Sustainable Artificial Intelligence: Assessing Performance in Detecting Fake Images

Othman A. Alrusaini
Department of Engineering and Applied Sciences-Applied College, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Detecting fake images is crucial because they may confuse and influence people into making bad judgments or adopting incorrect stances that might have disastrous consequences. In this study, we investigate not only the effectiveness of artificial intelligence, specifically deep learning and deep neural networks, for fake image detection but also the sustainability of these methods. The primary objective of this investigation was to determine the efficacy and sustainable application of deep learning algorithms in detecting fake images. We measured the amplitude of observable phenomena using effect sizes and random effects. Our meta-analysis of 32 relevant studies revealed a compelling effect size of 1.7337, indicating that the model's performance is robust. Despite this, some moderate heterogeneity was observed (Q-value = 65.5867; I^2 = 52.7344%). While deep learning solutions such as CNNs and GANs emerged as leaders in detecting fake images, their efficacy and sustainability were contingent on the nature of the training images and the resources consumed during training and operation. The study highlighted adversarial confrontations, the need for perpetual model revisions due to the ever-changing nature of image manipulations, and data scarcity as technical obstacles. Additionally, the sustainable deployment of these AI technologies in diverse environments was considered crucial.**

Keywords*—Artificial intelligence; image validation; deep learning; deep neural networks; fake images; image forgery; image manipulations

I. Introduction

Technological advancements in graphics design continue to receive unprecedented improvements with each new software release. Such advancements are beneficial and detrimental, as they can positively and negatively impact people. On the positive side, repaying old images and restoring them to their former state is now possible. Using software such as Photoshop, a designer can clone sections of an image using the clone stamp tool and replicate the pattern in a different area to make it appear real and authentic [1, 2]. Users can also add interesting features to their photographs to add aesthetics that were previously not present in their photographs. This aesthetic appeal from edited images can improve an image's appeal and introduce an element of fantasy into the work.

Nevertheless, this technology has seen its application stretch beyond serving people's genuine needs to improve photographic appeal. Many, if not most, of its application has been doctoring images to trick people into believing falsehoods [3, 4]. One of the fields that have suffered immensely is academics. People can now engage in document forgery to create certificates that look exactly like an institution's legally issued credentials [5]. Most recent versions of the graphics editing software use artificial intelligence (AI) to edit images, making the finishing even more illustrious [6]. This fact makes it quite difficult to detect fake images from the real ones using the naked eye, thereby creating an extra layer of complexity to the process.

The primary objective of this research is to undertake a meta-analysis study of deep learning tools and technologies used to detect fake images, with a focus on both their effectiveness and sustainability. The research questions guiding this analysis are as follows:

1) How effective are deep learning algorithms in detecting fake images, and how do their effectiveness and sustainability correlate?

2) What are the most reliable evaluation metrics for evaluating the performance and sustainability of deep learning algorithms in detecting fake images?

3) What are the technical challenges in the sustainable detection of fake images using deep learning techniques?

While several studies have engaged in the primary research of creating and evaluating deep learning models to detect fake images, few have done it in a meta-analytical way that also considers the sustainability of these technologies. This approach effectively synthesizes the field's gains in developing the algorithms. It also exposes the weaknesses, gaps, and sustainability concerns that need filling to improve algorithmic formidability. It is expected that this research and its analysis will add to the existing fake images detection with AI literature by providing a better understanding of the different models and various datasets.

The paper contains six sections: introduction, background, methods, results, discussion and conclusion. The introduction in Section I sets the stage for what the paper will deliberate on and establishes the researcher's rationale. The background in Section II, the paper delves deep into the current solutions and technological overview to give the reader a good vantage point from which to appreciate the gains and weaknesses in the field of fake image detection using deep learning technologies. The methods in Section III takes the reader through a setup of the meta-analytical approach, including search strategies, data sources, inclusion and exclusion criteria, and data analysis approaches. The results in Section IV comprehensively analyses the findings made in the analysis. Finally, the paper discusses these findings to synthesize the results and also summarises along with suggested areas for further investigation in Section V and Section VI respectively.

II. BACKGROUND

A. Types of Fake Images

The diversity of fake images has made their detection all the more challenging because of the intricacy that comes with each type. Image splicing is one of the many types contributing to the fake image ecosystem [7, 8]. It refers to the result of combining parts of different images to create a new but deceptive version. It is almost similar to morphing and blending two or more images. In both cases, tampering with the final image is difficult to identify with the naked eye. Sometimes, creating fake images may involve hiding some aspects within the image to make it look different [9]. One way to do so is by deleting the unwanted element, which counts as the removal technique. This technique is paired with 'insertion,' introducing a new element into the hitherto non-existent image. The second method to hide elements within an image is steganography, a more technical form of hiding the undesired elements [10].

Some techniques neither remove nor add elements to the original image. Instead, they work on the image's appearance to change certain aspects, such as lighting, contrast, color, and texture. The most popular methods are bundled under the group' filter-based manipulations' techniques [11, 12]. It is worth noting that designers often use filter-based manipulation techniques with others to create a new 'cooler' image. 3D rendering is another technique that changes the image from its 2D orientation to feature the third dimension of depth. While it does not introduce new elements, some shadows are likely to appear to give the illusion of a 3D image [13]. Regardless of the type of change and the intention behind such changes, analysts must be able to detect the changes programmatically for a more reliable consumption of these digital products. Deep learning is heavily equipped to check for even the mildest inconsistencies within the image structure and report them instantly [14].

B. Traditional Solutions to Detecting Fake Images

The challenge of detecting fake images predates the deep fake technologies, as there has been image doctoring beforehand. One solution that most analyses preferred using was engaging in image forensics. It is a collection of techniques involving statistical and pixel-level image analysis to identify inconsistencies [15, 16]. Some of the aspects sought after are differing noise patterns and lighting anomalies. This method is advantageous because it is not computationally expensive but falters in detecting high-end forgeries and is not easily scalable. Watermarking has also been a key technique in ensuring viewers can tell fake images from the original. While it is effective in copyright protection, it is less applicable to images whose originals do not have watermarks [17, 18]. Another traditional method is meta-analysis, which involves inspecting the images' metadata to detect fakes. While it provides contextual data, the metadata can be easily altered with the right technologies [15].

C. AI-Based Solutions to Detecting Fake Images

With the advent of artificial intelligence, so much technological progress has come about and has permeated the field of image analytics. One such technology is convoluted neural networks (CNNs), a specific network architecture for deep learning algorithms [15, 19]. Its highly-rated image analytical capabilities can automatically and adaptively learn spatial hierarchies of features in an image. It is highly accurate and is mostly applicable when there are complex patterns. However, their large dataset requirements imply they are computationally expensive [14]. Generative adversarial networks (GANs) are another technology consisting of a generator and discriminator working against each other, widely used for deepfake detection. Like CNNs, they thrive in complex patterns [8, 20]. Nevertheless, their nature as unsupervised learning models makes them susceptible to being unstable and generating false positives. Recurrent neural networks are similar to other technologies but are mostly applicable in video forensics because of their strength in analyzing sequential data [21, 22].

Ensemble methods, transfer learning, and zero-shot learning represent advanced AI approaches that address different aspects of fake image detection. Ensemble methods involve the combination of multiple AI models to improve predictive accuracy and robustness, although they come at the cost of computational expense and increased model complexity [23]. On the other hand, transfer learning provides an efficient approach by applying pre-trained models to new but similar tasks, effectively saving time and computational resources; however, its applicability is constrained to tasks that closely resemble the original training data [24]. Lastly, zero-shot learning presents a frontier in AI-based fake image detection, offering the ability to recognize types of fake imagery that the model has not been specifically trained on [25]. While this method is versatile and adaptable, it is still an area under active research, and thus its reliability is not fully established. Each method has advantages and disadvantages, emphasizing the need for ongoing research to refine these techniques and possibly integrate them for more effective and efficient fake image detection.

D. Fake Image Detection Process using AI

This section describes the methodological steps involved in detecting fake images through the use of AI to collect and analyze data. The flowchart for these steps is shown in Fig. 1.

Fig. 1. AI Fake image detection process flowchart.

1) Data collection: Data collection is fundamental in building supervised machine learning models. It is the first stage in fake image detection using AI. In this phase, high-dimensional data, often in image matrices or tensors, is gathered [24]. This dataset, comprising RGB values or even grayscale pixel intensities, is crucial for subsequent feature engineering. The data must also be annotated through manual labeling or semi-supervised methods to create ground truth labels distinguishing genuine images from artificially manipulated or deepfake counterparts.

2) Feature extraction: Feature extraction involves transforming raw image data into a reduced dimensionality

format using algorithms that capture essential patterns. Techniques such as convolutional neural networks (CNNs) are often employed here, leveraging filter banks to highlight vital image attributes such as edges, textures, and regions of interest [26]. When convolved with input data, these filters output feature maps highlighting image characteristics.

3) Feature selection: Not all features extracted hold discriminative power for the classification task at hand. Feature selection focuses on refining the feature set, eliminating redundant or irrelevant features to mitigate the curse of dimensionality, and optimizing model performance [27]. Algorithms such as Recursive Feature Elimination (RFE) or techniques leveraging mutual information can be utilized to determine the most salient features.

4) Classification: With a refined feature set, the classification phase employs algorithms to map these features to their respective labels. Deep learning architectures, like CNNs or more complex models like Residual Networks (ResNets), are trained using backpropagation [28]. The objective is to adjust weights and biases to minimize the loss function, typically cross-entropy loss for classification tasks.

5) Evaluation: Post-training, the model's robustness, and generalizability are gauged using accuracy, precision, recall, and the F1 score on a holdout validation or test dataset. Techniques like k-fold cross-validation can ensure the evaluation is comprehensive, and if underfitting or overfitting is detected, hyperparameter tuning, regularization methods, or architecture adjustments might be necessitated [20, 29].

III. Methods

A. Literature Search

In our investigation into artificial intelligence and image validation, we selected a set of keywords to guide our data extraction process. Central to our inquiry was "Deep learning," which is intrinsically tied to "Deep neural networks." To delve into the specific area of counterfeit or fake imagery, we utilized terms such as "Fake images," "Image forgery," "Image tampering," and "Image authenticity." Recognizing the significance of assessing the capability of algorithms, we incorporated "Effectiveness" and "Performance" into our search parameters. The term "Image detection" was chosen to understand the broader mechanisms behind image recognition and validation. Additionally, the "F1 score" was included as a metric of interest to gain insights into evaluation methods. Its inclusion is because of its widespread application in balancing precision and recall in binary classification problems. Through these keywords, we aimed to ensure a comprehensive exploration of the current state of deep learning techniques in detecting fake images.

B. PRISMA Flow Chart

The research was undertaken following the guidance of the PRISMA flowchart. It is a flow diagram reporting the stages articles go through to determine whether they are fit for inclusion in a meta-analysis [30]. The PRISMA flowchart in Fig. 2 delineates the sequence of a meta-analysis process. Initiating with the identification phase, a search was conducted across 10 databases, unearthing a total of 317 studies. From this collection, preliminary screening reduced the number to 226 records. The reasons for this reduction were multi-fold: 54 records were identified as duplicates, 23 were found ineligible based on certain criteria, and 14 were removed due to other specified reasons. The subsequent phase saw 187 of these 226 screened records being selected for detailed report retrieval. Of these, 44 reports could not be retrieved, which left a pool of 143 reports. These reports were then subjected to a comprehensive eligibility assessment. In the final count, several reports were excluded from the 143 due to a range of reasons: a lack of full-text availability in 20 reports, 45 not matching the necessary keywords, 27 not meeting the quality appraisal standards, and 19 being deemed irrelevant to the study's focus. After these exclusions, the evaluation was refined to a set of 32 studies that were considered relevant and included in the meta-analysis.

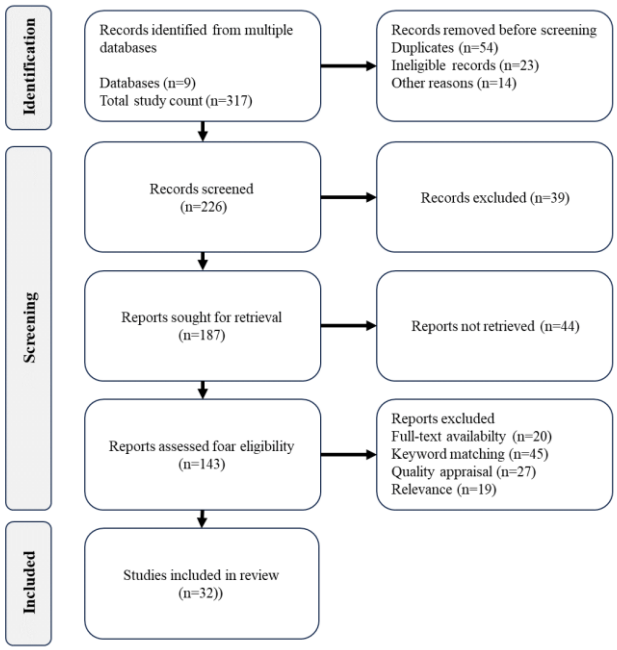

Fig. 2. PRISMA flowchart.

C. Distribution of Journals

Regarding the journals from which these final papers were derived, 'IEEE Access' from IEEE had the most articles, with eight total. MDPI's 'Applied Sciences' received five submissions, as illustrated in Table I. MDPI's 'Journal of Imaging' and 'Sensors' received three each. The paper source 'ACM Conference papers' appeared twice.

Other journals on the list had a source each, as observed in Table I. These numerous sources demonstrate how many fields are interested in applying artificial intelligence to detect fake images.

D. Computing Effect Sizes

Effect sizes are indispensable in meta-analysis, serving as a standardized metric to gauge the magnitude of observed phenomena and facilitating comparisons across diverse studies. This study focused on the F1 score, which measures a model's accuracy. The researcher also focused on the sample

size, denoted by the number of images used in individual studies. The choice was made to employ the random-effects model, considering the potential variability in study designs and regions covered. The determination of effect sizes hinged on a transformation suitable for F1 scores, as opposed to the conventional Fisher's Z transformation tailored for correlation coefficients [31]. Both Q-statistics and I^2-values were deployed to comprehend the variability or heterogeneity of the results among the different studies [32]. In meta-analysis, heterogeneity indicates the differences in outcomes across the incorporated studies. If heterogeneity is high, it implies that the studies' results vary considerably [31]. Such computations empower this study, offering a rigorous quantitative consolidation of literature on using artificial intelligence to detect fake images.

TABLE I. DISTRIBUTION OF STUDIES BY JOURNAL

Journal	Publication	#
IEEE Access	IEEE	8
Applied Sciences	MDPI	5
Journal of Imaging	MDPI	3
Sensors	MDPI	3
ACM Conference papers	ACM	2
International Journal of Advanced Computer Science and Applications	The Science and Information Organization	1
Journal of Visual Communication and Image Representation	Elsevier	1
International Journal of Scientific Research in Computer Science Engineering and Information Technology	IJSRCSEIT	1
The Visual Computer	Springer	1
Journal of Cybersecurity and Privacy	MDPI	1
Entropy	MDPI	1
PeerJ Computer Science	PeerJ	1
Neural Computing and Applications	Springer	1
IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS)	IRACST	1
International Journal of Information Technology	Springer	1
Electronics	MDPI	1

1) Fisher's Z transformation: Fisher's Z is a statistical method that transforms Pearson correlation coefficients into a normally distributed variable [31]. The transformation is used because Pearson's r does not have a normal distribution, which makes its confidence intervals asymmetric. Fisher's Z transformation aids in stabilizing this variance. A higher absolute value of Fisher's Z indicates a stronger relationship between variables.

$$Fisher's\ Z\ =\ 0.5\ \times ln\left(\frac{1+f}{1-f}\right) \quad (1)$$

where, f = F1-score

2) The Weight value for each study (w_i): The w_i value represents the inverse of the variance of the effect size for study i [31]. It gives more weight to studies with more precise estimates (i.e., smaller variance), allowing them to influence the combined effect size more than those with less precise estimates [2].

$$w_i = \left(\frac{z_i}{Var(z)}\right) \quad (2)$$

Where, z_i is a study's effect size measured by Fisher's Z index.

3) The Overall effect size (z_+): The z_+ formula calculates the combined effect size in meta-analyses using weighted individual study effect sizes [32]. This Equation allows the aggregation of individual study results to derive an overall effect, giving more weight to studies with larger sample sizes or more precise measurements. A greater z_+ value suggests a larger overall effect size across the analyzed studies.

$$z_+ = \frac{\Sigma_{i=1}^{N}(w_i\ \times\ z_i)}{\Sigma_{i=1}^{N}\ w_i} \quad (3)$$

where, N = number of studies

4) The Q statistic: The Q statistic measures the heterogeneity or variability of effect sizes across studies in a meta-analysis [31]. Determining heterogeneity is essential to deciding on the type of meta-analytic model (fixed-effects vs. random-effects) to use. A significant Q value indicates substantial heterogeneity among the included studies, suggesting that differences in effect sizes aren't solely due to sampling error.

$$Q\ =\ \sum_{i=1}^{N} w_i\ \times (z_i - z_+)^2 \quad (4)$$

5) The I^2 metric: I^2 is a metric that quantifies the percentage of total variability in study estimates attributable to heterogeneity rather than chance [32]. It provides insights into the consistency of findings across studies, independent of the number of studies included. An I^2 value close to 100% indicates high heterogeneity, suggesting that the results of the studies are diverse.

$$I^2 = \left(\frac{Q - N + 1}{Q}\right) \times\ 100 \quad (5)$$

6) Bounds: This bounds equation calculates the confidence interval around the mean effect size based on the standard deviation of study weights [32]. The bounds provide a range within which the true effect size is expected to fall, offering a measure of the precision of the effect size estimate. Narrower intervals denote more precise estimates, while wider intervals indicate more uncertainty around the effect size.

$$bounds\ =\ f_z\ \pm\ \alpha \times \sigma \quad (6)$$

where, f_z is the mean F1 score given all the studies included in the meta-analysis, and α is the level of

significance, while σ refers to the standard deviation of the study weights.

7) Fail-Safe N: The Fail-safe N method estimates the number of unpublished or "missing" studies required to nullify the effect observed in a meta-analysis [31]. It addresses the potential publication bias in meta-analyses. A high Fail-safe N suggests the meta-analysis results are robust against potential publication bias.

$$Fail-safe\ N\left(N_{fs.05}\right) = \frac{[(\sum z)^2\ -(N * \bar{z}^2)]}{\alpha^2} \qquad (7)$$

Where, ∑z is the summation of all Fisher's z indices, and $\bar{z}$ is the mean of all Fisher's z indices.

8) Critical value: The formula gives a threshold number for robustness against publication bias in meta-analyses [31]. Considering the number of studies in the analysis, it assesses the risk of overestimating effects due to publication bias. A higher critical value implies greater robustness of the meta-analysis results against potential biases.

$$Critical\ value\ = 5 \times N\ + 10 \qquad (8)$$

IV. Results

A. Study Statistics

Table II shows the summary statistics for the 32 studies included in the meta-analysis. The statistics include sample size, F1-score, Fisher's z, and their weights.

TABLE II. Individual Study Statistics

	Study	Year	Purpose	Dataset	Sample Size	Model	F1-Score	Fisher's Z	Weight
1	[33]	2022	Image Classification	CIFAR-10	19579	CNN	0.94	1.74	10.5177
2	[1]	2023	Fake Image Detection	CELEBA	7373	RNN	0.89	1.42	8.60468
3	[6]	2022	Fake Image Detection	DeepFake	24060	RNN	0.85	1.26	7.60152
4	[34]	2021	Object Detection	COCO	5361	CNN	0.84	1.22	7.38984
5	[35]	2022	Image and Video Analysis	UCF101	31671	CNN	0.87	1.33	8.06704
6	[27]	2018	Fake Image Detection	FaceForensics++	21953	RNN	0.92	1.59	9.61588
7	[36]	2022	Fake Image Detection	DFDC	7628	LSTM	0.99	2.65	16.016
8	[9]	2018	Image Classification	ImageNet	26622	CNN	0.84	1.22	7.38984
9	[37]	2021	Object Tracking	GOT-10k	11570	CNN	0.95	1.83	11.0849
10	[38]	2021	Fake Image Detection	DeepFake	34626	LSTM	0.91	1.53	9.2437
11	[39]	2022	Fake Image Detection	DFDC	21287	RNN	0.99	2.65	16.016
12	[40]	2020	Fake Image Detection	FaceForensics++	4053	RNN	0.9	1.47	8.90903
13	[41]	2022	Fake Image Detection	DFDC	12914	LSTM	0.97	2.09	12.6614
14	[42]	2022	Fake Image Detection	BEGAN	27843	RNN	0.96	1.95	11.7755
15	[24]	2021	Image Generation	Pascal VOC	27983	GAN	0.95	1.83	11.0849
16	[43]	2023	Object Detection	DeepFake	6216	CNN	0.99	2.65	16.016
17	[44]	2022	Fake Image Detection	Kinetics	6750	LSTM	0.86	1.29	7.82658
18	[45]	2018	Video Classification	ADE20K	3742	CNN	0.86	1.29	7.82658
19	[5]	2020	Image Segmentation	CELEBA	29993	CNN	0.84	1.22	7.38984
20	[46]	2017	Fake Image Detection	DFDC	31835	GAN	0.93	1.66	10.0356
21	[47]	2021	Fake Image Detection	CIFAR-100	20979	RNN	0.91	1.53	9.2437
22	[10]	2019	Image Classification	DeepFake	26169	CNN	0.85	1.26	7.60152
23	[28]	2018	Fake Image Detection	YOLO	8365	LSTM	0.92	1.59	9.61588
24	[29]	2018	Object Detection	PCGAN	11032	CNN	0.85	1.26	7.60152
25	[14]	2023	Image Synthesis	CELEBA	22837	GAN	0.86	1.29	7.82658
26	[48]	2021	Fake Image Detection	FaceForensics++	8702	RNN	0.94	1.74	10.5177
27	[49]	2019	Fake Image Detection	GAN	16576	LSTM	0.89	1.42	8.60468
28	[50]	2019	Image Generation	AVA	13425	GAN	0.95	1.83	11.0849
29	[22]	2021	Image and Video Analysis	DeepFake	23896	CNN	0.87	1.33	8.06704
30	[25]	2018	Fake Image Detection	DFDC	6611	RNN	0.89	1.42	8.60468
31	[51]	2021	Fake Image Detection	CELEBA	30177	LSTM	0.94	1.74	10.5177
32	[26]	2023	Fake Image Detection	CIFAR-10	32538	RNN	0.96	1.95	11.7755

Table II showcases various studies from 2017 to 2023, spanning applications like themes in fake image detection. Commonly used models include CNNs, RNNs, LSTMs, and GANs. Notably, LSTMs achieved top F1 scores in several studies ([35],[37],[41]). The consistency in Fisher's Z values suggests a uniform significance level. The weight, mirroring the sample size, hints at the study's reliability. In essence, the table reflects both progress and challenges in AI research.

The datasets employed in the studies are all related to machine learning and artificial intelligence. CIFAR-10 and CIFAR-100 are widely used benchmarks for image classification, consisting of small images from several categories. ImageNet, a significant player in the image classification field, has been instrumental in driving progress in deep learning. CELEBA focuses on facial attributes and provides a wide range of annotated faces. DeepFake, DFDC, and FaceForensics++ focus on detecting fake images and videos, which are crucial for developing ways to combat misinformation. COCO and Pascal VOC are widely used in object detection, while UCF101 and Kinetics are popular for video classification and analysis. GOT-10k is specifically made for object tracking, while ADE20K is focused on semantic segmentation tasks. Generative models like BEGAN, PCGAN, and GAN datasets play a crucial role in image synthesis and generation.

B. Meta-Analysis Summary Statistics

Table III summarizes the statistics from the 32 studies involved in this meta-analysis. The statistics include z_+, Q, I^2, σ, lower bound, upper bound, critical Nfs, Nfs, CI, and statistical significance.

TABLE III. OVERALL STUDY STATISTICS

Statistic	Value
z_+	1.7337
Q	65.5867
I^2	52.7344
σ	0.0562
Lower Bound	1.6235
Upper Bound	1.8439
Critical Nfs	170
Nfs	2706.9451
CI	[1.6235, 1.8439]
Statistical Significance	$p < 0.001$

1) Overall effect size (z_+: 1.7337): The z_+ value represents the meta-analysis's pooled or combined standardized effect size. A value of 1.7337 indicates a positive and relatively strong overall effect size [18]. These findings suggest that from a general standpoint, the machine learning models employed in the studies performed well in detecting fake images. According to [33, 52], 0.2 is small, 0.5 is medium, and 0.8 or above is considered large. Hence, the z_+ value obtained in this case shows a large effect size.

2) Heterogeneity (Q: 65.5867, I^2: 52.7344%): Both Q and I^2 are measures of heterogeneity among the included studies. The high Q-value suggests significant variability in the effect sizes across studies [51]. Table II illustrates the differences in the F1 scores obtained from running different machine learning models in detecting fake images. This test statistic follows a chi-square distribution, and its threshold for significance depends on the number of studies (or degrees of freedom). A significant Q-value implies heterogeneity.

The I^2 value further quantifies this heterogeneity: about 53% of the observed variability in the effect sizes is due to genuine differences among studies rather than random sampling error [24]. It affirms the credibility and reliability of the studies included in the meta-analysis because, despite the variability, they all make similar inferences regarding AI's ability to detect fake images. I^2 values of 25%, 50%, and 75% are considered low, moderate, and high heterogeneity, respectively. The I^2 value of ~53% in this study suggests moderate to high heterogeneity among the included studies.

3) Precision of the effect size (σ: 0.0562, lower bound: 1.6235, upper bound: 1.8439, CI: [1.6235, 1.8439]): These statistics provide insight into the precision and reliability of the z_+ value. The standard deviation (σ) is low, which suggests a precise estimate [11]. A low standard deviation is synonymous with minimal differences in the overall sentiment expressed by the 32 studies included in this meta-analysis. The confidence interval (CI) is also reasonably narrow, ranging from 1.6235 to 1.8439. While there are no standard σ values, a narrow CI, like in our case, denotes high precision [41]. Consequently, this further indicates that the pooled effect size is estimated with high precision [45].

4) Publication bias (Critical Nfs: 170, Nfs: 2706.9451): In meta-analyses, fail-safe N (Nfs) and critical Nfs are used to evaluate the potential for publication bias. The critical Nfs represents the minimum number of studies with null results required to raise the p-value above a significance threshold (typically 0.05). There is no standard value or range for this statistic. However, a higher Nfs than the critical value indicates that many unpublished, non-significant studies would be required to nullify the observed effect [40]. Additionally, the statistics suggest that the meta-analysis results are robust against possible publication bias [32].

5) Significance of the effect ($p<0.001$): This p-value indicates the probability that the observed effect (or a more extreme effect) would occur by random chance alone if there were no real effects. The study found a p-value less than 0.001. For a study conducted at a 0.05 confidence level, a statistical significance of anything lower than 0.05 is acceptable [43, 49]. Consequently, the value affirms that it is highly statistically significant. It further provides strong evidence against the null hypothesis that machine learning models can effectively detect fake images [23].

6) Implication: The meta-analysis results indicate a robust, positive, and highly significant overall effect size. The small confidence interval and standard deviation demonstrate

the estimated results' precision. However, substantial heterogeneity among the included studies necessitates additional research to determine the causes of this variation. The results appear robust against the possibility of publication bias. While the results appear trustworthy, future meta-analyses should consider the high heterogeneity and strive to reduce it.

V. Discussion

A. Effectiveness and Sustainability of Deep Learning Algorithms in Detecting Fake Images

The first research question regarded the effectiveness and sustainability of deep learning in detecting fake images. Our study discovered that the accelerated development of digital manipulation techniques has increased the difficulty of detecting fake or fabricated images [35, 36]. Deep learning algorithms, particularly convolutional neural networks (CNNs) and generative adversarial networks (GANs), have been widely cited as the leading instruments for addressing this concern [3, 44, 52]. Many sources we consulted emphasized that these algorithms frequently outperform conventional image analysis techniques, with many studies reporting accuracy rates and F1 scores exceeding 0.8 [24, 50]. High performance was most commonly observed in experiments employing a larger number of images as samples [53, 54]. However, even the lowest-performing studies' F1 scores did not fall below 0.84.

However, the revision of the sources disclosed a recurring theme. These algorithms' variable efficacy was based on the type and quality of the fake images they were trained on, although this was not the case in all studies. Moreover, the sustainability of these algorithms in the context of varying image types and qualities emerged as a significant consideration. Several sources alluded to sophisticated techniques for occasionally generating fake images that could circumvent deep learning detectors, raising concerns about the long-term sustainability of these detection methods. This pattern is notably evident when training data lacks diversity [9, 41, 44]. In addition, our investigation revealed that adversarial assaults on these algorithms present a challenging obstacle but also raise questions about their sustainable effectiveness [38, 42]. On the effectiveness and sustainability of deep learning algorithms in detecting fake images, this analysis finds that although deep learning is a promising avenue, it may require integration with other detection techniques to obtain optimal and sustainable results.

B. Evaluation Metrics for Deep Learning Algorithms in Sustainable Fake Image Detection

The second research question examined the evaluation metrics mostly employed to appraise the performance and sustainability of deep learning models in detecting fake images. Our exhaustive analysis highlighted the critical significance of dependable evaluation metrics for assessing the performance and sustainability of deep learning algorithms in detecting fake images [3, 37, 38]. Findings suggested that the most reliable metric is the F1 score, though most studies also engaged with other metrics to be more comprehensive. Most of the studies we evaluated emphasized indicated that the F1 score is not comparable to accuracy because it is an aggregate of precision and recall, thereby giving it an edge over other performance measures [7].

Most studies employed precision, recall, and the F1-score metric to avoid using accuracy. The performance measure (F1 score) provides a more comprehensive perspective on the efficacy of an algorithm [24]. This meta-analysis utilized the F1 score as its primary metric and was also one of the selection criteria for the selected studies. The harmonic mean property of the F1-score, which considers both precision and recall, is useful when an optimal equilibrium between false positives and false negatives is essential. Using only precision or recall is frequently deemed insufficient, as it conceals a crucial aspect of model performance [36, 46].

C. Technical Challenges in the Sustainable Detection of Fake Images using Deep Learning

The final research question interrogated the technical challenges experienced during fake image detection using deep learning approaches. Specialists frequently face several technical obstacles when utilizing deep learning technologies for fake image detection, as uncovered by our exhaustive analysis [1, 42, 47]. Additionally, the sustainability of these technologies in the face of evolving threats and techniques is a critical concern. Many sources we consulted elaborated on the difficulty of adversarial assaults [34, 39, 42]. Typically, these assaults are ingeniously designed perturbations that not only pose a technical challenge but also raise sustainability issues, as they can trick deep learning models into misclassifying a fake image as authentic or vice versa [9, 48]. This issue illustrates the vulnerability and potentially limited sustainability of these algorithms under specific conditions. While many of the sources devised mechanisms to circumvent adversarial assaults, they acknowledged that such complexities could have a significant impact on the performance and sustainability of the models. In addition, the dynamic nature of image manipulation techniques necessitates constant model updates, highlighting the need for sustainable development practices in AI, as current methods may become obsolete in the face of newer and more complex image manipulation techniques.

In addition, most sources mentioned an 'arms race' between fake image generators and detectors, highlighting a sustainability challenge in this technological contest. As technologies for deep learning evolve, so do techniques for generating fake images, resulting in a continuous and potentially unsustainable development cycle for both [5, 14]. Given the novelty of both disciplines, it is uncertain which will ultimately prevail, raising concerns about the long-term sustainability of detection algorithms. This swift evolution underscores the need for sustainable development practices in the field. Our investigation also revealed that data deficiency hinders model training capabilities [35, 38]. This case is notably true for labeled datasets containing high-quality fake images. Consequently, detection model efficacy declines. The complexity of deep learning models also poses computational and sustainability difficulties [6, 36]. The typical solution is to demand substantial resources for instruction and deduction, which may not be sustainable, especially for real-time

applications that require efficient and environmentally conscious approaches.

VI. Conclusion and Future Research

Identifying fake images has become a top concern in the wake of a highly advanced era, underscoring the necessity for sustainable methods in digital content management. Numerous individuals with questionable motives have utilized technology to fabricate misleading photos and manipulate unsuspecting individuals. The prevalence of inaccurate information in online spaces has heightened the need to eradicate this problem through sustainable approaches. Our comprehensive analysis showcases the growing potential of deep learning technologies, particularly convolutional neural networks (CNNs) and generative adversarial networks (GANs), in addressing this issue sustainably. While these models have shown promising outcomes and made a notable difference, they encounter challenges in maintaining sustainability in their applications. It is important to be cautious when generalizing study conclusions due to the differences in methodology, types of images, and geographical factors, and sustainability considerations. The F1 score is necessary to evaluate how well these algorithms perform on the modified images they are trained on, with an emphasis on quality, diversity, and sustainability.

However, the landscape of fake image detection is fraught with obstacles that exceed the capabilities of algorithms, demanding sustainable solutions. As a result of adversarial assaults, there is an ongoing arms race between fake image creators and their detectors, demanding sustainable solutions. In addition, the frequent need for model updates, computational demands, and data scarcity indicate the need for ongoing research efforts. As technology advances, image manipulations become more complex, increasing the demand for powerful, adaptable, and sustainable deep-learning solutions. We need to work together and combine the knowledge and expertise of academia, industry, and policymakers, to develop effective and sustainable strategies to better protect ourselves against fake imag

References

[1] T. Goel, R. Murugan, S. Mirjalili and D. K. Chakrabartty, "OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19," Applied Intelligence, vol. 51, no. 3, pp. 1351-1366, 2021.

[2] N. K. Chowdhury, M. M. Rahman and M. A. Kabir, "PDCOVIDNet: a parallel-dilated convolutional neural network architecture for detecting COVID-19 from chest X-ray images," Health information science and systems, vol. 8, no. 1, pp. 1-14, 2020.

[3] WHO, "WHO Coronavirus (COVID-19) Dashboard," 17 December 2021. [Online]. Available: https://covid19.who.int/. [Accessed 17 December 2021].

[4] D. Singh, V. Kumar and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks," European Journal of Clinical Microbiology & Infectious Diseases, vol. 39, no. 7, p. 137, 2020.

[5] M. Elgendi, M. U. Nasir, Q. Tang, R. R. Fletcher, N. M. C. Howard and S. Nicolaou, "The performance of deep neural networks in differentiating chest X-rays of COVID-19 patients from other bacterial and viral pneumonias," Frontiers in Medicine, vol. 7, no. 1, p. 550, 2020.

[6] J. Civit-Masot, F. Luna-Perejón, M. Domínguez Morales and A. Civit, "Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images," Applied Sciences, vol. 10, no. 13, p. 4640, 2020.

[7] M. Shorfuzzaman and M. S. Hossain, "MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients," Pattern recognition, vol. 113, no. 1, p. 107700, 2021.

[8] L. Li, T. Shim and P. E. Zapanta, "Optimization of COVID-19 testing accuracy with nasal anatomy education," American journal of otolaryngology, vol. 42, no. 1, p. 102777, 2021.

[9] M. N. Esbin, O. N. Whitney, S. Chong, A. Maurer, X. Darzacq and R. Tjian, "Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for COVID-19 detection," Rna, vol. 26, no. 7, pp. 771-783, 2020.

[10] L. Wang, Z. Q. Lin and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," Scientific Reports, vol. 10, no. 1, pp. 1-12, 2020.

[11] M. Szmigiera, "Impact of the coronavirus pandemic on the global economy - Statistics & Facts," 23 November 2021. [Online]. Available: https://www.statista.com/topics/6139/covid-19-impact-on-the-global-economy/#dossierKeyfigures. [Accessed 17 December 2021].

[12] S. A. Quadri, "COVID-19 and religious congregations: Implications for spread of novel pathogens," International Journal of Infectious Diseases, vol. 96, no. 1, pp. 219-221, 2020.

[13] M. Lipsitch and N. E. Dean, "Understanding COVID-19 vaccine efficacy," Science, vol. 370, no. 6518, pp. 763-765, 2020.

[14] M. A. Pettengill and A. J. McAdam, "Can we test our way out of the COVID-19 pandemic?," Journal of clinical microbiology, vol. 58, no. 11, pp. e02225-20, 2020.

[15] F. M. Salman, S. S. Abu-Naser, E. Alajrami, B. S. Abu-Nasser and B. A. Alashqar, "Covid-19 detection using artificial intelligence," First Journal of Biomedical Research, vol. 1, no. 1, p. 1, 2020.

[16] C. Shorten, T. M. Khoshgoftaar and B. Furht, "Deep Learning applications for COVID-19," Journal of big Data, vol. 8, no. 1, pp. 1-54, 2021.

[17] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons & Fractals, vol. 140, no. 1, p. 110120, 2020.

[18] G. Gilanie, U. I. Bajwa, M. M. Waraich, M. Asghar, R. Kousar, A. Kashif and H. Rafique, "Coronavirus (COVID-19) detection from chest radiology images using convolutional neural networks," Biomedical Signal Processing and Control, vol. 66, no. 1, p. 102490, 2021.

[19] C. Ouchicha, O. Ammor and M. Meknassi, "CVDNet: A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images," Chaos, Solitons & Fractals, vol. 140, no. 1, p. 110245, 2020.

[20] K. S. Lee, J. Y. Kim, E. T. Jeon, W. S. Choi, N. H. Kim and K. Y. Lee, "Evaluation of scalability and degree of fine-tuning of deep convolutional neural networks for COVID-19 screening on chest X-ray images using explainable deep-learning algorithm," Journal of Personalized Medicine, vol. 10, no. 4, p. 213, 2020.

[21] P. R. Bassi and R. Attux, "A deep convolutional neural network for COVID-19 detection using chest X-rays," Research on Biomedical Engineering, vol. 1, no. 1, pp. 1-10, 2021.

[22] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, G. Danala, Y. Qiu and B. Zheng, "Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms," International journal of medical informatics, vol. 144, no. 1, p. 104284, 2020.

[23] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," Physical and Engineering Sciences in Medicine, vol. 43, no. 2, pp. 635-640, 2020.

[24] A. Makris, I. Kontopoulos and K. Tserpes, "COVID-19 detection from chest X-Ray images using Deep Learning and Convolutional Neural Networks," in 11th Hellenic Conference on Artificial Intelligence, Athens, City Publishers, 2020, pp. 60-66.

[25] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha and N. Shukla, "COVID-19 detection through transfer learning using multimodal imaging data," IEEE Access, vol. 8, no. 1, pp. 149808-149824, 2020.

[26] S. Vaid, R. Kalantar and M. Bhandari, "Deep learning COVID-19 detection bias: accuracy through artificial intelligence," International Orthopaedics, vol. 44, no. 1, pp. 1539-1542, 2020.

[27] H. Benbrahim, H. Hachimi and A. Amine, "Deep transfer learning with apache spark to detect covid-19 in chest x-ray images," Romanian Journal of Information Science and Technology, vol. 23, no. S, SI, pp. S117-S129, 2020.

[28] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani and G. J. Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," Medical image analysis, vol. 65, no. 1, p. 101794, 2020.

[29] I. D. Apostolopoulos, S. I. Aznaouridis and M. A. Tzani, "Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases," Journal of Medical and Biological Engineering, vol. 1, no. 1, p. 1, 2020.

[30] A. M. Alqudah, S. Qazan, H. Alquran, I. A. Qasmieh and A. Alqudah, "Covid-19 detection from x-ray images using different artificial intelligence hybrid models," Jordan Journal of Electrical Engineering, vol. 6, no. 2, pp. 168-178, 2020.

[31] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," Expert Systems with Applications, vol. 164, no. 1, p. 114054, 2021.

[32] N. W. S. Saraswati, N. W. Wardani and I. G. A. A. D. Indradewi, "Detection of Covid Chest X-Ray using Wavelet and Support Vector Machines," Int. J. Eng. Emerg. Technol, vol. 5, no. 2, pp. 116-121, 2020.

[33] A. Saygılı, "Computer-Aided Detection of COVID-19 from CT Images Based on Gaussian Mixture Model and Kernel Support Vector Machines Classifier," Arabian Journal for Science and Engineering, vol. 1, no. 1, pp. 1-19, 2021.

[34] D. C. R. Novitasari, R. Hendradi, R. E. Caraka, Y. Rachmawati, N. Z. Fanani, A. Syarifudin and R. C. Chen, "Detection of covid-19 chest x-ray using support vector machine and convolutional neural network," Commun. Math. Biol. Neurosci, vol. 1, no. 1, p. 202, 2020.

[35] G. Van Houdt, C. Mosquera and G. Nápoles, "A review on the long short-term memory model," Artif. Intell. Rev, vol. 53, no. 8, pp. 5929-5955, 2020.

[36] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, no. 1, p. 132306, 2020.

[37] Ş. Öztürk and U. Özkaya, "Residual LSTM layered CNN for classification of gastrointestinal tract diseases," Journal of Biomedical Informatics, vol. 113, no. 1, p. 103638, 2021.

[38] M. Z. Islam, M. M. Islam and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," Informatics in medicine unlocked, vol. 20, no. 1, p. 100412, 2020.

[39] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana and S. Alhyari, "COVID-19 prediction and detection using deep learning," International Journal of Computer Information Systems and Industrial Management Applications, vol. 12, no. 1, pp. 168-181, 2020.

[40] F. Demir, "DeepCoroNet: A deep LSTM approach for automated detection of COVID-19 cases from chest X-ray images," Applied Soft Computing, vol. 103, no. 1, p. 107160, 2021.

[41] H. Naeem and A. A. Bin-Salem, "A CNN-LSTM network with multi-level feature extraction-based approach for automated detection of coronavirus from CT scan and X-ray images," Applied Soft Computing, vol. 113, no. 1, p. 107918, 2021.

[42] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman and P. R. Pinheiro, "Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection," Ieee Access, vol. 8, no. 1, pp. 91916-91923, 2020.

[43] T. Mahmud, M. A. Rahman and S. A. Fattah, "CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization," Computers in biology and medicine, vol. 122, no. 1, p. 103869, 2020.

[44] G. Jain, D. Mittal, D. Thakur and M. K. Mittal, "A deep learning approach to detect Covid-19 coronavirus with X-ray images," Biocybernetics and biomedical engineering, vol. 40, no. 4, pp. 1391-1405, 2020.

[45] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," Informatics in Medicine Unlocked, vol. 19, no. 1, p. 100360, 2020.

[46] A. I. Khan, J. L. Shah and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," Computer Methods and Programs in Biomedicine, vol. 196, no. 1, p. 105581, 2020.

[47] K. K. Singh, M. Siddhartha and A. Singh, "Diagnosis of coronavirus disease (covid-19) from chest x-ray images using modified xceptionnet," Romanian Journal of Information Science and Technology, vol. 23, no. 657, pp. 91-115, 2020.

[48] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," Computers in biology and medicine, vol. 121, no. 1, p. 103792, 2020.

[49] B. Abraham and M. S. Nair, "Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier," Biocybernetics and biomedical engineering, vol. 40, no. 4, pp. 1436-1445, 2020.

[50] A. Narin, C. Kaya and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," Pattern Analysis and Applications, vol. 1, no. 1, pp. 1-14, 2021.

[51] M. Z. Che Azemin, R. Hassan, M. I. Mohd Tamrin and M. A. Md Ali, "COVID-19 deep learning prediction model using publicly available radiologist-adjudicated chest X-ray images as training data: preliminary findings," International Journal of Biomedical Imaging, vol. 1, no. 1, p. 1, 2020.

[52] S. Toraman, T. B. Alakus and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," Chaos, Solitons & Fractals, vol. 140, no. 1, p. 110122, 2020.

[53] S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo and H. Lee, "Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging," Frontiers in medicine, vol. 7, no. 1, p. 427, 2020.

[54] S. Hassantabar, M. Ahmadi and A. Sharifi, "Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches," Chaos, Solitons & Fractals, vol. 140, no. 1, p. 110170, 2020.