Received: 25 November 2021 | *Revised: 04 January 2022* | *Accepted: 12 January 2022*
DOI: 10.37917/ijeee.18.1.7 *Vol. 18| Issue 1| June 2022*

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Semantic Segmentation of Aerial Images Using U-Net Architecture

Sarah Kamel Hussein[1*], Khawla Hussein Ali[2]
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, Iraq

Correspondence
* Sarah Kamel Hussein
College of Education for Pure Sciences,
Education College for Pure Sciences,
University of Basrah, Basrah, Iraq
Email: cepsm510003@avicenna.uobasrah.edu.iq
khawla.ali@uobasrah.edu.iq

Abstract

Arial images are very high resolution. The automation for map generation and semantic segmentation of aerial images are challenging problems in semantic segmentation. The semantic segmentation process does not give us precise details of the remote sensing images due to the low resolution of the aerial images. Hence, we propose an algorithm U-Net Architecture to solve this problem. It is classified into two paths. The compression path (also called: the encoder) is the first path and is used to capture the image's context. The encoder is just a convolutional and maximal pooling layer stack. The symmetric expanding path (also called: the decoder) is the second path, which is used to enable exact localization by transposed convolutions. This task is commonly referred to as dense prediction, which is completely connected to each other and also with the former neurons which gives rise to dense layers. Thus it is an end-to-end fully convolutional network (FCN), i.e. it only contains convolutional layers and does not contain any dense layer because of which it can accept images of any size. The performance of the model will be evaluated by improving the image using the proposed method U-NET and obtaining an improved image by measuring the accuracy compared with the value of accuracy with previous methods.

KEYWORDS: U-Net, Deep Learning, Image Processing, Semantic Segmentation, Convolutional Neural Network (CNN), Feature Extraction.

I. INTRODUCTION

Deep learning is one of the most controversial technologies at this time, as its energy and ability to simulate the human mind is very strange and frightening, deep learning is a technology invented by humans in order to try to imitate the way the human mind works, deep learning tries to simulate the mind human being in all his abilities, of which; Seeing, understanding speech, composing it, hearing, and other powerful abilities that our human mind possesses and is not rivaled by anything else. Not only did it stop at this point, but scientists have studied the human brain and how it works in order to design algorithms and programs capable of simulating it, and for this reason, we find that these algorithms are inspired by human medical and neurological studies and try as much as possible to imitate them, but in computer ways not biological [1].

Deep learning is a section or part of machine learning that is part of the larger science called Artificial Intelligence. Neurons or Neural Networks have been replaced by a computer to become a perceptron or Artificial neural network, of which we now have many types such as: Convolutional Neural Network, or Recurrent Neural Network [2].

Deep learning is a relatively recent term; As we did not really care about it until we had a lot of data during the modern technological revolution, and it is necessary in order to make difficult decisions through big data, without any human intervention or limiting the data or properties, deep learning as we will see is the one who takes care of all these Things are just like the human mind. Deep learning is a new science to a large extent, but it has roots in human knowledge, and it has stages through which it has developed since the forties of the last century until the present day, these stages are the stage of cybernetics: which extended from the forties of the twentieth century until the nineties, The stage of Connectionism, which was in the nineties and eighties of the last century, The stage of Deep Learning: It is what we now know as deep learning, and it started from 2006 and extended until now [3][4].

Semantic image segmentation is a form of dense segmentation task in computer vision in which the model outputs a dense feature map of the input RGB image with the same dimensions (height and width) as the input image. The

output feature map consists of many channels where there are classes for each pixel to predict from [5][6]. So, Semantic Segmentation is assigned class to each pixel in the image to a class label [7].

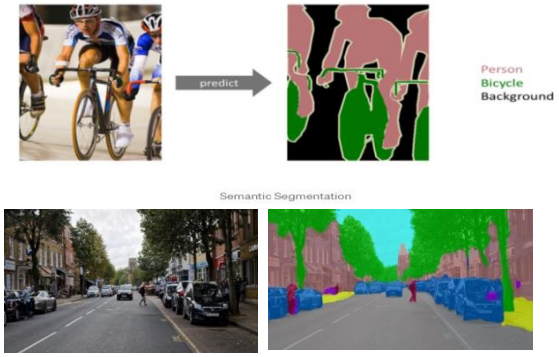

Fig. 1: Show is Semantic Segmentation

The degree of complexity of the issue varies from image to image, depending on the job and according to the degree of complexity that you train the neural network. The purpose of semantic image segmentation is to produce more than just labels and bounding box parameters as expected output [8]. The output is a high-resolution image (usually the same size as the input) with each pixel classified into a different class.

Computer vision is a subfield of computer science that tries to create intelligent applications that can comprehend the information of images in the same way that humans can. Where picture data can be in a variety of formats, including sequential images (video), scenes captured by several cameras, and data with multiple dimensions obtained from a medical imaging device [9][10]:

- Recognition: one or some of the objects that were previously marked to the computer are recognized, often with their different positions or different camera angles.
- Select: select a single match for the defined object. For example: identifying the face of a particular person or identifying the fingerprint of a particular person or a vehicle of a particular type.
- Investigation: The image data is searched to find a specific object.
  Example: investigating the presence of diseased cells in a medical picture, investigating the presence of a car on a highway.
- Image retrieval based on content: Images stored in a specific database are retrieved based on the content and concepts similar to the query from within the database. One of the most popular query methods in CBIR systems is the Query Image query, where an image is entered and the output is a set of similar images.
- Contributions aerial images containing 72 satellite images of Dubai were first applied with the proposed method U-Net to find out the efficiency of U-Net for semantic segmentation, and it was compared with CNN. Efficient results were obtained in this field and this was proven with CNN.

II. Related Works

The most important problem with computer vision is what is called semantic segmentation [11]. It's widely utilized in image processing to get a full picture of a situation. Because of the rapid advancement In recent years, deep learning architectures have been used to solve the majority of semantic segmentation difficulties [12]. Convolutional neural networks are the most efficient and accurate deep learning designs.

Since 2012, various Convolutional Neural Network based architectures like VGG16 [13], ResNet [14], Mobile Net [15], U-Net, as well as the recently developed Efficient Net, have evolved and set standards in picture classification. The application of these CNNs as feature extractors has lately made substantial progress in the field of semantic segmentation. Fully Convolutional Neural Networks were used in one of the first attempts at semantic segmentation using CNN (FCN) [16]. The loss of spatial information of small and thin objects is hampered by the CNNs' progressive down sampling of the original image resolution. The notion of dilated convolution was invented in to solve this problem [17] to increase the resolution of the feature map while maintaining the receptive field of the neuron. Dilated residual networks were proposed by Yu et al., which solved the problem of gridding artifacts [18].

III. Data Sets Preparing

Computer vision systems vary greatly, ranging from large and complex systems that perform general and comprehensive tasks, and between small systems that perform simple and customized tasks. But most computer vision systems mainly include the following components:

- Image acquisition: From the image sensors we get the image used, these include many cameras with light sensors, distance sensors, radiographic devices, radar, ultrasound cameras, and others. Depending on the type of sensor, the resulting image can be 2D, 3D, or a series of sequential images. The value of each pixel in the image depends on one or more light intensity levels (gray scale images or color images) and can indicate many physical measurements such as absorption, reflection of electromagnetic waves or distance.
- Pre-processes: It is necessary to ensure that the data provides the specific data onto the algorithm before applying the computer vision algorithm to the image and then obtaining the required information, including resetting the resolution and clarity to ensure the correctness of the image coordinates system. Second, Minimize noise in order to ensure that the sensor is not giving any false information. Third, Increase the variance

in order to ensure that the desired information will be obtainable.

- Feature extraction: Image features at different resolutions are obtained from the same image data. These landmarks are classified into :Global features such as color and shape.

It is possible to get more complex features related to colors and shapes in the image.

A. *Segmentation Architecture*

Image segmentation is an important stage of digital image processing, which is the process of the segment images of an area (image processing) interconnected and homogeneous regions according to a specific criterion for each color. The union of these regions should result in a reconstruction of the original image. Slicing is an important stage that allows extracting qualitative information about the image, as it provides a high-level description, as each region is linked to its neighboring regions within a network of nodes in which each node represents a region in the image. This node carries a card containing qualitative information about the region such as its size, color, shape, Orientation, and the brackets that connect the nodes can be marked with information about the relationship between adjacent areas, such as for example, an area whose content is in another, or it is below or above it, and so on. The level of complexity in network configuration varies depending on the slicing technique used [17].

B. *Fully convolutional networks (FCN)*

The FCN, a variant of the CNN, was one of the most significant improvements in the process of image segmentation [19]. The FCN differs from a traditional CNN in that the completely linked layers at the end of the CNN are converted into convolution layers. This results in a network that computes a nonlinear filter for each layer's output vectors. As a result, the completed network can function on inputs of any size and produce outputs with the same spatial dimensions. The classification network may now generate a heat map of the selected item class. Adding layers and a spatial loss to the network results in an efficient machine for end-to-end dense learning[20].

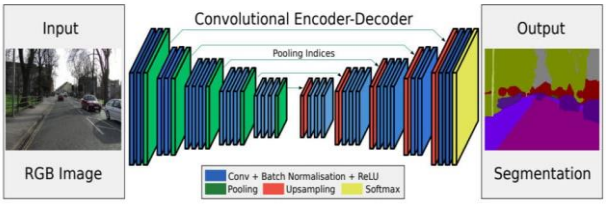

Fig. 2: Convolutional Encoder-Decoder for U-Net Architecture.

IV. PROPOSED METHOD

A. *U-Net architecture*

The U-Net was developed by Ronneberger et al. [4]. Training this network relies on data intensively to use the suggested images efficiently. Contributions aerial images containing 72 satellite images of Dubai were initially used to test the efficiency of the proposed approach U-Net for semantic segmentation, and the results were compared to CNN. In this field, effective outcomes were obtained, as demonstrated by CNN. The architecture consists of:

1. A feature map encoder that shortens the input image.
2. A decoder that uses deconvolutional layers learning to enlarge the feature map to the size of the input image.
3. The U-Net architecture's key contribution is the creation of shortcut connections. We noticed in FCN that when we down sample a picture as part of the encoder, we lose so, it captures finer information whilst also keeping the computation at low.

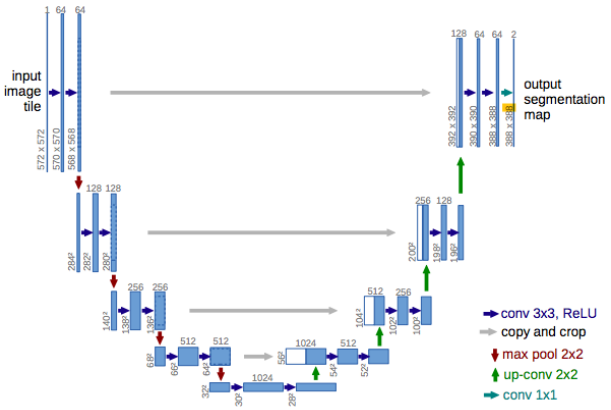

Fig. 3: U-Net architecture.

The architecture contains two paths: First: encoder (contraction path) which is used to capture the context of the image. The encoder is just a traditional stack of convolutional and max pooling layers [21][22]. Second: decoder (symmetric expanding path) which is used to enable precise localization using transposed convolutions.

B. *Loss Function for Segmentation*

It is the effect of the loss function on the hash output results. Where three different loss functions are used in the training procedure. So let's say p is the output value of each pixel in the image. Then we define the studied loss functions in this case as follows:

- Cross-Entropy Loss

Loss of the log where it calculates the logarithmic value of the output, i.e. for each pixel in the output tensor (and because we are talking about images).The term alpha is a measure of overweight for different classes and is a means of balancing the loss for unbalanced classes. In Equation 1 we show the final weighted entropy loss equation.

$$CE = -\alpha_c . \log \hat{y}_i \quad (1)$$

- *Focal Loss*

Focal loss is the best solution to the problem of unbalanced data set. Where he adds another label to reduce the impact of correct predictions and focus on incorrect examples. Gamma is a hyper parameter that determines how strong this reduction is. This loss affects network training on the unbalanced data set and can improve segmentation results.

$$FL = -\alpha_C(1-\hat{y}_i)^y.\log\hat{y}_i \quad (2)$$

- *IoU Loss (Jacquard index)*

Loss of IoU It is the last option for unbalanced segmentation and has fewer hyperparameters than other types. It can be explained in equation (3).

$$IOU = \frac{\text{Area of overlap}}{\text{Areaof Union}} \quad (3)$$

The above shows that the filter is the overlap between the masks of expected and ground truth, and the denominator is the union between them. The IoU is calculated by dividing the first by the second, with values closer to one indicating more accurate predictions.

The purpose of the optimization is to get a more accurate IoU of the image and has a value between 0 and 1, so the loss function is defined as:

$$\text{IOU Loss} = 1 - IOU \quad (4)$$

We trained U-Net with all three loss functions of the mentioned data set. As only 65 images were used for training and 7 images for verification, so we cannot expect perfect results. But this number of data is sufficient for the purpose.

V. Results And Analyses

In this paper, we review the problem of semantic segmentation on unbalanced type binary masks. Focal loss and mIoU are presented as loss functions for tuning network parameters. Finally, we train the U-Net implemented in PyTorch on the semantic segmentation method using aerial images.

A. Dataset

The dataset used here is a semantic segmentation set of aerial images containing 72 satellite images of Dubai, United Arab Emirates, divided into 6 categories. Classes include water, land, roads, buildings, plants, and the unnamed.

```
filename = ''/content/drive/MyDrive/semantic segmentation dataset/classes.json''
  shutil.unpack_archive(filename, extract_dir, archive_format)
self.BGR_classes = {'Water' : [ 41, 169, 226],
'Land' : [246,  41, 132],
'Road' : [228, 193, 110],
'Building' : [152,  16,  60],
'Vegetation' : [ 58, 221, 254],
'Unlabeled' : [155, 155, 155]} # in BGR
 self.bin_classes = ['Water','Land','Road','Building','Vegetation', 'Unlabeled']
```

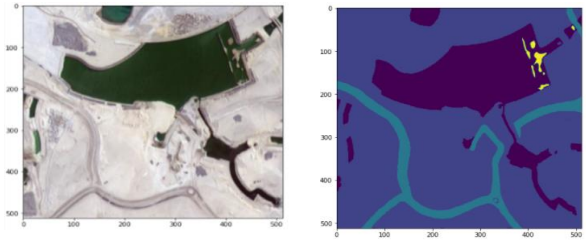

Fig. 5: Sample of Dataset

B. Loss Results

1) Cross Entropy Loss

As you can see, cross-entropy has a problem segmenting small areas and has the worst performance among these loss functions.

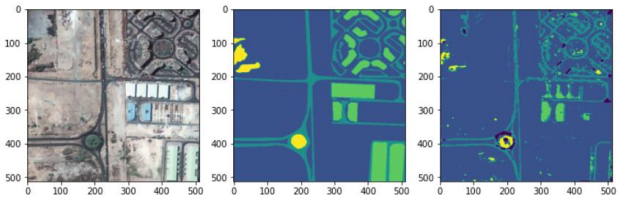

Fig. 6: Segmentation results by entropy loss.

As we note, cross-entropy has a small-space segmentation problem and has the worst performance among these loss functions.

2) Focal loss

Focal loss can achieve better results, especially in small regions, but it still needs some hyper parameter tuning through trial and error.

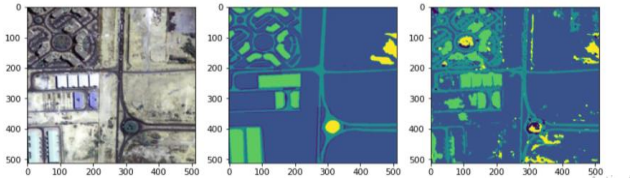

Fig. 7: Segmentation results using Focal loss

3) IOU (Loss)

Finally, we can see that IoU loss also does a great job in segmentation, both for small and large areas.

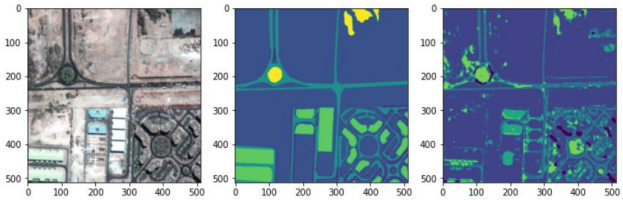

Fig. 8: Segmentation results using IOU

C. Training the Model

1) Compare U-Net and CNN

The semantic segmentation purpose is used to label all pixels in an image with an appropriate class. Original U-Net paper dimension is 572X572X3. Here in this work the initial image dimension used is 128X128X3. All models

were trained for 100 epochs with 500 patches of 72 images where used 7 image in testing and 65 image used in training for “Aerial Segmentation”.
Following hyper parameters are used number-channel=3 , number-classes= 6 ,optimizer is Adam , Learning rate is 001, Batch size is 17 .

TABLE I
Results Training the model (U-Net)

loss	accuracy	Val-loss	Val-accuracy
0.54446	0.76	0.52749	0.77
0.52756	0.77	0.52701	0.77
0.53875	0.76	0.53456	0.76
0.52270	0.77	0.50296	0.78
0.55374	0.76	0.53489	0.77

TABLE II
Results Training the model (CNN)

loss	accuracy	Val-loss	Val-accuracy
0.4892	0.9613	0.4533	0.9632
0.4539	0.9659	0.4502	0.9720
0.4397	0.9676	0.4302	0.9719
0.4307	0.9686	0.4302	0.9668
0.4227	0.9699	0.4556	0.9750

2) Evaluate the model

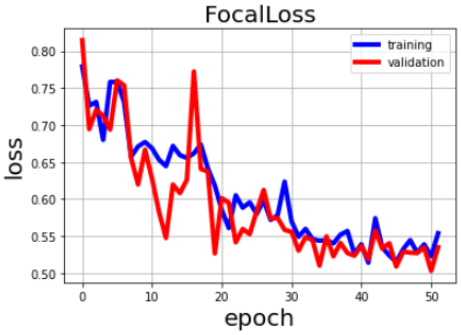

Fig. 9: Focal loss U-Net model

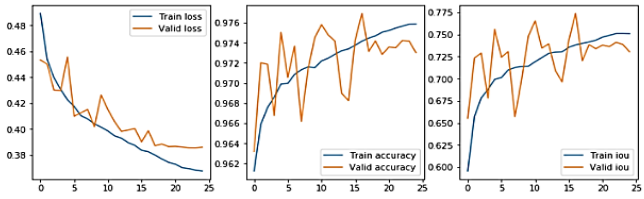

Fig.10: Training CNN Model

D. Testing the Model

Important to note that there were only 65 images for training and 7 for validation, so we can’t expect great results. But this number of data is enough for our purpose.

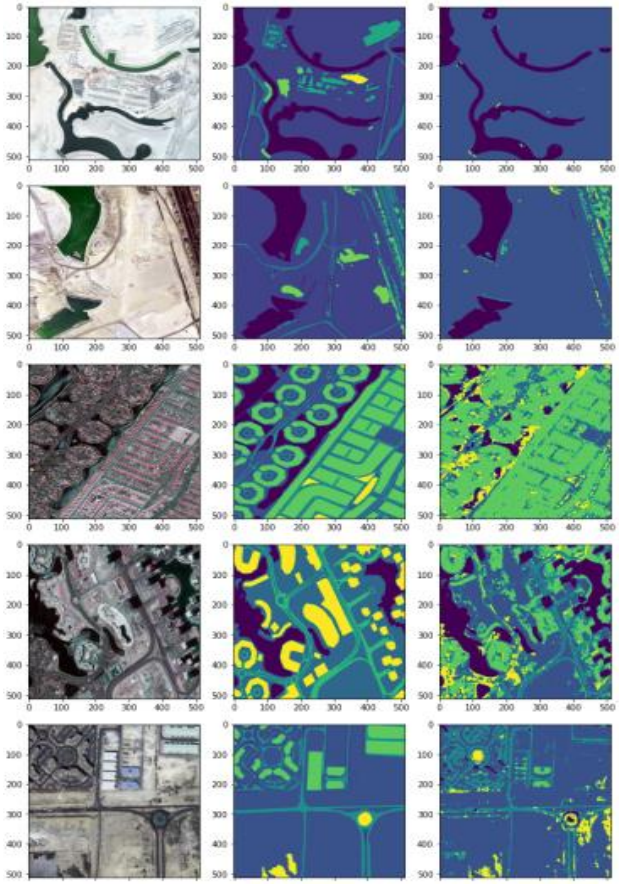

Fig. 11: Cross-entropy on the left, focal loss in the middle, and IoU loss on the right.

VI. Conclusions

The proposed algorithm is U-Net, which has been applied to aerial images. It works to give different color to each category, and it is possible to assign a category to each pixel in the image, such as the label with the word car or plane, and this is called semantic. The process of adding the corresponding pixels is performed directly with the previous operations, which are very smart operations, so it differs from FCN. It has preserved its spatial information because it contains the copy & crop process. We show that such a network can be trained end-to-end from very few images and outperforms the prior best method (fully convolutional network) .Moreover, the network is fast. Segmentation of a 572x572 images taken less than a second on a recent GPU. As a result, we obtained a matrix of the same dimensions for the input image, so U-Net was applied to the aerial images, and the process of prediction of pixels in the border region was accurate and fast through the results that was applied by the pytorch library.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

REFERENCES

[1] K. H. Ali and T. Wang, "Learning features for action recognition and identity with deep belief networks," in *2014 International Conference on Audio, Language and Image Processing*, 2014, pp. 129–132.

[2] J. Marshall, "Learning with technology," *Evid. that Technol. can, does, Support Learn.*, 2002.

[3] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016, pp. 565–571.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[5] Hameed Abdul-Kareem Younis, Marwa Kamel Hussien, "Multi-Frame Video Compression Scheme Using Three Step Search (TSS) Matching Algorithm", The islamic college university journal, Issue 29, pp. 49-68, 2014.

[6] M. K. Hussien and H. A.-K. Younis, "Wavelet-Based Video Compression System Using Diamond Search (DS) Matching Algorithm," *J. kerbala Univ.*, vol. 1, pp. 249–258, 2013.

[7] J. Fang, Q. Zhou, and S. Wang, "Segmentation Technology of Nucleus Image Based on U-Net Network," *Sci. Program.*, vol. 2021, 2021.

[8] N. Darapaneni, A. Jagannathan, V. Natarajan, G. V. Swaminathan, S. Subramanian, and A. R. Paduri, "Semantic Segmentation of Solar PV Panels and Wind Turbines in Satellite Images Using U-Net," in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 7–12.

[9] H. A. Younis and M. K. Hussein, "Adaptive Video Compression Technique Based on Wavelet Transform and NTSS Matching Algorithm NTSS", Journal of College of Education for Pure Sciences, vol. 4, no. 1, pp. 203–214.

[10] H. A. Ali and A. J. J. M. K. Hussein, "Secure Data Hiding Technique Using Video Steganography," *Des. Eng.*, pp. 6208–6217, 2021.

[11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv Prepr. arXiv1704.06857*, 2017.

[12] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 2, pp. 87–93, 2018.

[13] B. J. Bhatkalkar, D. R. Reddy, S. Prabhu, and S. V Bhandary, "Improving the performance of convolutional neural network for the segmentation of optic disc in fundus images using attention gates and conditional random fields," *IEEE Access*, vol. 8, pp. 29299–29310, 2020.

[14] C. Huang, H. Han, Q. Yao, S. Zhu, and S. K. Zhou, "3D U2-Net: a 3D universal U-Net for multi-domain medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 291–299.

[15] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge," in *International MICCAI Brainlesion Workshop*, 2017, pp. 287–297.

[16] M. K. Hussein, K. R. Hassan, and H. M. Al-Mashhadi, "The quality of image encryption techniques by reasoned logic," *TELKOMNIKA*, vol. 18, no. 6, pp. 2992–2998, 2020.

[17] M. K. Hussein, A. J. Jalil, and A. Alhijaj, "Face Recognition Using The Basic Components Analysis Algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 928, no. 3, p. 32010.

[18] M. K. Hussein, "The optimum encryption method for image compressed by AES," *GSJ*, vol. 8, no. 4, 2020.

[19] H. A. Ali, A. J. Jalil, and M. K. Hussein, "Vernam Encryption and Steganography of a Number of Images in the Digital Video," in *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2021, pp. 1–6.

[20] K. Hussein, E. Barges, and N. Jameel, "Security issues in wireless sensor networks," *J. Multi-Disciplinary Eng. Sci. Stud.*, vol. 3, no. 6, pp. 1798–1800, 2017.

[21] A. Danti and G. R. Manjula, "Secured data hiding of invariant sized secrete image based on discrete and hybrid wavelet transform," in *2012 IEEE International Conference on Computational Intelligence and Computing Research*, 2012, pp. 1–6.

[22] A. Z. Atiyah and K. H. Ali, "Brain MRI Images Segmentation Based on U-Net Architecture", Iraqi Journal for Electrical and Electronic Engineering, pp. 21-27, 2021. DOI: 10.37917/ijeee.18.1.3