DeepAIA model achieved a gain of 130% in terms of F-measure. Moreover, for the IAPR-TC12 dataset, compared to the best-scored model, the proposed DeepAIA model achieved a gain of 119% in terms of F-measure.

VI. CONCLUSION

With the explosive growth of digital images, the need to describe the images at a semantic level to facilitate indexing and arranging large-scale images has increased. Thus, the automatic interpretation and the uncovering of important information included in the image through annotating it with accurate labels is the main task of AIA. As well, accurate retrieving of images on demand is one among many real-life AIA applications. Consequently, different approaches varying from statistical methods to newly deep learning (DL) have been used to get the best possible performance on all kinds of datasets.

In this work, a DeepAIA model capable of automatically annotating large-scale images was proposed. The framework of the DeepAIA model adopts a well-known technique in image classification and annotation called CNN. In this research, the learning of CNN architecture was transfer learning of various pre-trained CNN architectures, that have proven to contribute to fairly increasing the model performance. Also, the benefit of data augmentation through a well-known technique called GAN was exploited, where data augmentation enriches the training set for better learning of the CNN architecture. The results of testing the Deep AIA on four different datasets were reported. By comparing with other models, DeepAIA outperformed state-of-the-art results on the four selected datasets.

The possible direction of research is to further experiment with different kinds of GAN architectures for data augmentation. In addition, experimenting with different pre-trained CNN architectures such as Googlenet, Lenet, and xception. On the more advanced level, the AIA task that DeepAIA solved can go beyond the simple annotation of image content to identify complex types. For instance, the task could be the identification and classification of weather conditions such as hot, cold, and humid or emotional states such as happy, sad, or confused. Besides, the task of DeepAIA can be extended to live annotating videos.

ACKNOWLEDGMENT

The Deanship of Scientific Research (DSR) at King Abdulaziz University (KAU), Jeddah, Saudi Arabia has funded this project, under grant no. (FP-223-43).

REFERENCES

[1] *Instagram by the Numbers: Stats, Demographics and Fun Facts*, Omnicore, Cincinnati, OH, USA, 2018.
[2] Y. Chen, X. Zeng, X. Chen, and W. Guo, "A survey on automatic image annotation," *Appl. Intell.*, vol. 50, pp. 1–17, Jun. 2020.
[3] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognit.*, vol. 45, no. 1, pp. 346–362, Jan. 2012.
[4] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, 2008, pp. 327–336.
[5] S.-B. Chan, H. Yamana, D.-D. Le, and S. Satoh, "Image annotation fusing content-based and tag-based technique using support vector machine and vector space model," in *Proc. 10th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2014, pp. 272–276.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2017.
[7] Y. Chen, L. Liu, J. Tao, X. Chen, R. Xia, Q. Zhang, J. Xiong, K. Yang, and J. Xie, "The image annotation algorithm using convolutional features from intermediate layer of deep learning," *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 4237–4261, Jan. 2021.
[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[9] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 4, pp. 588–598, Sep. 2017.
[10] F. Tancini, Y.-L. Wu, W. B. Schweizer, J.-P. Gisselbrecht, C. Boudon, P. D. Jarowski, M. T. Beels, I. Biaggio, and F. Diederich, "1,1-Dicyano-4-[4-(diethylamino)phenyl]buta-1,3-dienes: Structure-property relationships," *Eur. J. Organic Chem.*, vol. 2012, no. 14, pp. 2756–2765, May 2012.
[11] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3767–3780, Feb. 2019.
[12] F. Wu, Z. Wang, Z. Zhang, Y. Yang, J. Luo, W. Zhu, and Y. Zhuang, "Weakly semi-supervised deep learning for multi-label image annotation," *IEEE Trans. Big Data*, vol. 1, no. 3, pp. 109–122, Sep. 2015.
[13] T. Kiyokawa, K. Tomochika, J. Takamatsu, and T. Ogasawara, "Fully automated annotation with noise-masked visual markers for deep-learning-based object detection," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1972–1977, Apr. 2019.
[14] F. Feng, R. Liu, X. Wang, X. Li, and S. Bi, "Personalized image annotation using deep architecture," *IEEE Access*, vol. 5, pp. 23078–23085, 2017.
[15] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, and Q. Zhang, "Large scale automatic image annotation based on convolutional neural network," *J. Vis. Commun. Image Represent.*, vol. 49, pp. 213–224, Nov. 2017.
[16] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.
[17] X. Ke, J. Zou, and Y. Niu, "End-to-end automatic image annotation based on deep CNN and multi-label data augmentation," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2093–2106, Aug. 2019.
[18] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Netw.*, vol. 124, pp. 202–212, Apr. 2020.
[19] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.
[20] J. Vijayabhaskar, "Multi-label image classification tutorial with Keras ImageDataGenerator," Tech. Rep., 2019.
[21] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "COVIDGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020.
[22] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 6, 2017, pp. 4043–4055.
[23] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30, no. 1, p. 3.

[24] R. Verma, R. Mehrotra, C. Rane, R. Tiwari, and A. K. Agariya, "Synthetic image augmentation with generative adversarial network for enhanced performance in protein classification," *Biomed. Eng. Lett.*, vol. 10, no. 3, pp. 443–452, Aug. 2020.
[25] G. Swapna, S. Kp, and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," *Proc. Comput. Sci.*, vol. 132, pp. 1253–1262, Mar. 2018.
[26] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *Proc. IEEE 6th Int. Conf. Adv. Comput. (IACC)*, Feb. 2016, pp. 78–83.
[27] F. Chollet, *Deep Learning With Python*. Shelter Island, NY, USA: Manning, Nov. 2017.
[28] L. Wei, J. Cai, V. Nguyen, J. Chu, and K. Wen, "P-SFA: Probability based sigmoid function approximation for low-complexity hardware implementation," *Microprocessors Microsyst.*, vol. 76, Jul. 2020, Art. no. 103105.
[29] *Visual Concept Detection and Annotation Task 2011*, 2011.
[30] *Photo Annotation and Retrieval 2012*, 2012.
[31] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. Conf. Hum. Factors Comput. Syst. (CHI)*, 2004, pp. 319–326.
[32] M. Henning and D. Thomas, "The IAPR benchmark: ANew evaluation resource for visual information systems," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2006, pp. 13–23.
[33] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[34] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, "Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2018, pp. 392–402.
[35] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in *Proc. Asian Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatic, vol. 10115, 2017, pp. 198–213.
[36] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[37] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2005.
[38] H. Sahbi and X. Li, "Context-based support vector machines for interconnected image annotation," in *Proc. Asian Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 2011, pp. 214–227.
[39] M. Taileb and E. Alahmadi, "Multimodal automatic image annotation method using association rules mining and clustering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 11, pp. 678–684, 2018.
[40] A. Znaidia, "Handling imperfections for multimodal image annotation," Ph.D. dissertation, Ecole Centrale Paris, Gif-sur-Yvette, France, 2014.
[41] K. Nagel, S. Nowak, U. Kühhirt, and K. Wolter, "The Fraunhofer IDMT at image CLEF 2011 photo annotation task," in *Proc. CEUR Workshop*, vol. 1177, 2011, pp. 1–8.
[42] K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "The university of Amsterdam's concept detection system at ImageCLEF 2009," in *Proc. Workshop Cross-Lang. Eval. Forum Eur. Lang.* Berlin, Germany: Springer, 2009, pp. 261–268.
[43] B. Daróczy, R. Pethes, and A. A. Benczúr, "Sztaki@ ImageCLEF 2011," in *Proc. CLEF, Notebook Papers/Labs/Workshop*, 2011, pp. 1–6.
[44] N. Liu, Y. Zhang, E. Dellandréa, S. Bres, and L. Chen, "LIRIS-imagine at ImageCLEF 2011 photo annotation task," in *Proc. CLEF, Notebook Papers/Labs/Workshop*, vol. 11, 2011, p. 9.
[45] S. Nowak, K. Nagel, and J. Liebetrau, "The CLEF 2011 photo annotation and concept-based retrieval tasks," in *Proc. CEUR Workshop*, vol. 1177, 2011, pp. 1–25.
[46] A. Binder, W. Samek, M. Kloft, C. Müller, K. R. Müller, and M. Kawanabe, "The joint submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the image CLEF 2011 photo annotation task," in *Proc. CEUR Workshop*, vol. 1177, 2011.
[47] Y. Zhang, S. Bres, and L. Chen, "Semantic bag-of-words models for visual concept detection and annotation," in *Proc. 8th Int. Conf. Signal Image Technol. Internet Based Syst. (SITIS)*, Nov. 2012, pp. 289–295.
[48] N. Liu, E. Dellandréa, L. Chen, C. Zhu, Y. Zhang, C.-E. Bichot, S. Bres, and B. Tellez, "Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme," *Comput. Vis. Image Understand.*, vol. 117, no. 5, pp. 493–512, May 2013.
[49] E. Gadeski, H. Le Borgne, and A. Popescu, "CEA LIST's participation to the scalable concept image annotation task of ImageCLEF 2015," in *Proc. CLEF, Working Notes*, 2015, pp. 1–6.
[50] B. Daróczy, D. Siklósi, and A. A. Benczúr, "DMS-sztaki@ ImageCLEF 2012 photo annotation," in *Proc. CLEF, Online Working Notes/Labs/Workshop*, 2012.
[51] Y. Ushiku, H. Muraoka, S. Inaba, T. Fujisawa, K. Yasumoto, N. Gunji, T. Higuchi, Y. Hara, T. Harada, and Y. Kuniyoshi, "ISI at ImageCLEF 2012: Scalable system for image annotation," in *Proc. CEUR Workshop*, vol. 1178, 2012.
[52] B. C. Chien, G. B. Chen, L. J. Gaou, C. W. Ku, R. S. Huang, and S. E. Wang, "KIDS-NUTN at ImageCLEF 2012 photo annotation and retrieval task," in *Proc. CEUR Workshop*, vol. 1178, 2012.
[53] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2004, p. 2.
[54] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas, "Automatic image annotation using group sparsity," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3312–3319.
[55] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 316–329.
[56] B. Hariharan, L. Zelnik-Manor, S. V. Vishwanathan, and M. Varma, "Large scale max-margin multi-label classification with priors," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 423–430.
[57] Y. Wang, H. Dawood, Q. Yin, and P. Guo, "A comparative study of different feature mapping methods for image annotation," in *Proc. 7th Int. Conf. Adv. Comput. Intell. (ICACI)*, Mar. 2015, pp. 340–344.
[58] E. Kuric and M. Bielikova, "ANNOR: Efficient image annotation based on combining local and global features," *Comput. Graph.*, vol. 47, pp. 1–15, Apr. 2015.
[59] X. Zhang and C. Liu, "Image annotation based on feature fusion and semantic similarity," *Neurocomputing*, vol. 149, pp. 1658–1671, Feb. 2015.
[60] Z. Li, J. Liu, C. Xu, and H. Lu, "MLRank: Multi-correlation learning to rank for image annotation," *Pattern Recognit.*, vol. 46, no. 10, pp. 2700–2710, Oct. 2013.
[61] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 309–316.
[62] X. Ke, M. Zhou, Y. Niu, and W. Guo, "Data equilibrium based automatic image annotation by fusing deep model and semantic propagation," *Pattern Recognit.*, vol. 71, pp. 60–77, Nov. 2017.
[63] B. Wu, F. Jia, W. Liu, and B. Ghanem, "Diverse image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6194–6202.
[64] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, "Tagging like humans: Diverse and distinct image annotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7967–7975.

ABEER ALSHEHRI received the bachelor's degree in information technology from the College of Computer and Information Technology, Taif University, Saudi Arabia, in 2018, and the M.Sc. degree in information technology from the College of Computer and Information Technology, King Abdulaziz University, Saudi Arabia, in 2021. Her research interests include image annotations and deep learning.

MOUNIRA TAILEB received the Ph.D. degree in computer science from the University of Paris-Sud, Paris, France, in 2008. She is currently an Associate Professor at the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include artificial intelligence and machine learning.

REEM ALOTAIBI received the Ph.D. degree in computer science from the University of Bristol, Bristol, U.K., in 2017. From 2017 to 2018, she was a Visiting Lecturer with the Intelligent Systems Laboratory, University of Bristol. She is currently an Associate Professor with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. She is also the Supervisor of the Information Technology Department. Her research interests include artificial intelligence and machine learning.