

Lung cancer detection using image processing and deep learning

Asraa A. Abd Al-Ameer¹, Ghufraan Abdulameer Hussien², Hajer. A. Al Ameri¹

¹Department of Mathematics, Faculty of Education, Al-Zahraa University for Women, Karbala, Iraq

²Department of Business Administration, Al-Mustaqbal University College, Babel, Iraq

Article Info

Article history:

Received Apr 21, 2022

Revised Aug 12, 2022

Accepted Aug 31, 2022

Keywords:

Classification

Deep learning

Histopathology

Image compression

Image processing

Lung cancer detection

ABSTRACT

This project is about the detection of lung cancer by training a model of deep neural networks using histopathological lung cancer tissue images. Different models have been proposed for detecting lung cancer cells automatically involving Inception V3, Random Forest, and convolutional neural network (CNN). The deep convolutional neural network has been trained to extract important features that facilitate build detection and diagnosis of lung cancer cells more efficiently and accurately. The proposed method in this project has accomplished promising and satisfactory results in terms of accuracy, precision, recall, F-score, and specificity measure in lung cancer detection. Furthermore, it has been applied on dataset which contains 178,000 photos. The accuracy values that are obtained are accuracy 97.09%, precision 96.89%, recall 97.31%, F-score measure 97.09%, and specificity measure 96.88%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Asraa A. Abd Al-Ameer

Department of Mathematics, Faculty of Education, Al-Zahraa University for Women

Karbala, Iraq

Email: asraa.abd.alhussien@alzahraa.edu.iq

1. INTRODUCTION

Cancer is the most common malignancy worldwide, caused primarily by smoking and exposure to substances such as radon, arsenic, and asbestos. Furthermore, the use of preservative-laced manufactured foods [1]. Lung cancer is cancer that starts in the lungs and spreads throughout the body. The lungs, which receive oxygen through inhalation and release carbon dioxide through exhalation, are too spongy in the chest [2]. It is the main cause of cancer death in both men and women in the United States and many other countries [3], [4]. Every year, about 150,000 people die from lung cancer, while another 200,000 people are diagnosed with the condition [5].

Like all cancers, cancer of the lung is resulting from an abnormality in the body's basic unit of life [6]. Manually, lung cancer is detected by observing lung tissue images, but the manual process of detection is time consuming [7], [8]. The better approach is to build a model which can detect whether a person is having cancer or not in a few minutes [9]. Such a model can be built with the help of image processing and neural networks. So, the aims and the objectives of the proposed work are to construct a program used for the detection of lung cancer using python machine learning, so it could be able to: i) Decrease the rules for testing, ii) To reduce the time and cost required for various excessive medical tests, iii) For improving the manual analysis, iv) Early-stage detection of cancer, and v) Increasing the survivability of the patient.

Image processing explained in [10] has been used with the abilities of deep learning neural networks explained in [11] to perform this task semi-automatically by the computer quickly and efficiently. The findings imply that deep learning models can provide fast, accurate, and low-cost cancer detection to both professionals and patients, and hence have a substantial impact on cancer detection.

There are many works that tried to detect and diagnose lung cancer via image processing abilities and deep learning algorithms. Yang *et al.* [12], presented a convolutional neural network (CNN) architecture, and classification accuracy for the original image of lung cancer nodules using a geometrical dataset and the same transformations for the module shape. They built a block of convolutional neural network (CNN) to map the input data to an output layer. To perform the classification task, this CNN uses convolutional layers, maximum pooling layers, activate layers, and a soft Max layer to extract and lead features map. The team focused on comparative studies with different datasets. They trained CNN using smaller regions, are downsampled the images to half. Their training set consisted of 81,000 images divided into (40,500) images of cancer, and (40,500) non-cancerous images. The CNN architecture is not unique, but the convolutional layer filters' parameters and the size of the Max pooling operators must be consistent. A set of 160 images with equal size for both cancer and non-cancerous was used as a validation set. They found that training the CNN with real clinical lung computed tomography (CT) images and associated labels, and data augmentation optimized the diagnostic performance. While simple geometric transformations of real nodules are indeed effective.

Convolutional neural network have been also used by Coccia *et al.* [13], in order to help pathologists to differentiate between various histologic subtyping by assisting them in lung neuroendocrine neoplasms recognition. Their proposed work uses the technique of hybrid segmentation to separate the lung cancer cells to diagnose lung tumor. This technique combines Fuzzy *c* with active contour. CNN was used to train the segmented portion in order to classify the segmented region as normal or abnormal. The implemented technique verifies a good result where it provides 96.67% accuracy. While Malik *et al.* [14] are based on deep learning (DL) model implemented a multi-classification algorithm for identifying the diseases of lung cancer, COVID-19, and pneumonia. They proposed a model with the combination of Vgg-19 and CNN named BDCNet and applied it to various publically available benchmark datasets. The experiments showed that the model gave a remarkable performance, and provide important assistance for health experts and diagnostic radiographers. The whole-slide images available in the Cancer Genome Atlas (TCGA) have been used by Dehkharghanian *et al.* [15] for developing a deep-learning model for the automatic analysis of tumor slides where in order to recognize tumor in lung versus normal tissue. After testing and evaluating the deep-learning model by the training they found that artificial intelligence technology based on deep-learning models can help pathologists in cancer subtype detection or gene mutations in any cancer type with a save of costs and time. Jin *et al.* [6] built a model that used CNNs for classification task in the computer aided design (CAD) system. Their approach achieved 84.6% accuracy, 82.5% sensitivity, and 86.7% specificity. During the extraction stage, the system applied the circular *r* filter in focused regions, which helped to lower the overall cost of the detection and training stages. While Kalaivani *et al.* [16], designed a model that built a CNN to minimize the number of parameters and adjust the architecture of the network for the purpose of image classification. The CNN is made up of a series of layers each with its features and functionality. They collected fresh data that allow them to maximize the size and add uncertainty within the dataset. They created a website to store the dataset of the patients and allowed the patients to log in to their page at any time for further references. The dataset consisted of about 201 images divided into (171 for the training set, and 30 for testing set). The trained set was classified in the classification stage as either normal or abnormal, and the output would be displayed. From the previous works, the importance has been noted of deep learning in extracting features important in facilitating build diagnosis and detection of lung cancer cells in a more efficient way. Therefore the proposed work used deep learning and image processing techniques to detect lung cancer in order to help pathologists.

2. THE COMPREHENSIVE THEORETICAL BASIS

The proposed research is useful in terms of medical and health. It's a basis of many concepts in order to finally provide a model for lung cancer detection. The following subsections are theoretica information for the proposed work methods background.

2.1. Digital image processing

Is the process of utilizing a digital computer for processing digital images using an algorithm [17]. Digital image processing is a collection of technologies and concepts that may be used in a variety of situations. The goal of early image processing was to enhance the image's quality. Image processing is a common type of image processing. Image processing techniques include enhancement, encoding, restoration and compression [18].

2.2. Medical imaging

The method and practice to image the interiors of a body during clinical examination and medical intervention, while the visual depiction of a function of particular organs or tissues is known as medical imaging. Medical imaging is used to identify and treat disease by revealing interior structures that are covered

by skin and bones. Medical imaging also creates a database on normal anatomy and physiology, allowing anomalies to be detected [19].

2.3. Image compression

The method of portraying essential information with the smallest amount of data feasible [20] is known as image compression. The forms of compression are lossless and lossy, depending on whether the original picture can be recovered from the compressed file. It is the process of reducing the size of a graphics file in bytes without compromising the image's quality. Compression aids in the reduction of storage costs and the more efficient transmission of pictures.” joint photographic experts group (JPEG)” and “motion picture experts group (MPEG)” are two examples of compressed pictures [21].

2.4. Image classification

Sometimes known as image recognition, is a subset of computer vision in which algorithms classify images into categories. As input, it takes the image and gives a prediction on the image class. There are two types: supervised classification, in which the classification algorithm is trained on a collection of pictures and their labels, and unsupervised classification, in which the algorithm is taught solely on raw data [22].

2.5. Deep learning

Is a subject that works on artificial neural networks and is centered on learning and developing on its own by evaluating computer algorithms. Larger, more powerful neural networks have been enabled by big data analytics, allowing computers to monitor, understand, and react to complicated events more quicker than people. It can tackle any pattern recognition issue without the need for human interaction. Its an area of machine learning that deals with artificial neural networks which are algorithms inspired by the biological function of the brain [23]. It enables computer models with several processing layers for learning data representations at various degrees of abstraction. Speech recognition with visual object identification, object detection, and many other disciplines such as drug development and genomics have all benefited from these technologies. Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks (both supervised and unsupervised) have been applied to fields such as computer vision, natural language processing, drug design, machine translation, speech recognition, bioinformatics, material inspection, medical image analysis, and board game programs, where they have produced results that are comparable to those of supervised neural networks [24].

2.6. Convolutional neural networks (CNNs)

These are one of the machine learning neural networks. It's most commonly used as an image classifier in computer vision [6]. They're a multi-layer neural network made up of two types of layers: convolution layers “(c-layers)” and sub-sampling layers “(s-layers)”, which are connected alternately and make up the network's center section [12].

3. METHOD

This work is about building a computerized system that has the ability of lung cancer detection by means of a training model of deep learning using histopathological lung cancer tissue images. The traditional approaches for detecting lung cancer are extremely costly and time-consuming. As a result, this project has been created to help speed up and facilitate the process.

3.1. Dataset

The dataset that is used for this project is an image dataset, consisting of 220,025 RGB lung cancer tissue images, obtained from Kaggle. Figure 1 shows the dataset of cancerous and non-cancerous tissue images. The first step is to divide the dataset into two regions, one consists of (130,908) images for training with respective labels and another 89,117 images for testing out of the total training images. 41% are cancerous images and the remaining are noncancerous. From the pie chart shown in Figure 2, it can be interpreted that the dataset is imbalanced.

In the second step, the concept of downsampling or under sampling is used by reducing the size of the noncancerous class, in order to overcome the dataset imbalance since an imbalance affects accuracy. The dataset has downsampled to a total of 17,800 images, divided the entire dataset into training and validation folders as shown in Table 1. So 90% of the dataset, which are 160,200 images, is used for training the model, and the remaining 10%, which are 17,800 images, is used for validation.

The dataset file structure is consist of 3 folders, folder for training, for testing, and for validation as depicted in Figure 3. Both the training and validation folders, labeled "train" and "val," include malignant and non-cancerous photos that will be utilized to train and validate accuracies and losses in each epoch during a

training process. The third folder named “test” involves completely unseen cancerous and non-cancerous images that will be used for the testing process that shows how accurate the model is.

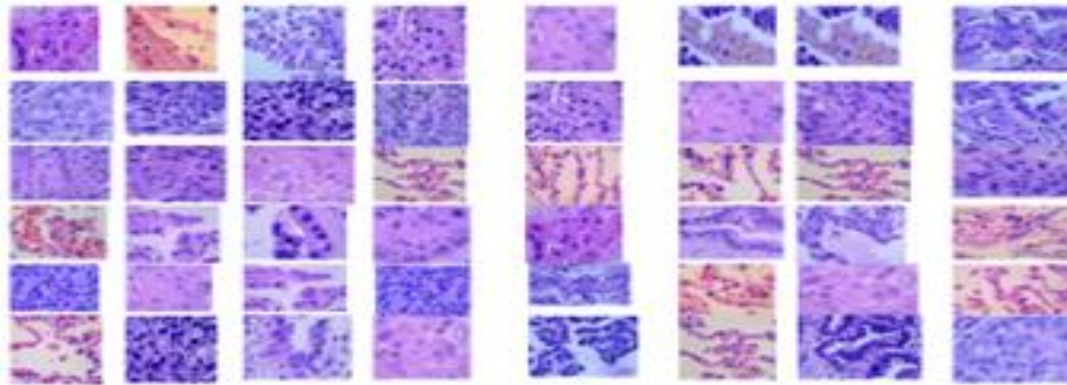


Figure 1. Dataset of cancerous and non-cancerous tissue images

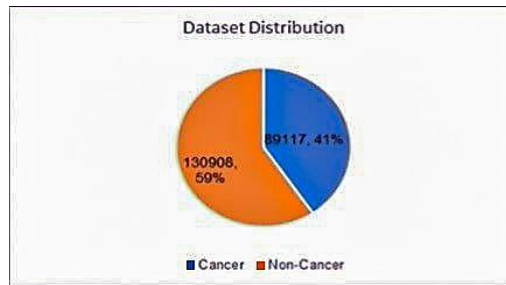


Figure 2. Distribution of dataset

Table 1. Dataset distribution

Division \ Type	Cancer	Non-Cancer	Total
Training	80100	80100	160200
Validation	8900	8900	17800
Total	89000	89000	178000

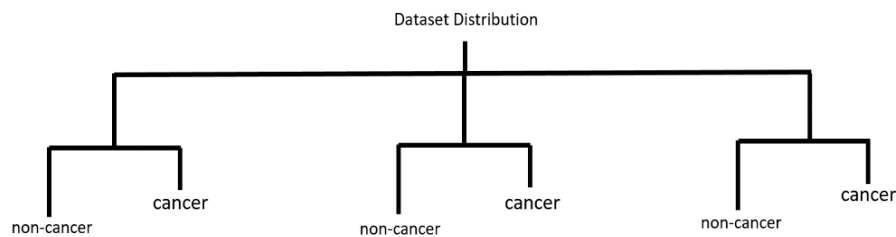


Figure 3. Distribution of dataset folder into train, test and validation subsets

3.1. Pre-processing of the images

On our training dataset, a CNN model has been trained to categorize (jpeg) photos of lung tissue into two groups (cancerous, non-cancerous). A color picture with dimensions of 224,224 is fed into the CNN model that has been pre-trained. So, all the images have been resized to the required size. Then the following variables have been derived. A normalization process was made to all of the items by dividing them by 225.0, and because the images were stored in folders ImageDataGenerator has been used in order to label the corresponding dataset. The labels in our case are (cancerous, non-cancerous). Table 2 contains one-dimensional arrays called “y-train, y-val, and y-test”.

Table 2. Variables of the model

x_train	y_train	x_test	y_test	x_val	y_val
NumPy arrays of the images of the training dataset	Labels of the training dataset	NumPy arrays of the images of the testing dataset	Labels of the testing dataset	NumPy arrays of the images of the validation dataset	Labels of the validation dataset

3.2. Implementation

The third step is the most important process which is generating the model. The CNN works best for classification. As a classifier, it can be used in two ways: i) Transferred Learning where pre-trained CNN is applied for feature extraction or classification task; ii) Specific convolutional networks. The two techniques to employ pre-trained architectures in a project for feature extraction are feature extraction and fine tuning [25], [26]. The proposed model is built on the Inception v3 architecture. Inception modules are made up of a variety of convolutions with variable kernel sizes and a max pooling layer [26].

After that VGG16 and Inception V3 will be used to extract the features that will be used as input data to the Neural Networks and random forest algorithms, which are considered as classifiers in the implemented model. The model deals with multi-class classification problem with two categories (cancer, non-cancer). The optimizer that used with the model is called “adam optimizer” to decide the best learning rate. The proposed work steps for detecting lung cancer in an image are shown in Figure 4.

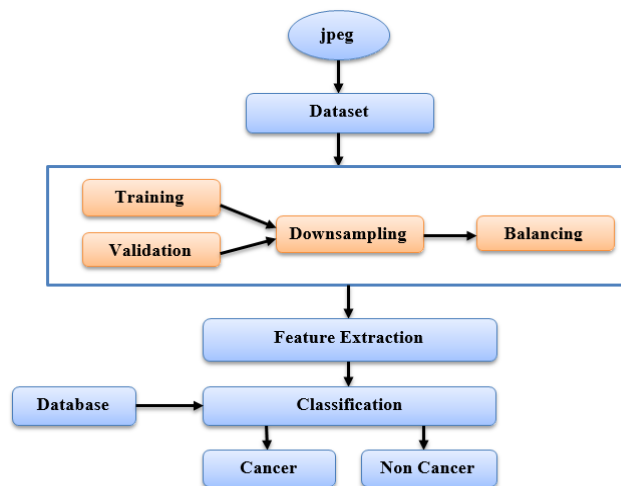


Figure 4. Model of detecting lung cancer in an image using deep learning (DL)

4. RESULTS AND DISCUSSION

In the first implementation, if the validation metrics are significantly worse than the training metrics in the initial implementation, the model is likely overfitting and does not generalize well from observed data to unseen data, which is called overfitting [27]. If, on the other hand, the validation measures are required more than the training metrics, then it indicates that the validation measurements are unsuitable. Table 3 shows the comparison of the results for different architectures.

Table 3. Comparison of results of different architectures

Models	Training Accuracy	Validation Accuracy	Comparison	Output
InceptionV3 + Neural Networks	91.31	87.46	Training Accuracy > Validation Accuracy	Underfitting
InceptionV3 + Random Forest	100	83.9	Training Accuracy > Validation Accuracy	Underfitting
VGG16 + Neural Networks	87.37	93.75	Training Accuracy < Validation Accuracy	Overfitting
CNN	96.43	97.1	Training Accuracy equal Validation Accuracy	Fits Well

Inception-v3: is a convolutional neural network architecture from the Inception family that makes several improvements including using Label Smoothing.

VGG-16: is a convolutional neural network that is 16 layers deep.

In the second implementation, the designed model of the CNN consists of three layers, that is convolutional layers, Max pooling layers, and drop out layers. Many measures have been used to test the performance of the model. To adjust the efficiency of the CNN model, accuracy, precision, recall, F1 score, and specificity measures are computed and presented in Table 4.

Table 4. Proposed CNN results

Performance Measure	Result
Accuracy	97.09%
Precision	96.89%
Recall	97.31%
F1 Score	97.09%
Specificity	96.88%

As shown from the performance measures that the CNN model trained the given dataset very well and gave good results. Thus, more an intelligent model can be provided to pathologists which helps them in discovering whether the patient has lung cancer or not. The model can be extended to involve additional layers, and other feature extractors can be added so that the model can perform better and result in accurate results. A 16-layer VGG, architecture, was being processed. To obtain the most accurate results, the input is handled through each of the 16 layers. The accuracy and possibility of better results can be increased by using convolutional layers.

5. CONCLUSION

In this work, the categorization of lung cancer images using the inceptionV3 convolutional neural network outperforms strategies that include random forest classifiers, SVM classifiers, and traditional image processing tools. From the previous results, it can be observed that all the major values are greater than 96%. This proves that the CNN model train from scratch classifies the given dataset very well if we just implement the model. We noticed that even with the use of open-source tools, which eliminate the need to write machine learning code from scratch, and a computer suite containing thousands of processors, it requires a massive amount of memory, possibly several terabytes, and it could take a month or more to determine which cellular features the team should direct the image-analysis software to look for. Furthermore, after optimizing and fine-tuning the parameters for each cell type, there was a need to modify the software to work across all cells. In order to make it easy for the end-user to use the system, the project could be extended to create a user interface, a website could be built as a user interface. The slide image should be uploaded on the interface, the user should click on the predict button, it'll take some time to predict the result. It's important to early detect cancer and treats it for patient recovery, so highly recommend that people between the ages of 50 and 80 with a smoking history should consider screening every year. It's quick, painless, and could save a life.




REFERENCES

- [1] P. de Groot and R. F. Munden, "Lung cancer epidemiology, risk factors, and prevention," *Radiologic Clinics of North America*, vol. 50, no. 5, pp. 863–876, 2012, doi: 10.1016/j.rcl.2012.06.006.
- [2] S. S. Ramalingam, T. K. Owonikoko, and F. R. Khuri, "Lung cancer: New biological insights and recent therapeutic advances," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 91–112, Mar. 2011, doi: 10.3322/caac.20102.
- [3] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung cancer statistics," in *Lung Cancer and Personalized Medicine*, vol. 893, 2016, pp. 1–19.
- [4] E. A. Zang and E. L. Wynder, "Differences in lung cancer risk between men and women: examination of the evidence," *JNCI Journal of the National Cancer Institute*, vol. 88, no. 3–4, pp. 183–192, Feb. 1996, doi: 10.1093/jnci/88.3-4.183.
- [5] M. Mustafa, A. J. Azizi, E. Ilzam, A. Nazirah, S. Sharifa, and S. Abbas, "Lung cancer: risk factors, management, and prognosis," *IOSR Journal of Dental and Medical Sciences*, vol. 15, no. 10, pp. 94–101, Oct. 2016, doi: 10.9790/0853-15100494101.
- [6] X. Y. Jin, Y. C. Zhang, and Q. L. Jin, "Pulmonary nodule detection based on CT images using convolution neural network," in *Proceedings - 2016 9th International Symposium on Computational Intelligence and Design, ISCID 2016*, Dec. 2016, vol. 1, pp. 202–204, doi: 10.1109/ISCID.2016.1053.
- [7] M. S. Al-Tarawneh, "Lung cancer detection using image processing techniques," *Leonardo Electronic Journal of Practices and Technologies*, vol. 11, no. 20, pp. 147–158, 2012.
- [8] J. J. Chabon *et al.*, "Integrating genomic features for non-invasive early lung cancer detection," *Nature*, vol. 580, no. 7802, pp. 245–251, Apr. 2020, doi: 10.1038/s41586-020-2140-0.
- [9] I. M. Nasser and S. S. Abu-Naser, "Lung cancer detection using artificial neural network," *International Journal of Engineering and Information Systems (IJEAIS)*, vol. 3, no. 3, pp. 17–23, 2019.
- [10] B. Chitradevi and P. Srimathi, "An overview on image processing techniques," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 11, pp. 6466–6472, 2014.
- [11] J. Schmidhuber, "Deep learning in neural networks - an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.




- [12] H. Yang, H. Yu, and G. Wang, "Deep learning for the classification of lung nodules," *arxivpreprints*, Nov. 2016, [Online]. Available: <http://arxiv.org/abs/1611.06651>.
- [13] M. Coccia, "Artificial intelligence technology in cancer imaging: Clinical challenges for detection of lung and breast cancer," *Journal of Social and Administrative Sciences www.kspjournals.org*, vol. 6, no. 2, pp. 82–98, 2019, [Online]. Available: www.kspjournals.org.
- [14] H. Malik, T. Anees, and Mui-zzud-din, "BDCNet: multi-classification convolutional neural network model for classification of COVID-19, pneumonia, and lung cancer from chest radiographs," *Multimedia Systems*, vol. 28, no. 3, pp. 815–829, Jun. 2022, doi: 10.1007/s00530-021-00878-3.
- [15] T. Dehkharghanian *et al.*, "Biased data, biased AI: deep networks predict the acquisition site of TCGA Images," *Research Square*, pp. 1–17, 2021, doi: <https://doi.org/10.21203/rs.3.rs-943804/v1>.
- [16] N. Kalaivani, N. Manimaran, S. Sophia, and D. D. Devi, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," in *IOP conference series: materials science and engineering*, 2020, vol. 994, no. 1, p. 12026.
- [17] I. Horace H-S, *Digital image processing and computer vision*, John Wiley & Sons, UK, vol. 8, no. 3, 1990.
- [18] P. B. Kutade and P. S. A. Bhalotra, "A survey on various approaches of image steganography," *International Journal of Computer Applications*, vol. 109, no. 3, pp. 1–5, Jan. 2015, doi: 10.5120/19165-0620.
- [19] D. Ganguly, S. Chakraborty, M. Balitanas, and T. Kim, "Medical imaging: a review," in *Communications in Computer and Information Science*, vol. 78 CCIS, 2010, pp. 504–516.
- [20] K. V. Gomathi and R. Lotus, "Digital image compression techniques," *International Journal of Research in Engineering and Technology*, vol. 03, no. 10, pp. 285–290, Oct. 2014, doi: 10.15623/ijret.2014.0310044.
- [21] A. Roman-Gonzalez and K. Asalde-Alvarez, "Image processing by compression: An overview," *Lecture Notes in Engineering and Computer Science*, vol. 1, pp. 650–654, 2012.
- [22] M. Ramprasath, M. V. Anand, and S. Hariharan, "Image classification using convolutional neural networks," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 17, pp. 1307–1319, 2018.
- [23] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: a brief review," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, 2018, doi: 10.1155/2018/7068349.
- [24] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [25] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, pp. 1043–1048, doi: 10.1109/ICDM.2008.28.
- [26] M. Ali, D.-H. Son, S.-H. Kang, and S.-R. Nam, "An accurate CT saturation classification using a deep learning approach based on unsupervised feature extraction and supervised fine-tuning strategy," *Energies*, vol. 10, no. 11, p. 1830, Nov. 2017, doi: 10.3390/en10111830.
- [27] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

BIOGRAPHIES OF AUTHORS






Asraa A. Abd Al-Ameer    is currently an Assistant Lecturer at Al-Zahraa University for Women, Karbala, Iraq. She received her Bachelor's degree in Information Technology from Babylon University, Babylon, Iraq, in 2016, and her Master's degree in Information Technology from Babylon University, Babylon, Iraq, in 2019. She is currently a Ph.D student at the department of information network, college of information technology, Babylon University, Babylon, Iraq. Her current research is focused on aspects that include SDN, security, network security, privacy, machine learning, and deep learning. She can be contacted at email: asraa.abd.alhussien@alzahraa.edu.iq.



Ghufran Abdulameer Hussien    is currently an Assistant Lecturer at Al-Mustaqbal University College, Babel, Iraq. She received her Bachelor's degree in Information Technology from Babylon University, Babylon, Iraq, in 2015, and her Master's degree in Information Technology from Babylon University, Babylon, Iraq, in 2018. She is currently a Ph.D student at the department of information network, college of information technology, Babylon University, Babylon, Iraq. She can be contacted at email: ghufran_abdulameer@mustaqbal-college.edu.iq.



Hajer. A. Al Ameri    is currently an Assistant Lecturer at Al-Zahraa University for Women, Karbala, Iraq. She received her Bachelor's degree in computer science from Karbala University, Karbala, Iraq, in 2016, and her Master's degree in Information Technology from Babylon University, Babylon, Iraq, in 2020. Her research Master is focused on aspects that include image processing, data mining. She can be contacted at email: hajer.alamery@alzahraa.edu.iq.