Mesopotamian journal of Big Data
Vol. **(2024)**, 2024, **pp.** 185–198
DOI: https://doi.org/10.58496/MJBD/2024/013 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Deep Learning Approaches for Gender Classification from Facial Images

Mustafa Abdulfattah Habeeb[1], Yahya Layth Khaleel[1], Reem D. Ismail[1], Z.T.Al-Qaysi[1], Fatimah N. Ameen [2,*]

[1] *Department of Computer Science, Computer Science and Mathematics College, Tikrit University, Tikrit 34001, Iraq*

[2] *Institute of Automation and Info-Communication, Faculty of Mechanical Engineering and Informatics,University of Miskolc, Miskolc, Hungary*

ARTICLE INFO

Article History
Received 17 Jun 2024
Accepted 11 Sep 2024
Published 11 Oct 2024

Keywords
Deep Learning
Gender Classification
Facial Images
Classification models
Man Image
Woman Image

ABSTRACT

Gender recognition on the facial level is considered one of the most important technologies that finds use in such fields as a personalized marketing plan, safe systems of authentication, and effective human-computer interfaces. However, it has the following challenges; variation of lighting, facial movement, and ethnic/age face images. AI and DL has been improving on the effectiveness, flexibility, and speed of the gender classification system. AI enables complex and automatic feature learning in Data, while DL is tailored for handle variants in vision-based data. In this paper, we evaluated several architectures including Efficient Net_B2, ResNet50, ResNet18, and Lightning whilst determining the performance of the architectures in gender classification tasks. Self-assessment criteria included accuracy, precision, recall, and the F1-score. As for the performance, we found that ResNet18 had the highest scores on all the metrics, with the validation accuracy of above 98%, closely accompanied by the ResNet50 that, although it performed well as well, needed more epochs for convergence. The implications of this study for the development of future work in the gender classification technology include the discovery of ethnical, dependable, and effective techniques. Through the consideration of the state of the art and case studies, stakeholders can optimise the efficacy and the accountability of such systems, and thus support societal gains as a result of the improvement in technology.

1. INTRODUCTION

In the last few years, with the growth in the number of images and the wide use of machine learning systems, gender classification from facial images has become an attractive and challenging task [1, 2]. This task has gained a lot of attention from several researchers due to its various applications, ranging from human-computer interface to intelligent video surveillance, forensic science, security control, and entertainment [3, 4]. Facial images provide many semantic clues, visual information, domain-specific features, and important visual characteristics that can be used for many applications, where gender classification is one of the most important among them [5, 6]. There are several approaches for gender classification such as plain machine learning techniques and deep learning techniques [7-9]. Recently, deep learning techniques have shown their significant success in many applications due to the flexibility and ease of use [10], demonstrating comparable, or even superior, classification and recognition accuracies on several tasks [11, 12].

There are several deep learning types such as Deep Belief Networks (DBN) [13], Deep Autoencoders (DAE) [14], deep Convolutional Neural Networks (CNN) [15, 16], deep Neural Architecture Search (NAS) [17], Multilayer Perceptron (MLP) [18], Generative Adversarial Networks (GAN) [19], and Restricted Boltzmann Machines (RBM) [20], among others, where the choice of the model depends on the domain, the problem complexity and the data type. Deep learning models introduce some very good properties to the model's classifiers, such as allowing the deep network to act as a feature extraction operator and work seamlessly with a single large annotated dataset or with a supervised fine-tune learning technique [21-23]. Moreover, deep learning can support the essential level of invariance to the domain data and domain generalization, not only the invariance to basic image operations such as rotation and scaling [24, 25]. Therefore, these classifiers can be used in several areas to reach many goals.

[*] *Corresponding author. Email: ameen.fatima.nadhim@student.uni-miskolc.hu*

There are various reasons for improving gender classification techniques, with the possibilities for using the results in individualized marketing, enhancing security and other parameters of customized authentication systems for a given user, and more flexible interaction between human and computers among the interests [26, 27]. The mathematical algorithms enable systems to determine the gender of an individual from the facial images and, therefore, provide the services that match the clients' requirements more closely and meet their needs better, and also make the work more effective [28, 29]. Nonetheless, face-based gender classification poses tremendous difficulties still. Some challenges include illuminations changes, associated with different facial expressions and influenced by light conditions, occlusions, etc algorithms [30]. Also, the performance of these systems across the ethnic and age diversity of users is also big matter [31]. Other issues which present significant challenges but comprise ethical dilemmas about the privacy and misuse of gender data also present remarkable challenges that require a special evaluation. It is possible to fight these challenges with the goal of creating accurate and non-bias gender identification systems [32-34].

Since gender classification systems have become prevalent in many sectors, any vulnerability that the gender classification systems may have to adversarial attacks becomes a threat [35, 36]. For instance, when it comes to security, an image can be manipulated in a way that it will make a system misclassify the gender of a person and this can lead to security risks or failures in identification [37, 38]. In marketing such attacks can alter the demographic data used in promotions, which leads to waste of resources or wrong direction of marketing efforts [39]. Furthermore, the concerns of ethical nature that can be associated with gender classification and their connection to privacy and misuse of gender data are compounded by the possibility of adversarial interference [40, 41].

The use of deep learning approaches for gender classification from facial images has gained significant attention in recent years due to its potential applications in various fields such as security, healthcare, and marketing. These approaches have shown promising results in accurately identifying gender based on facial features and have sparked interest in further research and development in this area. In this paper, we will explore the various deep learning techniques that have been utilized for gender classification from facial images. We will also discuss the challenges and future directions in this field. The introduction will provide an overview of the current state of gender classification from facial images. It will also discuss the importance of deep learning approaches in this field. The introduction will provide an overview of gender classification from facial images. It will also discuss the importance of deep learning approaches in this field. The introduction of deep learning methods has revolutionized the field of gender classification from facial images. These approaches have shown significant improvements in accuracy and robustness compared to traditional methods.
Our main objectives are to:

1. Discuss in detail the current techniques of face-based gender classification.
2. Identify the weakness and constrain of face-based gender classification technique.
3. Describe realistic examples of the usage of the gender classification system in various sectors.
4. Explain how ethical issues can be managed and show ideas on how to deal with prejudices and infringement of patient's privacy.
5. Design and evaluated four various techniques in dace gender classification (EfficientNet_B2, ResNet50, ResNet18, and Lightning).

Most current research on face-based gender classification systems would not be possible without two critical technologies: Artificial Intelligence (AI) [29, 42] and Deep Learning (DL) [43, 44]. In detail, AI helps important pattern matching and feature learning, enabling the systems to analyze facial images without much supervision and affect by lighting and other factors. DL is exceptionally well suited to the analysis of shapes because during training feature extraction for facial images is performed automatically. This leads to better performance, randomness such as face movements and obstruction, and the necessary flexibility. DL models [45, 46] are capable of handling big data, s, with pre-trained models for transfer learning, improving applicability and require a lot of computational power. Combined, the two provide more reliable, timely, and applicable gender categorization systems necessary for uses such as security, advertisement and interactive solutions [33, 47].

The importance of this study is in the possibility to contribute to the further studies and development of the gender classification methods. Thus, knowing major trends and applications, the stakeholders will be able to use these systems more efficiently and more responsibly. Overall, therefore, this paper aims at striving towards an improvement of effective gender identification solutions that may improve multiple technological applications.

2. RELATED WORKS

From the current face-based classification technologies, the slight improvement is seen particularly on the aspect of gender classification and each of them has his or her outstanding contribution to offer to the subject.

From the work of [48] the proposed facial recognition model is actually the integration of Seg-Net with the pre-existing model which uses Support Vector Machines (SVM) for gender and age estimation. This approach has proven effective in real-time applications, achieving high accuracy rates across various datasets: As to, so far as I know, 88.3% on Adience, 95.1% on IOG, 94.1% on FEI while 91.8% on proprietary datasets. In other words, the enhancement of using Seg-Net for segmentation and SVM for classification to make human-computer interaction and commercial contexts.

Towards this purpose, a new network, namely, gender CNN is devised for gender classification in [49] where IMDb-WIKI is used also. The goal of this work is to extract facial features from real-world facial data; dropout, which randomly drops-off some data, and data augmentation to prevent overfitting are also part of this. Presumably, the optimized architecture allows to reach 84.8% of the correspondent value in the basic architecture. New methods allow to classify the age group beyond the demonstrated sample, as illustrated by the classification of the age group that constitutes only 84.8% of the population with 52.3% accuracy. The CNN2ELM network design has been proposed to achieve only with all the best advanced training techniques integrated in the network design.

The authors in [50] tried hybridizing CNN with other machine known as Extreme Learning Machines (ELM). Since ELMs use the benefits of CNNs to perform classification, CNNs are used for feature extraction in the CNNs' benefit. In accordance with the experimental results above, this hybrid architecture obtains higher recognition rate and shorter time for gender classification from both MORPH-II and Adience Benchmark datasets, because of the detailed parameter setting and the overfitting issues.

Gender of human images in the form of 2D body images is classified using deep neural networks according to the elucidation provided in [51]. The end-to-end design prevents the inclusion of extra biometric features that are always useful in maintaining a high accurate gender classification system. It includes the pipeline for preprocessing the body images, as well as a 2D labeled data set. This just shows that gender determination from body images is as effective as from face images since the modified ResNet50 has higher accuracy.

Last, [52] has proposed an Ensemble CNN for the real-time gender recognition. For the gender classification, the system achieves 95% of accuracy On IMDB dataset. The Real-time model incorporates actual-time processing with the effects offered. The average response time for both single face image input and multiple face image input for the quantity and the quality of the response is about 0.5 seconds.

In total, these papers contribute to the improvement of the face-based classification in terms of time, quality and adaptability of the approach in other sets and real-world settings for the purpose of the insights and methodologies in further research and applications.

3. PROPOSED METHODOLOGY

The methodology for classifying gender through facial recognition involves several crucial stages. First, facial images are preprocessed to standardize and improve their quality. A meticulously curated and augmented dataset is critical for successful model training. DL models are trained on this dataset to identify distinguishing gender features. The model's performance is then assessed using validation and testing datasets to ensure its accuracy and reliability. Once trained, the model can accurately classify gender in new facial images, making it a valuable tool for various applications, including security and personalized marketing. Fig. 1 illustrates the steps in this methodology.

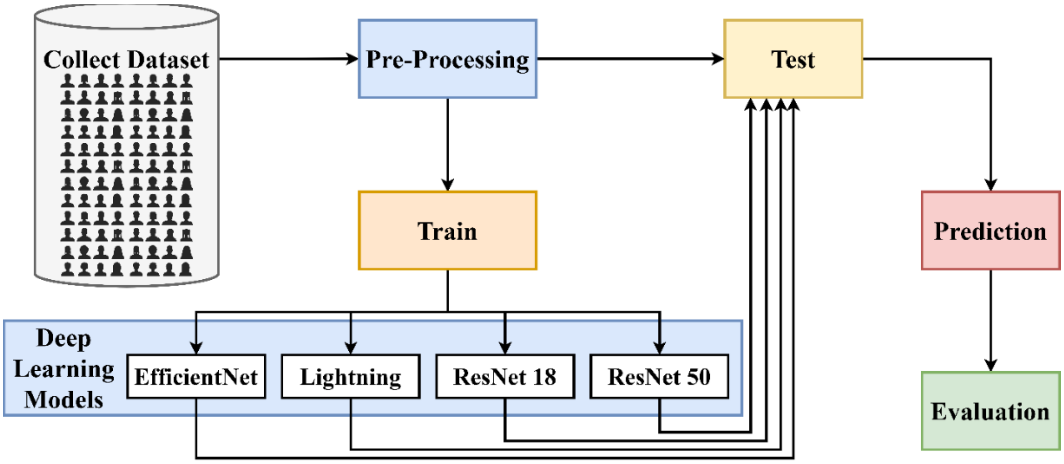

Fig. 1. Steps of Proposed Methodology

3.1 Dataset Exploration

The "Gender Classification: “Male and Female Images” is a quite balanced dataset which is developed specifically for gender recognition tasks, containing a set of a sufficient number of different male and female images downloaded from [19]. This dataset has 1,195 images and the classes distribution is as follows ; 54.1% men and 45.9% women which is shown in the Fig. 2, although it is also equally balanced for training and testing the DL models. It consists the image of people at different ages so it cover the big data necessary for building models as well as for learning more on appreciating differences in age. The dataset also contains different ethnicities—which is important for building models in one, ensuring that they are good for all. In addition, the photos are taken in different environments, indoor, and outdoor, which adds to the realism and variability needed to improve model performance.

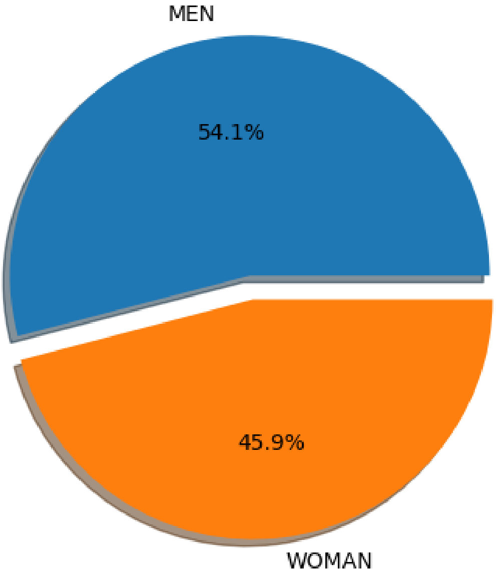

Fig. 2. Class distribution in dataset

Also, the dataset consists of the images of a persons captured from different angles, being serious or smiled, which means that the type of picture provided in the dataset is far from being an easy set of stereotype samples to be used as the material for gender classification. All images are well annotated by the binary labels ‘Male’ or ‘Female,’ so the dataset can be used for supervised learning tasks. The "Gender Recognition: Appropriate usage of the “Male and Female Images” dataset is in practicalities consisting of building the model for gender identification from facial features for assigning gender.

3.2 Data Preprocessing

Preprocessing is a crucial step in preparing the "Gender classification". The first among the steps is converting the set of vectors which is set in the format of the data frame to cater for the image data as well as its related labels. This is of great advantage, moreover, the ability to handle, analyze and the visualization of the dataset.
The next process is where the image is classified on the right group it belongs to which is either ‘Man Image’ or ‘Woman Image’ as classified in Table 1. That involves all the other steps are outlined above for the tasks of supervised learning the data should be labeled properly so the model can learn gender labels connected with the images.

TABLE 1. THE FIRST FIVE FIELDS IN A DATASET

ID	**Image_path**	**Label**
0	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN
1	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN
2	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN
3	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN
4	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN

Another very critical step you must take is rearranging the images to an equal dimension. By a standard way in dimensioning all images, the model works out the data in a much superior manner. This step helps in standardizing the size of images which otherwise can be a nuisance to the system when trying to train our model or improper portrayal of the image can influence the result [53].

To increase variability in the training dataset, data augmentation strategies are employed [54]. These transformations entail rotation, flipping, cropping as well as colour balance. Such variations artificially increase the size of the given dataset and the model gets to learn various possibilities as can be seen in Fig 3.

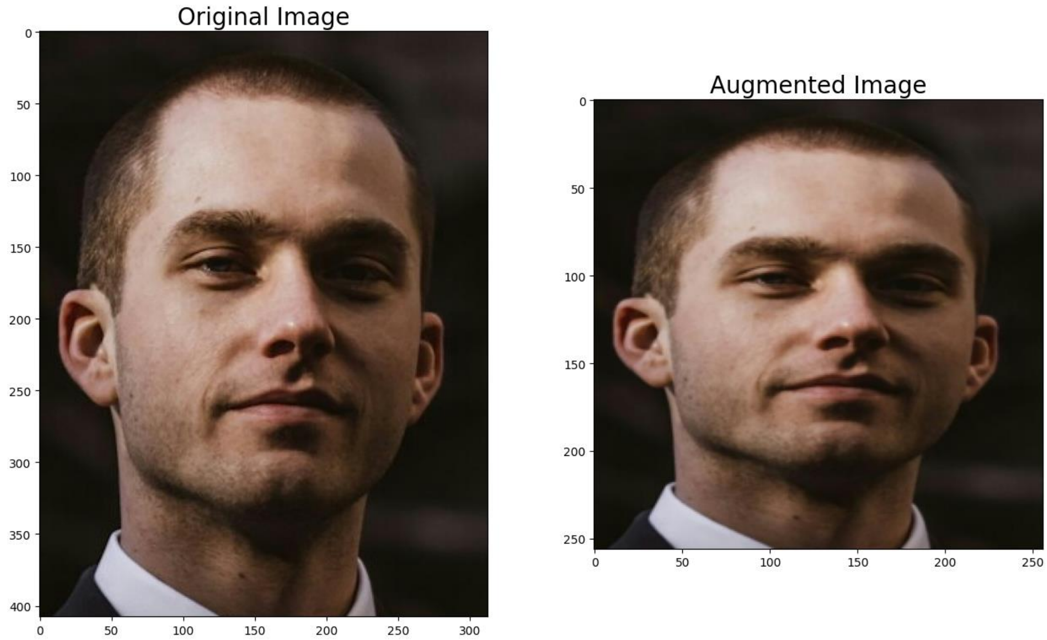

Fig. 3. Difference between Original Image and Augmented Image

Finally, the dataset is split into three subsets: training data (70%), validation data (15%) and test data (15%) as was represented in Fig. 4. The training set involves feeding the models with data that will enable them learn about the desired network, the validation set is used to improve the parameters’ model in order to avoid overfitting and the test set simply tests the performance of the model based on data which it has not been trained on. This splitting helps in the validation of the model and the model performance.

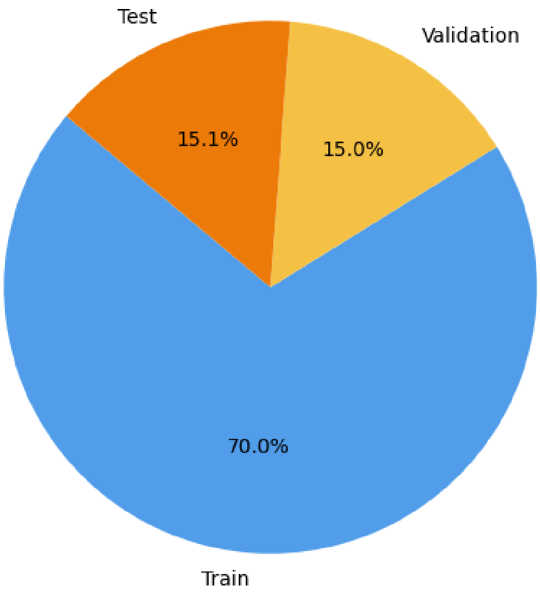

Fig. 4. Splitting dataset

3.3 Classification models

The process of constructing models for gender classification using DL involves the use of several architectures which are known to give the best results. This paper looks into the applicability of EfficientNet_B2, ResNet50, ResNet18, and Lightning for the correct gender prediction.

- EfficientNet_B2: is a latest and optimal CNN architecture efficient in terms of depth, width, and resolution of the network. They employ a compound scaling method, which scales all dimensions of the network by factor that is a set of scaling coefficients. This leads to a very efficient model which can perform as well or even better than the current best models with many fewer parameters or with fewer computational requirements [55]. It implies that in gender classification where facial features contribute a lot of factors, EfficientNet_B2 can capture them particularly on gender differences.
- ResNet50: is another popular DL model, known for its residual learning framework, which addresses the vanishing gradient problem in deep networks. With 50 layers, ResNet50 can learn complex patterns and features from the input images. The use of residual connections allows the model to maintain high accuracy and generalization by enabling the training of very deep networks [56]. This makes ResNet50 a powerful choice for gender classification tasks, as it can extract detailed and hierarchical features from facial images.
- ResNet18: is a shallower version of the ResNet family, and much faster and lighter than much deeper models such as ResNet50. has 18 layers, ResNet18 is deep enough to have favorable model complexity-compute characteristics. It inherits the benefits of residual learning, but is more appropriate for situations where there are limited computational resources [57]. For gender classification ResNet18 can still perform well accuracy by learning the important facial features with low-resource, thus it's a perfect model to be deployed on device with limited resources.
- Lightning: PyTorch Lightweight is a high-level API built on top of PyTŤorch that will allow you train DL models with ease. It is an incredibly bendy, high-speed for building and training models with standard industry practices to enable reproducibility and scale [58]. With Lightning, it's much simpler to develop and experiment with models of gender classification. It frees you to focus on the details of the model architecture and the hyperparameters as best you can without the minutiae of the training loop getting in the way.

Table 2 presents the model parameters and the optimization techniques applied for the four models applied in this research. Here, the various parameters which include the methods of selection and tuning of the parameters are enumerated as follows to each model to ensure that it is trained well to give high accuracy and reliability in gender classification.

TABLE 2. THE MAIN PARAMETERS AND OPTIMIZATION FOR MODELS

Models	Optimizer	Batch Size	Epochs	Learning Rate
EfficientNet_B2	Adam	32	15	0.001
ResNet50	Adam	64	75	0. 00001
ResNet18	SGD	32	25	0.001
Lightning	Adam	32	30	0.001

Also, Table 2 shows the most significant parameters and techniques used in the course of fine-tuning gender classification models' training. All the hyper parameters in each of the models are tuned in such a way that the time taken to train the model is in balance with the performance of the model. For EfficientNet_B2 the Adam optimizer is used this is an optimization algorithm that is adaptive in step size and very efficient. It was trained using a batch size of 32 and for 15 epochs and a learning rate of 0. 001. This makes it possible for EfficientNet_B2 to be able to learn from the data that is fed to it and at the same time, it uses lesser computation.

ResNet50 used batch size of 64 to assist in the deeper structure which the model will have as compared to the others. It also employs the Adam optimizer, the learning rate is much lower, it is 0. 00001. A smaller learning rate and using a higher number of epochs that are 75 allows ResNet50 to better train and fine-tune its parameters for the feature learning accuracy.

ResNet18 employs the Stochastic Gradient Descent (SGD) as the optimizer and this is one of the best techniques for training deep neural networks. In this work we have used a batch size of 32 and a learning rate of 0.001 and it is done for 25 epochs. This is good as far as the time it will take to train the model is concerned as well as the quality of the model that will be produced and therefore ResNet18 is appropriate for gender classification.

The Lightning framework uses the Adam optimizer it is an optimizer that adjusts the learning rate for each update of the gradients. The training of the model is done using a batch size of 32, for 30 full passes through the entire data set (30 epochs) at a learning rate of 0. 001. Lightning can be used to compare the effectiveness of different strategies of optimization when different classification techniques are used for the prediction of gender, therefore making the tool useful for the task at hand.
Therefore, it is possible to create highly effective and precise models of gender identification by these methods.

3.4 Model Evaluation

Evaluating advancements in face-based gender classification using DL involves a comprehensive analysis of the model's performance through various metrics and techniques. Accuracy is a primary metric, indicating the proportion of correctly classified instances out of the total instances. This metric is particularly useful for balanced datasets where the distribution of gender classes is relatively equal. Precision measures the ratio of true positive predictions to the total predicted positives, which is crucial in applications where the cost of false positives is high, such as in automated security systems. Recall, or sensitivity, quantifies the ratio of true positive predictions to the actual positives, making it important in contexts where it is critical to identify all instances of a specific gender, such as in demographic analysis for targeted marketing. The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure that is especially useful when there is a need to balance the trade-off between precision and recall [59, 60].

The confusion matrix is useful for a deeper analysis of the performance because it demonstrates the exact quantity of true positives, true negatives, false positives, and false negatives. From the confusion matrix, the researchers can easily identify the types of errors made by the model, for instance, whether the model has a tendency to misclassify one gender in place of the other [59]. This division helps in noting the weaknesses of the model with an aim of enhancing the model in the future.

It is also important to look at the fitness of the model to the data in the evaluation of the DL models. Learning curves that are, the plot of the performance of the model on the training and validation set at different epochs is very useful in identifying underfitting and overfitting. This is known as underfitting and such a model is overly simplistic and will present high levels of error when tested on both training and validation data. Overfitting on the other hand occurs when the model is complex and learns from noise in the training data hence producing low training error but high validation error [61]. This is the bias variance trade off and it is the key to distinguishing a good model from a great one in terms of its ability to perform well on new data. Table 3 shows the Terminology and Equation of these metrics.

TABLE 3. PERFORMANCE METRICS AND TERMINOLOGY FOR CLASSIFICATION MODELS

Terms	Description	
TP	Number of samples or the population of samples which was correctly segmented as malicious	
TN	The total number of samples that were correctly classified onto the benign class	
FP	Number of incorrectly classified samples as a result of the observed distinction of malice	
FN	Erroneously classified samples in minority class	
Confusion Matrix	TP	FP
	FN	TN
Accuracy	(TP+TN)/(TP+TN+FP+FN)	
Precision	TP/(TP+FP)	
Recall	TP/(TP+FN)	
F1-Score	2*((precision*recall)/ (precision +recall))	

4. RESULTS AND DISCUSSION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

4.1 Authors and Affiliations

The performance of various DL models for gender classification based on face recognition was evaluated using key metrics: accuracy, precision, recall, and F1-score. The results demonstrate significant differences in the effectiveness of each model. EfficientNet_B2 achieved an accuracy, precision, recall, and F1-score of 0.9333, indicating a high level of consistency across all metrics. This model shows reliable performance, with balanced precision and recall, reflecting its robustness in classifying both genders correctly.

ResNet50 performed slightly better with an accuracy of 0.9540, a precision of 0.9519, a recall of 0.9429, and an F1-score of 0.9474. This model exhibited a strong balance between precision and recall, with a slightly higher accuracy than EfficientNet_B2. The improved recall suggests that ResNet50 is particularly effective at identifying the minority class correctly, thereby reducing false negatives.

The highest performance was observed with ResNet18, which achieved an exceptional accuracy of 0.9969, precision of 0.9932, recall of 1.0000, and an F1-score of 0.9966. The perfect recall indicates that ResNet18 identified all instances of the positive class correctly, making no false negatives. This model's outstanding performance across all metrics suggests it is highly effective for gender classification tasks.

On the other hand, the Lightning model showed comparatively lower performance with an accuracy of 0.8250, precision of 0.8467, recall of 0.8250, and an F1-score of 0.8222. Although its precision is higher than its recall, indicating a relatively lower number of false positives, the overall lower metrics suggest that the Lightning model is less reliable than the other models evaluated. Table 4 summarizes the results of four models in this study.

TABLE 4. RESULTS OF CLASSIFICATION MODELS

Model	Accuracy	Precision	Recall	F1-score
EfficientNet_B2	0.9333	0.9333	0.9333	0.9333
ResNet50	0.9540	0.9519	0.9429	0.9474
ResNet18	0.9969	0.9932	1.0000	0.9966
Lightning	0.8250	0.8467	0.8250	0.8222

The results highlight the varying effectiveness of different DL models for gender classification based on facial recognition. ResNet18 emerged as the most effective model, achieving near-perfect performance across all evaluated metrics.

ResNet50 also demonstrated strong performance, though slightly below ResNet18. The balance between precision and recall in ResNet50 suggests that it is a reliable model, particularly in scenarios where reducing both false positives and false negatives is crucial. EfficientNet_B2, while performing well, did not reach the high levels of ResNet18 and ResNet50, but still offers a robust solution with consistent metrics.

The Lightning model, with the lowest performance, highlights the importance of choosing the right architecture for specific tasks. Its lower precision, recall, and F1-score suggest that it may not be as effective for gender classification based on facial recognition as the other models. The confusion matrices of the four models are shown in Fig. 5.

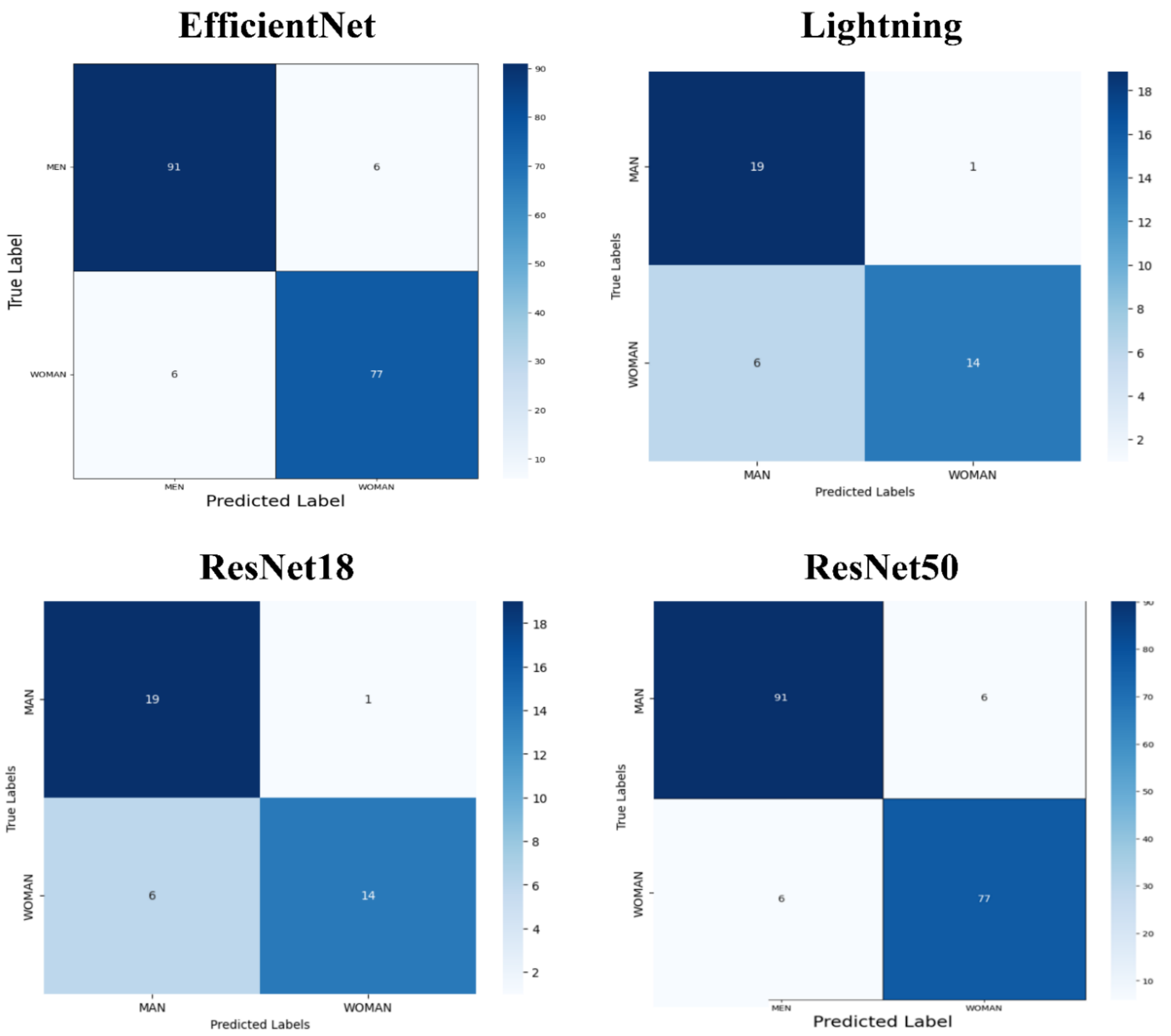

Fig. 5. The confusion matrices of the DL models

The superior performance of ResNet18 and ResNet50 models suggests that deeper architectures with more complex feature extraction capabilities are better suited for this type of classification task.

Now, we need to examine the fitting of the two highest-performing models, ResNet18 and ResNet50 as that presented in Fig. 6, to determine which one is the best for gender classification based on facial recognition. This process involves analyzing the learning curves of both models, which plot their performance on the training and validation sets over successive epochs.

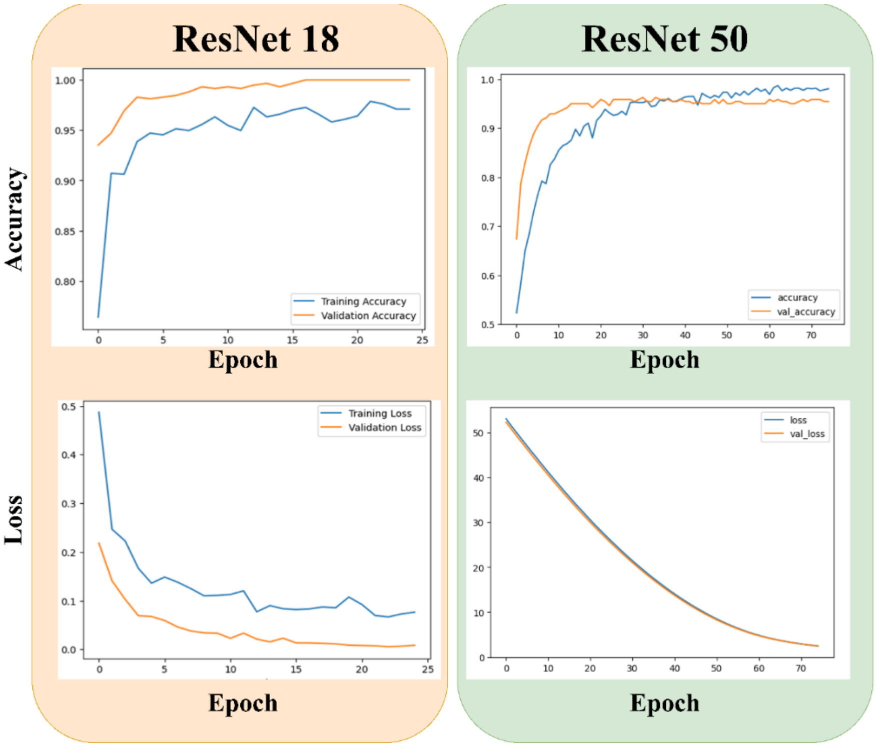

Fig. 6. The fitting of ResNet18 and ResNet50 models

On the left side, for ResNet 18, the top plot shows accuracy over 25 epochs. Both training accuracy and validation accuracy rapidly improve, reaching above 95% within the first few epochs. Validation accuracy slightly surpasses training accuracy, stabilizing around 99%. The bottom plot illustrates the corresponding loss values. Both training loss and validation loss decrease swiftly, indicating effective learning, with validation loss generally lower than training loss, which stabilizes after 10 epochs.

On the right side, for ResNet 50, the top plot depicts accuracy over 75 epochs. Training accuracy starts at about 55%, rapidly increases, and surpasses 95% around epoch 30. Validation accuracy follows a similar trend, stabilizing above 95%. The bottom plot shows the loss values for ResNet 50, with both training and validation loss starting high and gradually decreasing, showcasing consistent learning throughout the epochs. Both losses show a similar pattern, reflecting effective training without overfitting. Both architectures demonstrate high accuracy and effective learning, with ResNet 50 requiring more epochs to stabilize compared to ResNet 18.

The performance of the gender classification model (ResNet18) was evaluated using a dataset containing images labeled as "WOMAN" and "MEN," encoded as 1 and 0, respectively. The model's predictions were compared to the actual labels to assess accuracy. Table 5 below summarizes a sample of the dataset, showcasing the image paths, gender labels, encoded labels, and model predictions:

TABLE 5. SUMMARY OF DATASET SAMPLE WITH IMAGE PATHS, GENDER LABELS, ENCODED LABELS, AND MODEL PREDICTIONS.

	image_path	*label*	*label_encoded*	*model_prediction*
0	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN	1	1
1	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN	1	1
2	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	MEN	0	0
...	...	...	...	...
177	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN	1	0
178	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	MEN	1	1
179	/content/drive/MyDrive/Colab Notebooks/Gender-classification-dataset...	WOMAN	1	1

Out of the 180 samples, most entries demonstrate correct predictions where the label and model prediction match. For instance, in rows 0-4, 175-177 and 179, the model correctly identified the gender. However, some instances of misclassification were observed. For example, in row 177, an image labeled as "WOMAN" (encoded as 1) was incorrectly predicted as "MEN" (0). Overall, the results highlight the model's high accuracy in gender classification, with a one case of misclassification, which will be further analyzed for improvement in future work.

5. CONCLUSION AND FUTURE WORKS

In this paper, we have conducted a comprehensive review and analysis of contemporary DL approaches in face-based gender classification. Our study highlights the significant advancements made in this field, primarily driven by the capabilities of DL architectures such as ResNet18 and ResNet50. These models have demonstrated remarkable accuracy and robustness in gender classification tasks, with ResNet18 achieving the highest performance across various evaluation metrics. The integration of these technologies into real-world applications holds great promise for enhancing security systems, marketing analytics, and healthcare diagnostics. However, it is imperative to address ethical considerations, particularly concerning biases and privacy, to ensure the responsible deployment of these systems. Our findings provide valuable insights for stakeholders aiming to develop and implement effective gender classification technologies.

Despite the promising results, our study acknowledges several limitations. One major challenge is the variability in lighting conditions, facial expressions, and occlusions, which can significantly impact the accuracy of gender classification models. Additionally, biases in training data, particularly concerning ethnicity and age, can lead to skewed results and reduced generalizability across diverse populations. Another limitation is the computational complexity and resource requirements of DL models like ResNet50, which may not be feasible for deployment in resource-constrained environments. Furthermore, our evaluation focused primarily on accuracy and related metrics, potentially overlooking other critical aspects such as interpretability and real-time performance.

However, this study has been able to show that deep learning models including ResNet18 and ResNet50 enhance the gender identification from facial images and the following fields for future research are suggested to enhance the mode robustness, general and usefulness:

- **Addressing Bias in Training Data:** the problem of defining the training sample, which can be racially, age, or anything else, biased. Subsequent research has to be focused on the approaches that reduce these biases or directions that search for the methods on how to reduce or eliminate the classification gender biases or how to improve the models' performances in the other demographical buckets. This could involve selection of datasets such as semi balanced datasets or for instances ushing of bias correction measures at times of training.
- **Improving Robustness to Variability:** Others that could be affecting the model even more, might be illumination change, facial movements and also occlusions. In the future work, other methods of data augmentation could be considered or models which would not be influenced by such changes in the image could be developed. Also, if it is necessary to increase the efficiency of the algorithm at certain conditions, the application of several sources of information, for example facial images and information about the context, lighting, etc.
- **Optimizing Computational Efficiency:** The applicability of other architectures like ResNet50 or DenseNet121 to infer demands in terms of resources could as well be a hurdle for the network capability on a limited computational space. More research should be done in this direction; for instance, lightweight architectures; or the model compression that aims at achieving a higher classification rate, with far fewer computations. This would empower gender classification systems to be moved to the edge devices or real-time utilizations.
- **Enhancing Interpretability and Explainability:** This is more relevant to gender classification models, since these are gradually incorporated in more important and delicate applications including security and health. Regarding further work, the efforts should be made in the direction of enhancing interpretability of such models in order to explain to the concerned users and the other stakeholders as to how the decision for gender prediction has been arrived at. This may in turn require the creation of other geometries or other kinds of diagram to explain to the ordinary layman how the model works.
- **Investigating Ethical Implications:** Because such matters are sensitive and there might be some ethical issues which can be associated with gender classification systems, further research should be done based on the ethical aspect of the system. The question of optimal use of gender classification systems is not settled, and when gender classification systems are widely used, it is crucial to recognize the increase in the amount of societal-level harm that results.

In these areas, future work will allow for the development of better, fairer and more international standards for gender identification technologies and enhance the relevant field and its applications.

Conflicts Of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Acknowledgment

The author acknowledges the support and resources provided by the institution in facilitating the execution of this study.

References

[1] O. Agbo-Ajala and S. Viriri, "Deep learning approach for facial age classification: a survey of the state-of-the-art," *Artificial Intelligence Review,* vol. 54, pp. 179-213, 2021.

[2] M. Hadid, Q. M. Hussein, Z. Al-Qaysi, M. Ahmed, and M. M. Salih, "An overview of content-based image retrieval methods and techniques," *Iraqi Journal For Computer Science and Mathematics,* vol. 4, pp. 66-78, 2023.

[3] J. Gan, L. Xiang, Y. Zhai, C. Mai, G. He, J. Zeng*, et al.*, "2M BeautyNet: Facial beauty prediction based on multi-task transfer learning," *IEEE Access,* vol. 8, pp. 20245-20256, 2020.

[4] A. S. Albahri, R. A. Hamid, J. K. Alwan, Z. Al-Qays, A. Zaidan, B. Zaidan*, et al.*, "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review," *Journal of medical systems,* vol. 44, pp. 1-11, 2020.

[5] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker, "How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis," *Proceedings of the ACM on Human-computer Interaction,* vol. 4, pp. 1-35, 2020.

[6] S. Kumar, S. Rani, A. Jain, C. Verma, M. S. Raboaca, Z. Illés*, et al.*, "Face spoofing, age, gender and facial expression recognition using advance neural network architecture-based biometric system," *Sensors,* vol. 22, p. 5160, 2022.

[7] O. Albahri, A. Zaidan, A. Albahri, B. Zaidan, K. H. Abdulkareem, Z. Al-Qaysi*, et al.*, "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *Journal of infection and public health,* vol. 13, pp. 1381-1396, 2020.

[8] S. Dargan, M. Kumar, and S. Tuteja, "PCA-based gender classification system using hybridization of features and classification techniques," *Soft Computing,* vol. 25, pp. 15281-15295, 2021.

[9] M. B. Omar, M. A. Ahmed, H. A. Hussein, Z. Al-Qaysi, M. M. Salih, S. Hamddi*, et al.*, "Taxonomy, Open Challenges, Motivations, and Recommendations in Driver Behavior Recognition: A Systematic Review."

[10] Z. Al-Qaysi, M. Suzani, N. bin Abdul Rashid, R. A. Aljanabi, R. D. Ismail, M. Ahmed*, et al.*, "Optimal Time Window Selection in the Wavelet Signal Domain for Brain–Computer Interfaces in Wheelchair Steering Control," *Applied Data Science and Analysis,* vol. 2024, pp. 69-81, 2024.

[11] R. Abdul Ameer, M. Ahmed, Z. Al-Qaysi, M. Salih, and M. L. Shuwandy, "Empowering Communication: A Deep Learning Framework for Arabic Sign Language Recognition with an Attention Mechanism," *Computers,* vol. 13, p. 153, 2024.

[12] M. Ahmed, M. D. Salman, R. Adel, Z. Alsharida, and M. Hammood, "An intelligent attendance system based on convolutional neural networks for real-time student face identifications," *Journal of Engineering Science and Technology,* vol. 17, pp. 3326-3341, 2022.

[13] J. Naskath, G. Sivakamasundari, and A. A. S. Begum, "A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN," *Wireless personal communications,* vol. 128, pp. 2913-2936, 2023.

[14] T. Xayasouk, H. Lee, and G. Lee, "Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models," *Sustainability,* vol. 12, p. 2570, 2020.

[15] A. Derry, M. Krzywinski, and N. Altman, "Convolutional neural networks," *Nature Methods,* vol. 20, pp. 1269-1270, 2023.

[16] O. N. Haggab and Z. Al-Qaysi, "Detecting Defect in Central Pivot Irrigation System using YOLOv7 Algorithms," *Al-Salam Journal for Engineering and Technology,* vol. 3, pp. 38-49, 2024.

[17] J. Mellor, J. Turner, A. Storkey, and E. J. Crowley, "Neural architecture search without training," in *International conference on machine learning*, 2021, pp. 7588-7598.

[18] A. H. Fath, F. Madanifar, and M. Abbasi, "Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems," *Petroleum,* vol. 6, pp. 80-91, 2020.

[19] A. You, J. K. Kim, I. H. Ryu, and T. K. Yoo, "Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey," *Eye and Vision,* vol. 9, p. 6, 2022.

[20] A. S. Alphonse, K. Shankar, M. Jeyasheela Rakkini, S. Ananthakrishnan, S. Athisayamani, A. Robert Singh*, et al.*, "A multi-scale and rotation-invariant phase pattern (MRIPP) and a stack of restricted Boltzmann machine (RBM) with preprocessing for facial expression classification," *Journal of Ambient Intelligence and Humanized Computing,* vol. 12, pp. 3447-3463, 2021.

[21] S. Haseena, S. Saroja, R. Madavan, A. Karthick, B. Pant, and M. Kifetew, "Prediction of the age and gender based on human face images based on deep learning algorithm," *Computational and Mathematical Methods in Medicine,* vol. 2022, p. 1413597, 2022.

[22] M. Hadid, Z. Al-Qaysi, Q. M. Hussein, R. A. Aljanabi, I. R. Abdulqader, M. Suzani*, et al.*, "Semantic Image Retrieval Analysis Based on Deep Learning and Singular Value Decomposition," *Applied Data Science and Analysis,* vol. 2024, pp. 17-31, 2024.

[23] R. A. Aljanabi, Z. Al-Qaysi, and M. Suzani, "Deep Transfer Learning Model for EEG Biometric Decoding," *Applied Data Science and Analysis,* vol. 2024, pp. 4-16, 2024.

[24] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, pp. 4396-4415, 2022.

[25] M. A. Abed, Z. Al-Qaysi, and M. Suzani, "Improving Lumbar Disc Bulging Detection in MRI Spinal Imaging: A Deep Learning Approach," *Al-Salam Journal for Engineering and Technology,* vol. 4, pp. 1-19, 2025.

[26] A. De Keyser, Y. Bart, X. Gu, S. Q. Liu, S. G. Robinson, and P. Kannan, "Opportunities and challenges of using biometrics for business: Developing a research agenda," *Journal of Business Research,* vol. 136, pp. 52-62, 2021.

[27] A. Haleem, M. Javaid, M. A. Qadri, R. P. Singh, and R. Suman, "Artificial intelligence (AI) applications for marketing: A literature-based study," *International Journal of Intelligent Networks,* vol. 3, pp. 119-132, 2022.

[28] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society,* vol. 1, pp. 89-103, 2020.

[29] S. M. Samuri, T. V. Nova, B. Rahmatullah, S. L. Wang, and Z. T. Al-Qaysi, "Classification model for breast cancer mammograms," *IIUM Engineering Journal,* vol. 23, pp. 187-199, 2022.

[30] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics,* vol. 9, p. 1188, 2020.

[31] A. Krishnan and A. Rattani, "A novel approach for bias mitigation of gender classification algorithms using consistency regularization," *Image and Vision Computing,* vol. 137, p. 104793, 2023.

[32] L. L. Dhirani, N. Mukhtiar, B. S. Chowdhry, and T. Newe, "Ethical dilemmas and privacy issues in emerging technologies: A review," *Sensors,* vol. 23, p. 1151, 2023.

[33] A. Albahri, Y. L. Khaleel, and M. A. Habeeb, "The Considerations of Trustworthy AI Components in Generative AI; A Letter to Editor," *Applied Data Science and Analysis,* vol. 2023, pp. 108-109, 2023.

[34] Q. A. Hameed, H. A. Hussein, M. A. Ahmed, M. M. Salih, R. D. Ismael, and M. B. Omar, "UXO-AID: A new UXO classification application based on augmented reality to assist deminers," *Computers,* vol. 11, p. 124, 2022.

[35] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey toward private and secure applications," *ACM Computing Surveys (CSUR),* vol. 54, pp. 1-38, 2021.

[36] Z. Al-Qaysi, A. Al-Saegh, A. F. Hussein, and M. Ahmed, "Wavelet-based Hybrid learning framework for motor imagery classification," *Iraqi J Electr Electron Eng,* 2022.

[37] A. Hussein, M. E. Sallam, and M. Y. A. Abdalla , Trans., “Exploring New Horizons: Surgical Robots Supported by Artificial Intelligence ”, MJAIH , vol. 2023, pp. 40–44, Aug. 2023, doi: 10.58496/MJAIH/2023/008.

[38] Z. Al-Qaysi, M. A. Ahmed, N. M. Hammash, A. F. Hussein, A. S. Albahri, M. Suzani*, et al.*, "A systematic rank of smart training environment applications with motor imagery brain-computer interface," *Multimedia Tools and Applications,* vol. 82, pp. 17905-17927, 2023.

[39] Y. Deldjoo, T. D. Noia, and F. A. Merra, "A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks," *ACM Computing Surveys (CSUR),* vol. 54, pp. 1-38, 2021.

[40] Y. L. Khaleel, M. A. Habeeb, A. Albahri, T. Al-Quraishi, O. Albahri, and A. Alamoodi, "Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods," *Journal of Intelligent Systems,* vol. 33, p. 20240153, 2024.

[41] R. D. Ismail, Q. A. Hameed, and M. B. Omar, "An EEG based Physiological Signal for Driver Behavior Monitoring Systems: A Review," *Tikrit Journal for Computer Science and Mathematics*, vol. 1, pp. 38-54, 2023.

[42] F. K. H. Mihna, M. A. Habeeb, Y. L. Khaleel, Y. H. Ali, and L. A. E. Al-saeedi, "Using information technology for comprehensive analysis and prediction in forensic evidence," *Mesopotamian Journal of CyberSecurity*, vol. 4, pp. 4-16, 2024.

[43] Z. Al-Qaysi, A. Albahri, M. Ahmed, R. A. Hamid, M. Alsalem, O. Albahri*, et al.*, "A comprehensive review of deep learning power in steady-state visual evoked potentials," *Neural Computing and Applications*, pp. 1-24, 2024.

[44] A. Albahri, M. M. Jassim, L. Alzubaidi, R. A. Hamid, M. Ahmed, Z. Al-Qaysi*, et al.*, "A trustworthy and explainable framework for benchmarking hybrid deep learning models based on chest X-ray analysis in CAD systems," *International Journal of Information Technology and Decision Making*, 2024.

[45] A. Albahri, Y. L. Khaleel, M. A. Habeeb, R. D. Ismael, Q. A. Hameed, M. Deveci*, et al.*, "A systematic review of trustworthy artificial intelligence applications in natural disasters," *Computers and Electrical Engineering*, vol. 118, p. 109409, 2024.

[46] Z. Al-Qaysi, M. Suzani, N. bin Abdul Rashid, R. D. Ismail, M. Ahmed, R. A. Aljanabi*, et al.*, "Generalized Time Domain Prediction Model for Motor Imagery-based Wheelchair Movement Control," *Mesopotamian Journal of Big Data*, vol. 2024, pp. 68-81, 2024.

[47] Z. Al-Qaysi, M. Suzani, N. bin Abdul Rashid, R. D. Ismail, M. Ahmed, W. A. W. Sulaiman*, et al.*, "A Frequency-Domain Pattern Recognition Model for Motor Imagery-Based Brain-Computer Interface," *Applied Data Science and Analysis*, vol. 2024, pp. 82-100, 2024.

[48] W. G. Yass and M. Faris , Trans., "A Comprehensive Review of Deep Learning and Machine Learning Techniques for Real-Time Car Detection and Wrong-Way Vehicle Tracking", Babylonian Journal of Machine Learning, vol. 2023, pp. 78–90, Nov. 2023, doi: 10.58496/BJML/2023/013.

[49] R. Kumar, K. Singh, D. P. Mahato, and U. Gupta, "Face-based age and gender classification using deep learning model," *Procedia Computer Science*, vol. 235, pp. 2985-2995, 2024.

[50] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, vol. 275, pp. 448-461, 2018.

[51] M. Oulad-Kaddour, H. Haddadou, C. C. Vilda, D. Palacios-Alonso, K. Benatchba, and E. Cabello, "Deep learning-based gender classification by training with fake data," *IEEE Access*, vol. 11, pp. 120766-120779, 2023.

[52] A. Lahariya, V. Singh, and U. S. Tiwary, "Real-time Emotion and Gender Classification using Ensemble CNN," *arXiv preprint arXiv:2111.07746*, 2021.

[53] K. T. Ahmed, H. Afzal, M. R. Mufti, A. Mehmood, and G. S. Choi, "Deep image sensing and retrieval using suppression, scale spacing and division, interpolation and spatial color coordinates with bag of words for large and complex datasets," *IEEE Access*, vol. 8, pp. 90351-90379, 2020.

[54] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, pp. 545-563, 2021.

[55] Ü. Atila, M. Uçar, K. Akyol, and E. Uçar, "Plant leaf disease classification using EfficientNet deep learning model," *Ecological Informatics*, vol. 61, p. 101182, 2021.

[56] J. Y. I. Alzamily, S. B. Ariffin, and S. S. Abu-Naser, "Classification of Encrypted Images Using Deep Learning–Resnet50," *Journal of Theoretical and Applied Information Technology*, vol. 100, pp. 6610-6620, 2022.

[57] Y. Zhao, X. Zhang, W. Feng, and J. Xu, "Deep learning classification by ResNet-18 based on the real spectral dataset from multispectral remote sensing images," *Remote Sensing*, vol. 14, p. 4883, 2022.

[58] H. M. S. SALEEH, H. Marouane, and A. Fakhfakh , Trans., "A Novel Deep Learning Approach for Detecting Types of Attacks in the NSL-KDD Dataset", BJN, vol. 2024, pp. 171–181, Sep. 2024, doi: 10.58496/BJN/2024/017.

[59] M. A. Habeeb, Y. L. Khaleel, and A. Albahri, "Toward Smart Bicycle Safety: Leveraging Machine Learning Models and Optimal Lighting Solutions," in *The International Conference on Innovations in Computing Research*, 2024, pp. 120-131.

[60] M. Ahmed, Z. Al-Qaysi, A. Albahri, M. Alqaysi, G. Kou, O. Albahri*, et al.*, "Intelligent decision-making framework for evaluating and benchmarking hybridized multi-deep transfer learning models: managing COVID-19 and beyond," *International Journal of Information Technology & Decision Making*, p. 2350046, 2023.

[61] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, vol. 54, pp. 6391-6438, 2021.