IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 August 2024, accepted 1 October 2024, date of publication 9 October 2024, date of current version 30 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3477257

SURVEY

A Review of Deepfake Techniques: Architecture, Detection, and Datasets

PETER EDWARDS, JEAN-CHRISTOPHE NEBEL, (Senior Member, IEEE), DARREL GREENHILL, AND XING LIANG, (Member, IEEE)
School of Computer Science and Mathematics, Kingston University, KT1 2EE London, U.K.
Corresponding author: Xing Liang (x.liang@kingston.ac.uk)

ABSTRACT Driven by continuous advancements in artificial intelligence, especially deep learning, the level of realism associated with deepfake technology continues to improve year after year, which poses unprecedented challenges to the field of deepfake detection. The boundary between what we as humans can detect as real or fake becomes evermore blurred as new generations of algorithms such as Dall-E 3 and Stable Diffusion are released. This paper provides a comprehensive study into the landscape of deepfake detection, exploring in-depth the key challenges, recognising recent successes, and suggesting promising avenues for future research. A meta-literature review is conducted to identify the current challenges and future directions, which form the foundation of this work. They are investigated by analysing state-of-the-art research with a focus on the key components that are crucial to the design of a deepfake detector, i.e., the architecture, detection methods and datasets. A major challenge identified by this study is the lack of dataset diversity leading to unfair attribute representation. This must be addressed by improving standardisation on dataset ethics and privacy. This is one of the main reasons for the insufficient generalisation capability of current deepfake detectors as demonstrated by their unsatisfactory performance when faced with unseen data or data in the wild. This literature review provides deepfake detection researchers and practitioners with the latest information that will serve as a vital resource for their continued and important activity, now and in the future.

INDEX TERMS Deepfakes, deepfake detection, generative AI, deep learning, machine learning, artificial intelligence, datasets, survey.

I. INTRODUCTION

The technology behind deepfake media has come a long way since its first inception in 2017, where, according to a leading UK newspaper [1], the term originated from a social media user who created a series of pornographic videos by swapping the faces with those of celebrities. At first, the novelty of this technology led to the release of several face-swapping apps, including Facelab [2] and FaceApp [3], which allowed people to generate content for entertainment value. However, the world is starting to see the full potential of this technology and the darker side in which it may be used to cause harm. Advancements in Artificial Intelligence (AI), principally in the fields of Machine Learning (ML) and Deep Learning (DL), have been instrumental in the acceleration of this technology, contributing to the spread of fake media throughout our society.

Early incarnations of deepfake media were primitive and often associated with static imagery that was of low quality. In recent years, this has shifted towards higher-quality imagery and video content due to improvements in DL model training and the sharing of open-source algorithms for content generation. We are now approaching a pivotal point where the lines between real and fake media are becoming blurred. News headlines have illustrated the severity and risks involving high-profile public figures in which fraudulent acts depict disinformation through fake news. Evidence has highlighted that, during the conflict between Russia and Ukraine, in 2022, a deepfake of Ukrainian President Zelensky was distributed to incite the citizens of Ukraine to surrender [4]. Doubts over

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

the integrity and authenticity of the video ensured that it failed its objective. This deepfake was one of many examples where such technology, if in the wrong hands, can become a weapon in the digital age that we live in.

Academic research in the field of deepfakes has significantly grown since 2017, as illustrated in Figure 1 which shows statistical data collected from Dimensions [5] using the publication type 'Article' or 'Preprint' and a date range from '2017' to '2024'. Note that a linear trendline was used to extrapolate them until 2025. Unsurprisingly, considering the mass media attention around deepfakes and fake news, the statistics highlight that the volume of published research on deepfake detection is far outpacing research on deepfake creation illustrating the demand for solutions on this controversial topic.

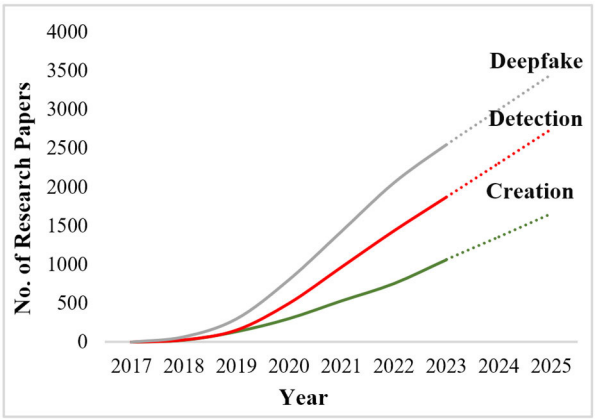

FIGURE 1. Statistics depicting the number of deepfake, deepfake creation or deepfake detection studies published over the last seven years (2017 to 2023).

Prior to deepfakes, the detection of manipulated imagery often focused on the semantic characteristics within the image in terms of what can be seen and its overall composure. For example, research by O'Brien and Farid [6] focused on the vanishing point in an image to establish the relationship with common reflections and determine the feasibility of the image containing forged content. Lighting and shadow details within the image are other examples of inconsistencies that may occur. Furthermore, Johnson and Farid [7] report that lighting and cast shadows can be used to approximate if the lighting source is consistent for all the objects within the image and, therefore, with reasonable accuracy, identify if manipulation has occurred. These techniques have successfully transitioned to the task of deepfake detection, with promising results. For example, Wu et al. [8] explain that subtle clues in the swapped face region can expose inconsistencies that do not match the composure of the image as a whole and are often the unintentional result of the deepfake creation pipeline. Additionally, this technique can be effective against video content as explained by Zhu et al. [9], where inconsistencies in the inter-frame sequence highlight abnormalities. However, quality and detail improvements in image and video media have led to the need for a more robust approach.

Artefacts uncovered in the spatial and frequency domain have exposed vital information relating to either the pixel formation (spatial domain) that makes up the overall image over time or the frequency representation, e.g., low or high-frequency components (frequency domain), which corresponds to the rate at which the pixel information is changing. For example, a disturbance in the surrounding pixel formation between the old and new content can provide valuable statistical information that may expose the boundary of where the manipulation occurred [10]. Additionally, facial blending inconsistencies, which are inherently transferred by the synthesis process during the creation of a deepfake, leave traceable artefacts in the image statistics [11]. The camera model NoisePrint has demonstrated success in applying the Photo-Response Non-Uniformity technique to extract and compare noise signatures from images in a manner similar to extracting a person's fingerprint [12]. Furthermore, utilising the spatial and frequency domain as handcrafted features for ML has paved the way forward with novel detection methods capable of inferring information based on complex information with limited human interaction. Typically used with a binary classifier and a fully connected layer, the process of selecting the features to be learned can present challenges, particularly when the underlying pipeline for deepfake creation is continuously evolving. Indeed, this can result in poor generalisations on unseen data.

An important milestone in the detection of deepfakes has been made possible through DL, whereby a model is able to learn complex multi-dimensional patterns from complex datasets using artificial neurons that replicate the way in which the human brain works [13]. In addition, this enables a richer representation of features to be learned in a way that would not be achievable through standard ML [14].

Since the taking off of DL before 2012, DL architectures have rapidly evolved, driving significant advancements in deepfake research, which can be observed in Figure 2. This progress began with Convolutional Neural Networks (CNNs), including VGG, CapsNet, which played a foundational role in shaping modern deepfake detection techniques and paved the way for future advancements with their novel approaches. However, as the years have advanced, increased levels of interest have shifted towards the development of hybrid architectures. In particular, this can be seen in the number of variant architectures, including Transformer and CapsNet, which have seen great interest from the academic community.

To establish the starting point, ten previously published literature review papers between 2022 and 2023 were evaluated in Section III to identify the key successes and challenges. Informed by them, four challenge themes, i.e., dataset, architecture and scalability, explainability and evaluation, were established to further complement this study and guide the reader towards the main themes. Section IV provides the reader with a high-level overview of the common

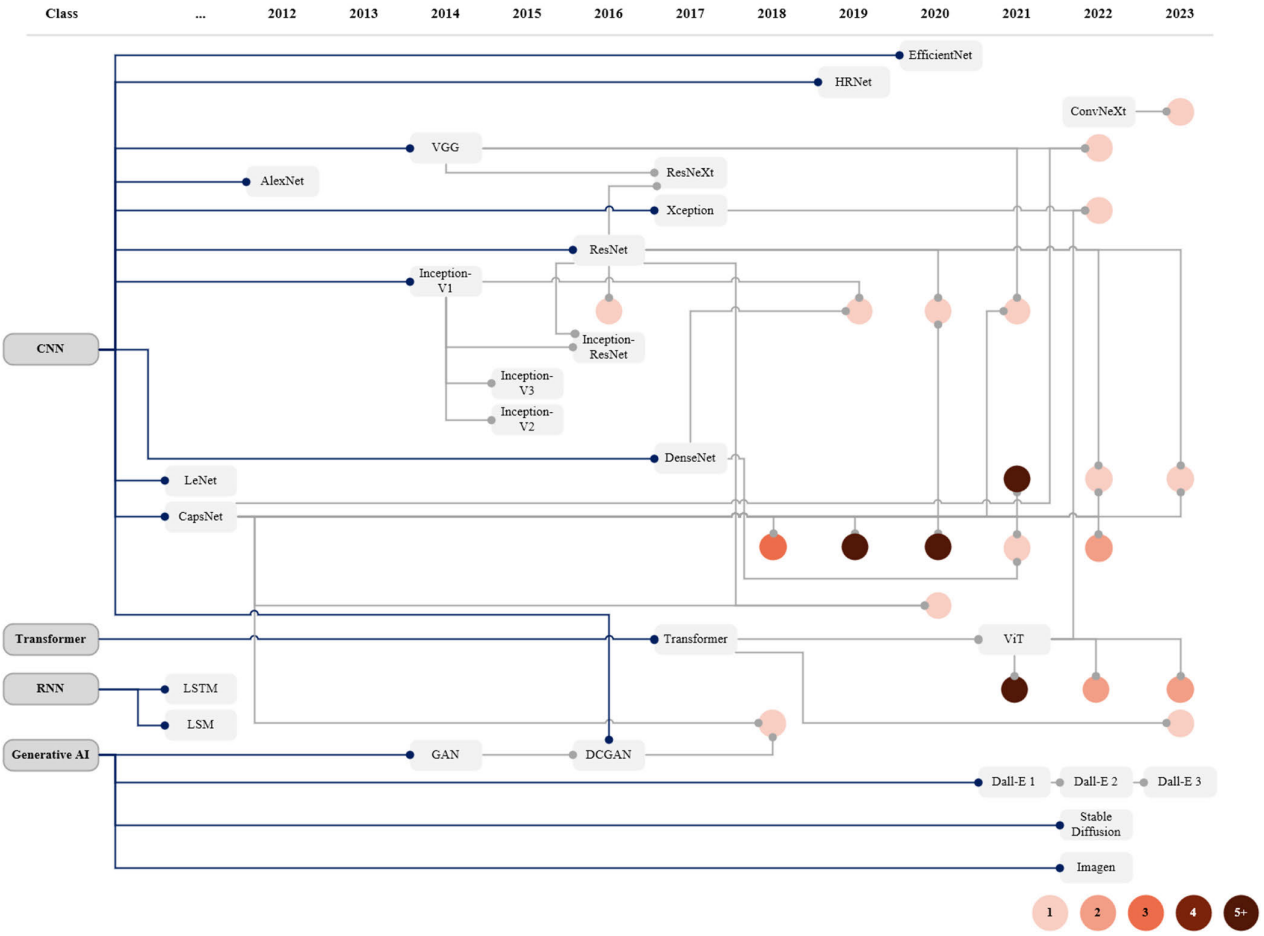

FIGURE 2. Timeline of architectures by class. The diagram highlights some of the main DL architectures and their associated variant architectures over time. The coloured bubbles indicate the number of research papers dedicated to new hybrid variants.

architectures used in deepfake detection whereas Figure 2 provides a timeline of the main architectural variants and shows where research activity has evolved. A comprehensive breakdown of the techniques used for deepfake detection is presented in Section V, followed by Section VI, which provides the reader with an overview of the datasets used for model training and evaluation. Finally, section VII reviews the observations and future trends exposed from the papers studied in Section V.

A. AIMS AND OBJECTIVES

The aim of this literature review is to provide the reader with an in-depth evaluation of the latest research on deepfake detection. This is achieved by exploring the various types of architectures and datasets that are used in order to understand how these crucial elements contribute to delivering State-of-the-Art (SOTA) detection. The primary objectives are as follows:

- Present an overview of key findings from recent literature reviews and produce a snap-shot view of the observed challenges associated with the key challenge themes from Section I-B.
- Evaluate the architectures used in deepfake detection and identify their strengths and weaknesses.
- Investigate and compare SOTA research papers on deepfake detection.
- Conduct a review of datasets used for deepfake detection and the impact they have on generalisation and bias.
- Identify and compare the key challenges against those observed in previous literature reviews.

B. CHALLENGE THEMES

To guide the categorisation and identification of current and future key challenges, this analysis is based on the most important themes highlighted in previous literature review papers (see Section III). Table 1 provides a breakdown of these challenges along with their description.

II. METHODOLOGY

This section provides an overview of the process used to undertake this review paper and details the research strategy and timeframe used.

A. RESEARCH STRATEGY

In order to cast a wide net over this research domain and deliver a comprehensive review and not be limited by any specific angle, no specific journal databases were used for the acquisition of research papers. In addition, since this review

TABLE 1. Challenge themes and their associated description.

Challenge Theme	Description
Dataset	Do publicly available datasets provide a fair representation of diversity (age, gender, ethnicity, etc.) in order to accurately train and evaluate models for deepfake detection? How does the practice of focusing on benchmarking algorithms against commonly used (or standard), but often 'old', datasets, instead of exploring new and innovative datasets, affect research progress? Are suitable measures put in place to handle pre-processing activities in terms of the ethical approach to collecting, processing and storing data?
Architecture and Scalability	Are known weaknesses or limitations sufficiently recognised? Are these addressed using suitable mitigations? With research shifting towards Deep Learning, are hybrid architectures able to overcome known design limitations while balancing the trade-off between architecture, computational resources and datasets? How can inference speed be expressed in relation to model complexity and computational resources?
Explainability	What efforts have researchers made to offer better model interpretability? Can the use of a confidence score or visual aid to highlight possible areas of manipulation help provide trust and explainability?
Evaluation	What performance metrics are considered important for evaluating a model's accuracy? How suitable is current consideration for conducting both intra and inter-dataset evaluations and the importance of demonstrating a model's ability in generalising to unseen data? Since the evaluation of the model's accuracy is artificial by nature, what consideration has been made for testing against data in the wild?

is focused on deepfakes detection-based methods, material specific to the creation pipeline of deepfakes was not included in the search strategy. Moreover, as the process for selecting detection papers was based on image and video techniques, publications associated with audio or text specifically were excluded.

B. TIMEFRAME

The following timeframe was applied throughout this review paper in order to ensure that a comprehensive review is conducted while providing only the most relevant and recent research.

- Meta-Literature Review (Section III) – due to the fast-moving pace of research in this field it was decided to prioritise the search of published literature review papers between 2022 and 2023 to provide coverage of the most recent work. Eventually, a subset of ten high-quality papers (see Table 2) was selected for review.
- Architecture (Section IV) – the search for papers on the subject of DL architectures was not limited by a date range but instead, the architecture needed to be associated with the learning of image media. The aim was to uncover the extensive range of research and the various key architectures and hybrid variants (see Figure 2 for a reduced diagram and Figure 5 in Appendix A, for a full diagram).
- Deepfake Detection (Section V) – using over one hundred papers, a subset of twenty-nine prominent papers published in 2023 were chosen for analysis. This was based on their novel approach or contribution to this research domain.
- Datasets (Section VI) – the search for papers on the subject of datasets was not limited by a date range but instead an extensive search was performed to identify as many key and novel datasets (see Figure 3 and Figure 4).

III. META-LITERATURE REVIEW

This section explores ten recently published review papers in the field of deepfake creation and detection, aiming to provide a snapshot view of the approach taken while understanding the common challenges and future directions. Table 2 provides a breakdown of the ten review papers selected, which cover the period between 2022 and 2023.

A. CURRENT CHALLENGES AND FUTURE DIRECTIONS

Based on the review papers listed in Table 2, current challenges and future directions have been clustered into the following four themes: dataset, architecture and scalability, explainability and evaluation. The aim is to cover the high-level trends affecting deepfake detection research while identifying ways to mitigate these challenges and support future directions.

TABLE 2. Recently published review papers (2022 and 2023).

Prior Review Papers	Research Type	Year	Generation Techniques	Detection Techniques	Image Datasets	Video Datasets	Challenge / Theme
Tyagi and Yadav [15]	Detection	2023	-	16	16	7	Dataset, architecture, explainability and scalability
Patil et al. [16]	Detection	2023	-	9	-	-	Explainability
Masood et al. [17]	Generation & Detection	2023	41	48	-	9	Dataset, architecture, explainability and scalability
Zanardelli et al. [18]	Detection	2023	-	13	11	3	Dataset, architecture and evaluation
Stroebel et al. [19]	Detection	2023	-	52	14	16	Dataset, architecture and evaluation
Nguyen et al. [20]	Generation & Detection	2022	-	24	-	-	Dataset, architecture and explainability
Seow et al. [21]	Generation & Detection	2022	50	44	-	11	Dataset, architecture and evaluation
Dagar and Vishwakarma [22]	Generation & Detection	2022	42	58	5	9	Dataset, architecture and explainability
Juefei-Xu et al. [23]	Generation & Detection	2022	91	118	11	16	Dataset and evaluation
Rana et al. [24]	Detection	2022	-	88	5	13	Dataset, architecture and evaluation

1) DATASET

The dataset theme is arguably the most important one, as this fundamentally determines how effective a trained model is at performing the task at hand. Access to good-quality data is therefore crucial. Although this often becomes a challenge for researchers, it can sometimes be overlooked.

A lack of availability to the public of global datasets, which provide a fair representation of forgery techniques and real media, has been observed as a critical limitation in the development of deepfake detection methods [15], [22], [23]. Researchers will often supplement their work by utilising custom datasets [18] and [23] for training and evaluation. However, they generally provide limited exposure to how effectively a model performs compared with other datasets and are often unavailable to the general public. A community-led global dataset [23] for AI-synthesised images could provide a global approach to benchmarking and evaluating model performance. Providing a structured approach [18] in the way in which data is used for training and evaluation would allow for improved measurable accuracy while offering a consistent and transparent approach. Existing datasets commonly used by researchers [24] for training and evaluation include FaceForensics ++ (FF++) and Celeb-DF. However, these alone do not provide enough of a challenge [23] to evaluate success as they do not represent current media found in the real-world.

The quality and size of the datasets are also observed [17], [19], [21], [22] as a key challenge, especially as some of the detection methods are trained on limited data [17] or data that is specific to a single creation technique [20]. A model's ability in generalising to unseen data becomes a greater challenge when the quality and size of the data are limited [19], this may lead the model to be ineffective against real-world data. This is particularly important in situations of fairness, whereby the risk of increasing bias towards certain attributes, including ethnicity, gender and age, has genuine consequences that can lead to racial profiling and discrimination if used in real-world applications [19]. The risk of adversarial attacks to avoid detection [21] is also raised as a serious concern, whereby the input permutations of the data are altered (sometimes referred to as data-poising) to prevent the model from learning true representations of the intended data. Moreover, the use of pre-processing techniques [9] has implications, particularly if the researcher has not fully understood the dataset in question.

The general consensus indicates that many of the proposed detection methods published are constrained to working in environments that do not reflect the real-world and therefore are likely to fail to generalise or infer correctly if applied to data in real-world applications. Continuous advances in technology [22] will eventually lead to the creation of full-body deepfakes, which will ultimately present new challenges

for managing new and emerging datasets going forward. As observed by Naitali et al. [25], the WildDeepfake [26] dataset provides imagery related to full-body deepfakes, yet the volume and diversity are still somewhat limited in their effectiveness for AI. Despite the limited amount of available full-body deepfake data, research by Hong et al. [27] highlights the progress being made to generate 3D moving full-body content, which in the future will aid the development of new datasets.

2) ARCHITECTURE

Early implementations of ML architectures exposed weaknesses in their ability to extract local and global features [15], making it difficult for the model to effectively learn from the selected handcrafted features. Indeed, selecting the relevant features [17] is a complex process, which has become more and more challenging as the quality of deepfake media improves. Combining DL approaches with traditional techniques (statistical analysis, for example) [18] has highlighted an overall improvement in model performance, suggesting that selecting handcrafted features is no longer sufficient for learning complex patterns in media content. A shift towards a hybrid approach [19], [21], where combining various architectures to overcome known limitations, has prompted a new research direction focused on DL. Although at present there is no architectural framework that provides a sufficiently stable platform, the Xception Network [18] has demonstrated promising results as a DL architecture. Indeed, recent literature demonstrates how DL approaches can outperform non-DL approaches [24]. For example, a multi-modal architecture is proposed to allow for combining features across audio, imagery and video media [17]. However, there is a lack of research on temporal aggregation [18] in video media, whereby the temporal consistency between frames is evaluated rather than providing a binary classification for the entire video sequence.

A negative trade-off in the success of DL approaches comes in the form of a steep rise in computational resource required as the models grow in complexity [15]. Fortunately, adopting pre-trained models based on existing architectures is one way of overcoming this challenge [21]. Pre-trained models can be fine-tuned for downstream tasking [21] while reducing the number of trainable parameters and overall computational effort. Managing a model's complexity to ensure its efficiency through optimisation is essential for the transition from research to capability deployment. However, ensuring optimised inference times [17] while expanding a model's size through new training data needs careful consideration. This is particularly true for DL architectures, where the number of parameters dramatically increases with network depth [15].

3) EXPLAINABILITY

The third and most challenging theme is explainability [15], [16], [17], [20], [22], as most of the proposed deepfake detection methods provide only a binary classification (real or fake) as their output. The lack of additional context around the model's decision process can lead to issues in trust and interoperability [22], which ultimately results in ambiguity [15] in understanding and evaluating a model's effectiveness. In particular, many of the reviewed methods fail to provide localised information [22] that would identify where the manipulation likely occurred. This is further compounded by the black-box [20], [22] nature in which a model is trained and tested and is therefore not adequate [16] in providing confidence to the user if an image or video is a deepfake. By addressing the black-box nature as a main area of research in DL, progress is being made that will eventually support deepfake detection. One way of adding explainability could be through the use of a multi-class classification combined with a confidence score [17], as this would present the user with the power to make an informed decision.

4) EVALUATION

The absence of a structured and uniform approach to evaluate, [19] and [23], deepfake detection methods is observed, with an emphasis on the creation of a consistent approach to benchmarking against other SOTA techniques. Although the accuracy metric [24] is identified as the most common measurement, this does not take into account the multitude of existing metrics, including precision, recall, and the Area under Receiver Operating Characteristic (AUC-ROC) Curve, which provide valuable insight into the state of a given model. In addition, reported evaluation metrics are at times overinflated [23] and do not clearly reflect the way in which the model was trained or evaluated, which in turn leads to inconsistencies in the author's work. As the measure of success is often based on achieving higher accuracy against other SOTA techniques while benchmarking against common datasets [21], there is little to no attention to real-world scenarios. The fact that many of the common datasets contain outdated image and video content, that does not adequately reflect new and emerging deepfake technology [19], diminishes the models' ability to perform well against data from the wild.

IV. ARCHITECTURE

To put into context the architectures that have been exploited in deepfake detection, the following section provides an overview of the architectures commonly used in the computer vision domain. Figure 2 provides a timeline of the common DL architectures. The coloured bubbles within the figure indicate where increased research has been dedicated to create new hybrid implementations based on a specific class of architecture

A. CONVOLUTIONAL NEURAL NETWORK

One of the most prominent contributions to research on Convolutional Neural Networks (CNNs) was conducted by LeCun et al. [28] in 1989 to overcome the challenge of performing numerical character recognition and classification

through ML. This subsequently led to the development of the LeNet-5 architecture in 1998 by LeCun et al. [29], which consisted of a seven-layer network using three convolution layers (a kernel of 5×5) with feature maps of size: 6, 16 and 120, two subsampling layers with 2×2 receptive fields for average pooling, and finally two fully-connected layers. According to LeCun et al. [29], the use of back-propagation in the network can help reduce the number of trainable parameters since the weights in the feature extractor can be learned through a shared scheme.

1) CAPSULE NETWORK

Proposed in 2011 by Hinton et al. [30], the Capsule Network (CapsNet) was designed to overcome a significant limitation in the way complex spatial relationships are learned by conventional CNNs. According to Hinton et al. [30], CNNs are non-equivariant and therefore do not take into account the precise positional relationship between facial features in both the local and global feature space. In essence, a CNN will learn the arrangement of facial features (eyes, nose, mouth, etc.) as individual components, respectively, and will not take into account the location of neighbouring facial components. To learn the global feature space, Kwabena Patrick et al. [31] explain how each capsule contains a series of neurons that focus on learning from the same feature space while individually learning distinctive properties from the feature. Further to this, the CapsNet uses two layers of capsules, represented as the lower and higher-level capsules, whereby the output from each capsule is in the form of a vector. Interest from the research community to modify and improve the architecture can be seen in Figure 2 where the growth of hybrid variants has increased between 2018 and 2021.

2) INCEPTION NETWORK

Szegedy et al. [32] proposed the Inception network in 2014 using a sparsely connected architecture combined with a CNN to not only achieve greater network depth but to also overcome the demand on computational resources. According to them, significant growth in network parameters can occur as a consequence of increasing the number of fully connected layers within the architecture due to the connectivity between each and every layer across the network. They suggest applying dimensionality reduction through a series of max and average pooling operation aggregations to help reduce the number of expensive computational operations while preventing the transfer of a large number of filters between the layers within the network. The authors claim a key design motivation is the scalability of the architecture to run on devices with potentially low computational resources, making it more viable for deployment in real-world situations.

3) XCEPTION NETWORK

In 2017, Chollet [33] presented the Xception network as a parameter-efficient adaptation to the Inception-V3 [34] architecture, where stackable depthwise separable convolutions are used as a replacement for the Inception module while achieving notably improved performance. According to them, the original hypothesis for the Inception network was based on the idea that cross-channel and spatial correlations should not be mapped together in the feature map as they are sufficiently separated. However, in contrast to this, they observe that by using a CNN, the cross-channel and spatial correlations can be separated in the feature map. Comparative testing on the ImageNet dataset concludes only marginal improvement in performance, yet the authors highlight that the parameter size is similar to the Inception-V3 network and that any improvement is likely to be contributed to the efficient use of the parameters and not the depth of the network.

4) DENSE CONVOLUTIONAL NETWORK

The Dense Convolutional Network, Huang et al. [35], was proposed in 2017 as a stackable feed-forward network designed as a more efficient and deeper CNN, where dense blocks act as the building blocks in the network. Inspired by the use of skip-connections in the Residual Neural Network (ResNet) architecture to overcome the challenge of gradient flow, the authors propose direct connections from any layer to all preceding layers, thereby removing the gradient flow problem. Furthermore, a convolution and down-sampling layer using average pooling is positioned between each of the dense blocks. The paper highlights competitive results against other SOTA architectures on tested evaluation datasets, with the potential to provide improved learning through feature sharing across the network.

B. TRANSFORMER

The Transformer was proposed by Vaswani et al. [36] as an alternative to CNN and Recurrent Neural Network architectures in 2017 and is based on the use of self-attention. According to them, the design incorporates an encoder and decoder structure and is based on a fully connected feed-forward network. Using learned embeddings, the input and output tokens are converted to vectors, with additional positional information embedded.

The Vision Transformer (ViT) was proposed by Dosovitskiy et al. [37] in 2021 and builds on the success of the Transformer for Natural Language Processing. Using Self-Attention as a key component, the architecture works by feeding an input image as a sequence of fixed-sized, one-dimensional patches with embedded positional information. The sequencing is additionally flattened before being linearly projected to a Transformer Encoder using Multiheaded Self-Attention. Furthermore, the classification is performed using a Multi-Layer Perceptron block. Even though impressive results are observed on the evaluated datasets, the authors highlight the lack of image-related inductive bias when trained on smaller datasets, which can result in overfitting or poor generalisation. However, the authors observe that on larger datasets, the model learns

the patterns from the data itself, and therefore inductive bias becomes less essential. Similarly, the ViT has gained increased interest from the research community as illustrated in Figure 2 where continuously hybrid variants have been developed.

C. GENERATIVE AI

The ability to generate realistic content based on generative AI has accelerated over the past several years with the introduction of advanced DL models that are capable of generating text, image, video, and audio content [38]. Dall-E 3 [39], Imagen [40] and Stable Diffusion [41] are just some of the prominent models available, which can produce new samples based on the patterns learnt by the models [42].

1) GENERATIVE ADVERSARIAL NETWORK

The concept of training two simultaneous models that compete against each other was proposed by Goodfellow et al. [43] in 2014 and is known as the Generative Adversarial Network (GAN). The two models, a generative (creator) and a discriminator (detector), are trained against each other, where the aim of the discriminator is to determine by estimating the probability that the sample produced by the generative came from either the training dataset or from the generative model itself. According to them, a key design feature is the implementation of backpropagation in the network for improved gradient flow, which is an alternative to the use of Markov chains. Furthermore, forward propagation can be used to sample from the generative model.

2) DIFFUSION MODELS

Diffusion Models (DMs), Yang et al. [44], are based on a type of probabilistic generative model that works by introducing noise into the input image as it traverses through the network in a forward pass. To generate new content, the model must learn to reconstruct the image by removing the noise, which is known as the diffusion process. According to the authors, the three key types of DMs are Denoising Diffusion Probabilistic Models [45], Score-based Generative Models [46] and Stochastic Differential Equations [47].

V. DEEPFAKE DETECTION

The aim of this section is to explore the most current literature on deepfake detection in order to establish a baseline view of the current SOTA methods available. Table 3 provides a breakdown of twenty-nine novel methods from literature in 2023 and each represents a unique approach to deepfake detection.

A. CONVOLUTIONAL NEURAL NETWORK

To improve generalisation, the authors in [51] use a ResNet18 as the base architecture and combine a K-Nearest Neigbors algorithm for the feature classification. To enhance the architecture, they implemented Error Level Analysis during the pre-processing stage and supplied it as the input to the ResNet18 for feature extraction. Although better generalisation was observed, the study is limited by the use of static image data. Indeed, analysis of temporal inconsistencies could deliver greater robustness. Study [53] employs a patch-based approach using the Gram-Net architecture proposed in [78]. According to the authors in [78], the Gram-Net incorporates a Gram Block positioned before the down sampling layer of a ResNet backbone to learn global features. However, it is worth considering that the reported 100% accuracy in the initial evaluation of the first two datasets [53] might require further validation to ensure robustness. In [60], the authors identify a potential weakness in the CNN architecture that could lead to spatial information being lost because of the design of the max pooling layer, whereby the model becomes unable to learn truthful representation from the original image features. Thus, they hypothesise that by adapting a Visual Geometry Group (VGG)-19 network as a feature extractor, spatial information in the lower layers of the network can be retained. To emulate how the human brain works, a CapsNet is used for the classification task as it uses neurons to learn key relationships between facial components and their orientation. Eventually, their new architecture delivers not only enhanced performance and convergence speed but also reduces the model complexity. In [55], the authors also address CNN's insufficiency in extracting spatial features, whilst bringing to attention the need for further research in temporal feature extraction. They tackle these shortcomings, by implementing pre-processing steps to enhance image quality using a Gaussian noise reduction and employing the Lucas-Kanade optical flow algorithm. As in previous work, the authors finalise their design using the VGG with a CapsNet to which they include a modified Dynamic Routing Algorithm. They report significant performance gains over other SOTA methods.

Methods to extract noise information from the frequency domain have demonstrated significant progress in the field of deepfake detection. Restoration techniques to restore low-quality artefacts from compressed images could prove valuable, especially when access to high-quality deepfakes is often limited. The authors in [73] consider magnitude and phase spectra to capture contour and textual information and their spectral relationship within the frequency domain. The design incorporates both a Regular and Irregular Local Fourier to increase local information, whilst utilising a Pointwise convolution to overcome the challenge of gradient loss and explosion. To extract multiple features and capture their interactions, the authors incorporate a cross-attention mechanism with multiple Binary Cross-Attention blocks. Although evaluation results demonstrate comparative results against other SOTA methods, particularly in cross-dataset evaluations, the model underperforms as the level of image compression increases through post-processing operations. To address this, it is proposed [62] to achieve greater robustness by leveraging low and high features that are often destroyed during the post-processing operation with a framework called TruFor [62]. Using a modified Noiseprint [12], they intend to expose essential image information about the

TABLE 3. List of deepfake detection methods reviewed in this study. NOTE: where applicable an average value for ACC and AUC was applied.

Reference	Concept	Architecture	Training Dataset	Evaluation Dataset	Performance ACC	Performance AUC
Cozzolino et al. [49]	Person-of-Interest (POI)	ResNet50	VoxCeleb2	DFDC Preview, FakeAVCelebV2, KoDF, DF-TIMIT	**86.70** 86.60 81.10 85.70	95.20 94.10 89.90 **99.20**
Wang et al. [50]	-	CNN + ViT	FF++	FF++, DFDC, Celeb-DF, DeeprForensics-1.0	**92.11** 65.76 63.27 62.46	**97.66** 73.68 72.43 78.19
Rafique et al. [51]	-	ResNet18 + KNN	Unknown	RFFD	**89.50**	
Heo et al. [52]	-	EfficientNet-B7 + ViT	DFDC	DFDC, Celeb-DF V2,		97.80 **99.30**
Soleimani et al. [53]	-	Gram-Net	Style/StarGAN + CelebA + FFHQ, StyleGAN + CelebA, StyleGAN2 + FFHQ, StyleGAN2 + FFHQ, PGGAN + FFHQ,	Style/StarGAN + CelebA + FFHQ, StyleGAN + CelebA, StyleGAN2 + FFHQ, StyleGAN2 + FFHQ, PGGAN + FFHQ	**100.0** **100.0** 99.70 90.50 84.90	
Ahmed and Sonuç [54]	-	CNN	Unknown	DFDC	**95.77**	
Lu et al. [55]	iCapsNet-TSF	VGG + CapsNet	Unknown	FF++, Celeb-DF	**94.07**	
Shao et al. [56]	DeepFake-Adapter	ViT	FF++	FF++, Celeb-DF, DFDC, DeeperForensics-1.0		**97.77** 71.74 72.66 86.50
Khalid et al. [57]	DFGNN	GNN + ResNet	FF++, Celeb-DF, DFDC Preview, WLRD	FF++, Celeb-DF, DFDC Preview, WLRD	**97.16** 87.60	**98.00** 95.00 92.00 92.00
Ke and Wang [58]	DF-UDetector	EfficientNet	FF++, Celeb-DF, DFDC, WildDeepfake	FF++, Celeb-DF, DFDC, WildDeepfake	 **78.38**	**81.32** 79.22 77.80
Wang et al. [59]	SFDG	GCN + EfficientNet-B4	FF++	FF++, Celeb-DF, DFDC, DeeperForensics-1.0, DFD, WildDeepfake	95.23 **99.22** 84.41	97.75 **99.96** 73.64 92.10 88.00 92.57
Ilyas et al. [60]	E-Cap Net	VGG-19 + CapsNet	FF++, DFFD, WLRD	FF++, DFFD, WLRD		99.52 **99.99** 98.31
Usmani et al. [61]	-	ViT	RFF, RFFD	RFF, RFFD	**92.15** 88.52	
Guillaro et al. [62]	TruFor	DNCNN + Transformer	CASIA 2.0, FantasticReality, IMD2020, COCO, RAISE	CASIA 1.0, COVERAGE, Columbia, NC2016, DSO-1, VIPP, OpenForensics, CocoGlide	81.30 68.00 **98.40** 66.20 93.00 76.10 63.90	91.60 77.00 **99.60** 76.00 98.40 82.00 75.20
Xu et al. [63]	TALL - Swin	Swin Transformer	Unknown	FF++, Celeb-DF, DFDC, DeeperForensics-1.0	**95.73**	97.22 90.79 76.78 **99.62**
Wang et al. [64]		Xception	ImageNet	FF++, WildDeepfake	**94.54** 84.08	**96.40** 91.33
Anas Raza et al. [65]	HolisticDFD	CNN + Transformer	FF++, DFDC, Celeb-DF	FF++, Celeb-DF, DFDC		94.15 **96.24** 92.60
Zhao et al. [66]	ISTVT	Xception + Transformer	FF++, Celeb-DF	FF++, Celeb-DF, DFDC, DeeperForensics-1.0, FaceShifter	97.57 **99.80** 92.10	 84.10 74.20 98.80 **99.30**

TABLE 3. *(Continued.)* **List of deepfake detection methods reviewed in this study. NOTE: where applicable an average value for ACC and AUC was applied.**

Tian et al. **[67]**	-	Xception	ImageNet, FF++	FF++, Celeb-DF, DFDC	97.74 | **99.27** **99.85** | 99.99 69.17 |
Yang et al. **[68]**	AVoiD-DF	ViT	DefakeAVMiT, FakeAVCeleb, DFDC	DefakeAVMiT, FakeAVCeleb, DFDC	**95.30** | **97.60** 83.70 | 89.20 91.40 | 94.80
Liang et al. **[69]**	-	U-Net + Swin Transformer	FF++, Biwi Kinect Head Pose Database	FF++, Celeb-DF	**94.76** | **96.80** | 72.30
Li et al. **[70]**	ADAL	GAN	FF++, Celeb-DF V2, DFDC	FF++, Celeb-DF V2, DFDC, DFD	93.22 | 95.38 94.51 | 97.37 **96.47** | **98.23** 87.47 | 92.14
Wang and Chow **[71]**	NoiseDF	Siamese Network + RIDNet	FF++	FF++, Celeb-DF, DFDC, DeeperForensics-1.0	**84.36** | **93.99** 70.10 | 75.89 59.87 | 63.89 67.49 | 70.88
Khormali and Yuan **[72]**	-	GCN + Transformer	FF++, Celeb-DF V2, WildDeepfake	FF++, Celeb-DF V2, DFDC, WildDeepfake, DeeperForensics-1.0, FaceShifter, Celeb-DF	97.46 | 98.15 **99.47** | **99.43** | 77.30 81.37 | 81.24 | 98.90 | 99.10 | 87.90
Yang et al. **[73]**	FDS_2D	CNN	FF++, Celeb-DF	FF++, Celeb-DF, DFDC	| 74.56 | 56.59 | **77.01**
Huang et al. **[74]**	-	CNN	FF++	FF++, Celeb-DF, DFDC, DFD,	| **99.32** | 83.80 | 81.23 | 93.92
Mundra et al. **[75]**	-	PCA	StyleGAN1-3	StyleGAN1-3	
Pang et al. **[76]**	MRE-NET	ResNet34	FF++, Celeb-DF, DFDC,	FF++, Celeb-DF, DFDC, WildDeepfake	94.68 | 98.06 86.59 | **97.35** | **99.75** 85.61 | 91.23
Dong et al. **[77]**	ID-Unaware	ResNet34	FF++	FF++, Celeb-DF, DFDC	| **99.70** | 91.15 | 71.49

manipulation history of an image to form a unique noise signature. Then, Contrastive Learning is used to compare the similarity of random patches extracted from the input image to learn anomalies based on their noise signature and in turn, infer if the image has been forged. Also observing that image post-processing techniques can result in the contamination of low-level features and, fundamentally, information loss, it is proposed to use more traditional data-centric techniques for feature enhancement such as sharpening [58]. Unfortunately, this leads only to minor performance gains. To restore low-level contaminated features to their original state, it is suggested [58] to employ an EfficientNet backbone using an adversarial learning strategy and discriminator, which is in contrast to the approach in [55]. Cross-dataset evaluation shows the value of this approach by outperforming other SOTA methods. Still, it is clear that further research in this area should be considered, particularly in the anti-forensic domain.

Research on learning temporal inconsistencies in deepfake video content is another area showing promising results. A ResNet with Contrastive Learning is used in [49] in which discriminative features are learnt through a multi-modal approach (video and audio) using two separate networks. Named 'Person of Interest', the models are trained on real video content using a ResNet50, delivering significant performance gains through the use of the combined feature embeddings, that are fused using a Multilayer Perceptron. As the study recognises a potential limitation that can result in reduced generalisation when only a single POI video is provided, the authors recommend that multiple videos of the target are used to ensure greater accuracy. Instead of focusing on a multi-modal approach, study [50] uses extracted video keyframes to overcome compression and image quality loss using a Deep Convolutional Transformer model, which utilises Convolutional Pooling and Re-Attention. A 17-layer CNN with a kernel size of 3×3, batch normalisation and

a GELU activation function is used to extract local features before being fed directly to a Pooling Transformer, using depth-wise separable convolution for learning global representations. To address the limitations of the previous method, study [71] uses a noise-based approach where the extraction of the I-Frame or intra-frame enables the extraction of a higher level of image quality post-compression. Furthermore, the authors utilise a Siamese network combined with a pre-trained Recursive Information Distillation Network (RIDNet) to extract noise patches from the face and background, using a Euclidean distance, ensuring the background region is the furthest away. In addition, a Multi-Head Relative-Interaction is designed as a replacement for the Cosine similarity used by Siamese networks to enhance the measurement of similarity between the noise patches, whilst overcoming limitations that the authors perceive as information loss and performance degradation. Overlapping face and background patches are observed for future research in the study as this could result in the model generating false positive classifications. A Multi-Rate Excitation Network is proposed in the study [76], where the spatial-temporal inconsistencies are learnt through bipartite groups to measure different sampling rates. The expectation is that sampling different rates can encourage the network to learn longer-distance temporal inconsistencies. A Momentary Inconsistency Excitation module is used to extract spatial artefacts and to force the network to learn short-distance cross-group temporal inconsistencies, while a Longstanding Inconsistency Excitation module focuses on long-term temporal relationships. The Evaluation highlights that spatial and temporal embeddings play an equal role in the effectiveness and robustness of a detection method.

The use of facial features to extract identity information is another research direction to achieve promising results. Unique facial characteristics acquired from the local and global feature space can prove valuable, especially in the area of identity leakage, towards improving generalisation. In the study [74], the authors devise a novel Implicit Identity Driven framework to measure the distance between the explicit and implicit identity of real and fake faces. This approach relies on an Explicit Identity Contrast (EIC) and Implicit Identity Exploration (IIE) loss using a CNN architecture as the backbone. The hypothesis is that, within the feature space, the fake face converges more closely with the implicit identity of the target face than the explicit identity of the source face during the fake swapping process. Therefore, the EIC loss can be used to separate samples within the feature space by creating discriminative feature embeddings, while the IIE loss helps to refine the implicit identity from the target face. Alternatively, the Spatial Interaction Network, proposed in [64], utilises a Region of Interest layer and a Recursive Feature Eliminator to generate a local feature map using coordinates from four facial regions (nose, mouth, left eye and right eye). It is suggested that the removal of the max pooling layer of an Xception network (backbone) keeps the loss of low-level features to a minimum. A Spatial-Aware Module learns the weighted importance of both the local and global features to compute the similarity between the features. A predicted score is then produced via a Mult-Layer Perceptron. Eventually, it is shown that leveraging local features from the global feature map can lead to a reduction in computational resources, whilst achieving competitive results. In contrast to the previous method, the authors [77] highlight that unintentional implicit identity information may be being learnt by binary classifiers, which could result in a model becoming biased and therefore increasing the risk of misclassification. The study evaluates several architectures pre-trained using FFHQ [79], and despite the model not being trained on several datasets, implicit identity leakage was discovered from the Celeb-DF [80] and LFW [81] datasets. To overcome this, the authors propose a multi-scale anchor to focus on local regions and in turn limit the model's exposure to global identity information. In [75], low-dimensional embeddings are extracted using Principal Components Analysis, Autoencoder and Fourier analysis to demonstrate how common features in the spatial domain can be used to distinguish between fake and real images. The expectation is that unique geometric attributes associated with the synthesis process can be used to exploit within-class similarities that relate to how the face is aligned, the pose and other relevant factors. Although alterations to the geometric aspects, such as cropping, could result in performance degradation, the authors believe their approach is less likely to be affected by anti-forensic or laundering attacks. Instead of focusing on features in the spatial domain, the approach taken in [54] uses Rationale-Augmented CNN to perform facial reconstruction for deepfake detection. The authors observe that substituting the cross-entropy loss for a triplet loss function of an Inception network would enable the model to quantify new facial features during the training phase. Experiments show the triplet loss has the potential to both improve the model's overall performance and provide computational efficiency when performing similarity matches between each face.

Another direction of research has focused on Graph Neural Network (GNN) as it is believed that they can learn a richer representation of features, particularly when it comes to interconnected facial landmarks, which could prove crucial in the absence of diverse deepfake datasets. In [57], the authors propose to combine a GNN and pyramid ResNet structure to enhance model interpretability. The input image is first spliced into patches, before being fed to the K-Nearest Neighbors algorithm to construct the graph. There have also been attempts to utilise individual pixels [82] for constructing the graph or integrating GNNs with the CapsNet and Long Short-Term Memory network, but these resulted in degraded performance due to model complexity [57]. Also motivated by the ability offered by GNN in terms of structuring vital information from complex facial features by constructing high-level relationships between inter-connecting nodes, it is proposed to improve the learning of long-term interdependencies of a Graph Convolutional Network (GCN)

by combining a pre-trained ViT network using Contrastive Learning [72]. In addition, since high-level features are supposed to be less prone to interference by the process of manipulation, they are key to providing greater generalisation to unseen data. The study observes the relationship between the graph convolution layers and the receptive field for learning long-term information. Performance degradation could result from a reduction in the number of layers. To address the limitations of the previous study, [59] a GCN architecture is enhanced by using a dynamic graph learning approach [59] since this enables the model to build the structure of each layer dynamically as the model continuously learns. Identifying the initial optimum number of neighbouring nodes for the construction of the graph is highlighted as a limiting factor in the model's design. In other words, the model could demonstrate signs of under or over-fitting of its data during the training phase. Despite this, the adaptability of the graph structure to change allows for a more flexible approach.

B. TRANSFORMER

The authors of the study [52] recognise a key weakness in traditional CNNs where limited coverage of global features has an impact on a model's ability to learn the image as a whole. The authors continue to explain how the convolutional filter applied in the pooling layer results in vital information being removed. To overcome the loss of this information, the authors apply a ViT architecture, explaining how embedded global information can be captured through the use of a Multi-Head Self Attention Layer using positional patch embeddings. To improve overall performance, the authors apply a hybrid approach by combining the EfficientNet architecture with a ViT to improve the representation of spatial information across the local and global feature space. In the study [56], low and high-level semantic information is used to evaluate how discriminative features provide separability amongst real and fake distributions, trained separately using a pre-trained Xception and ViT network. The outcome from the study concludes, using linear-probing, that the ViT presents a stronger representation of high-level features and has the potential for greater robust generalisation to unseen data. However, the authors observe that fine-tuning the complete backbone of parameters would result in efficiency challenges while identifying that limited deepfake datasets could have the potential to result in the model overfitting for downstream training. To overcome this limitation the study adopts a fine-tuning strategy to 19% or 16.92 million of the overall ViT-Base parameters while freezing the remaining backbone.

Access to datasets and expensive computational resources, observed in [56], were also highlighted as challenges in [61], in which the authors emphasise the importance of developing a lightweight model that is capable of operating under such conditions. In contrast to [56], the authors claim to have further reduced the number of trainable parameters of the ViT-Base from 86M to 5.2M or 6%. The evaluation of this novel approach, termed Shallow ViT, demonstrates competitive results against SOTA methods. However, only intra-dataset experiments were carried out and therefore it is not possible to determine how the model would generalise under cross-dataset conditions. Alternatively, a lightweight framework is achieved in the study [65], where the authors claim their model uses only 3% of the parameters normally associated with other SOTA methods. The authors fuse feature embeddings from spatial, temporal and spatiotemporal features as a holistic approach to deepfake detection, using a transformer-based architecture. Each respective embedding is processed through a pre-trained model using a combination of 2D and 3D convolutional layers with max pooling, which is then fed independently to a series of Transformer encoder layers using Multi-Head Self Attention and Mult-Layer Perceptron. The self-attention token embeddings are concatenated further using a sequence pooling technique before a binary classifier determines if the content is real or fake [65]. Despite concerns of spatial information being lost [60] through the max pooling layer, the authors [65], believe that their framework can lead to improved performance from the reduction of spatial information whilst maintaining competitive results to other SOTA methods.

Study [69], freezes the backbone parameters of a Swin Transformer during the training phase and subsequently fine-tunes the model for downstream classification. The authors hypothesise that during the creation of a deepfake, a loss of depth information can be used as a measurement to calculate the feature distance, whereby real faces will measure closer and real and fake faces will measure further away. Trained using the source, target and fake face, a depth map is created before being fed to a triplet loss network where the discriminative features are learnt concerning their latent feature space. As highlighted in the study, SOTA results are observed, however, further testing is required to demonstrate the effectiveness of the model under more challenging conditions.

On the other hand, study [63] presents a novel detection strategy named Thumbnail Layout (TALL) is combined with a Swin Transformer to enhance spatial and temporal awareness, with the aid of Self-Attention and a Shifted Window mechanism. The study suggests improved generalisation when TALL is used with a Transformer, however, due to the model-agnostic design, TALL can be adapted to work with other DL architectures. The TALL element employs a dense-sampling approach to extract four consecutive frames at random, which are then resized and presented as a thumbnail layout. A novel approach by decomposing the spatial-temporal features through a Transformer using self-attention and a self-subtracting mechanism is observed in the study [66]. The authors believe that during the creation process, spatial information is often treated in isolation and is therefore independent of the inter-frame sequence, leading to temporal inconsistencies being introduced. To capture these inconsistencies, the self-subtracting mechanism helps to guide the network to target important temporal features.

Based on the Layer-wise Relevance Propagation [83], the authors introduce explainability to their model by visualising the discriminative and salient areas.

The authors in the study [67], observe that poor generalisation can result from model training on datasets where the number of discriminative features is too subtle for a standard CNN architecture to interpret. However, to overcome this limitation, the authors propose a Frequency-Aware Attention Feature Fusion using an Xception network as the backbone. To improve the learning of local and global features, augmentation is applied to each RGB frame using a Discrete Cosine Transform and a series of learnable weights to capture a range of frequency domain information. The author subsequently converts the frequency information back to their spatial domain to produce a frequency-aware image. Visualising the model using a Gradient-weighted Class Activation Map (Grad-CAM) confirms that their approach can track regions with manipulation, whilst ignoring those that are in face pristine. In contrast, the study [68] proposes an audio-visual multi-model framework by fusing spatial and temporal feature embeddings to learn inter-frame joint relationships. The audio-visual embeddings are first encoded using a spatial and temporal encoder and are based on a ViT network, which the authors claim is better at capturing a stronger representation of temporal features over time. A modified decoder was employed to overcome the challenge of collaboratively learning from multi-modal distributions that are not the same, using a Bi-Directional Cross-Attention block.

C. GENERATIVE AI

Deep neural networks might be unable to discriminate subtle artefacts, particularly as generative models advance, which can result in redundant information being learnt. Based on this idea, an adversarial learning strategy to perform artefact disentanglement, using an encoder and decoder structure, is proposed to provide significant improvement during the feature extraction [70]. Here, the features, with the newly constructed fake image are re-processed through the encoder and decoder to learn the ground truth. The study highlights that adversarial learning can not only achieve stronger generalisation but also overcome some of the challenges of training based on specific datasets.

VI. DATASET

Determining the right type of data for any research topic is challenging, and AI is no exception. Data plays an essential part in the training and validation of any given model, irrespective of the task at hand. However, determining what type of data and how much is needed is something that requires careful consideration. Fortunately, access and availability of datasets for research in the field of deepfakes have become more straightforward thanks to the tremendous effort from both academia and industry, where rich and diverse data can be acquired with ease. Nonetheless, the ethical approach taken to acquire individual subjects and the consideration for their privacy should be factored into the thought process when determining a dataset's suitability. It can be seen from Table 3, which summarises the papers reviewed in Section V there is consistency regarding the datasets they used for model training and evaluation. This is reinforced by Figure 3 which highlights the most popular datasets and their release date. It is particularly remarkable that those published between 2019 and 2020 are still highly influential in benchmarking deepfake detectors whereas deepfake generation technology has progressed significantly since their release. While their popularity is associated with the need to demonstrate comparative results against other SOTA methods that use the same datasets, as the field evolves, the value placed on a dataset should decrease compared to modern and more advanced ones. Indeed, the most relevant benchmarking should reflect the current state of deepfake datasets to show true SOTA and prove true generalisation.

Finally, another consideration is understanding the makeup of a given dataset in terms of pre- and post-processing activities and data diversity (age, gender, ethnicity, etc.). Indeed, if it is not adequately considered, their usage is likely to result in an adverse effect on a model's ability to generalise. For example, Figure 4 illustrates the ratio between real and fake content, highlighting important variations between datasets. The following section provides an overview of the datasets commonly associated with the training and evaluation of deepfake detector methods. A comprehensive process of uncovering datasets from 2015 onwards can be seen in Figure 3. Datasets typically considered benchmarking datasets are reviewed in Sections VI-A to VI-D. Novel datasets are discussed in Sections VI-E to VI-H. Section VI-I explores the considerations of ethics, privacy, dataset diversity and the challenges of processing operations to datasets.

A. FACEFORENSICS++

The FaceForensics++ (FF++) [84] dataset was published in 2019 for training and benchmarking deepfake detectors. This is a revised version of the previously released FaceForensics (FF) [85] dataset from 2018. The accompanying paper [84] defines four techniques used to generate the content, which is based on two graphical techniques (Face2Face and FaceSwap) and two learning-based techniques (DeepFakes and NeuralTextures), containing over 1.8 million images from 1,000 video sources. To simulate real-world data, the content was subjected to post-processing using different compression rates and is denoted by High-Quality (HQ), Low-Quality (LQ) and the original raw format in its uncompressed state. At the time of its release, the average accuracy was 80.87% (calculated from the average accuracy of raw 95.50%, HQ 80.73% and LQ 66.38), showing that it was quite challenging for SOTA technology. However, the best method from Table 3 (based on the average of the accuracy scores for the evaluation dataset) has since achieved an accuracy of 94.51%. In addition, this does not

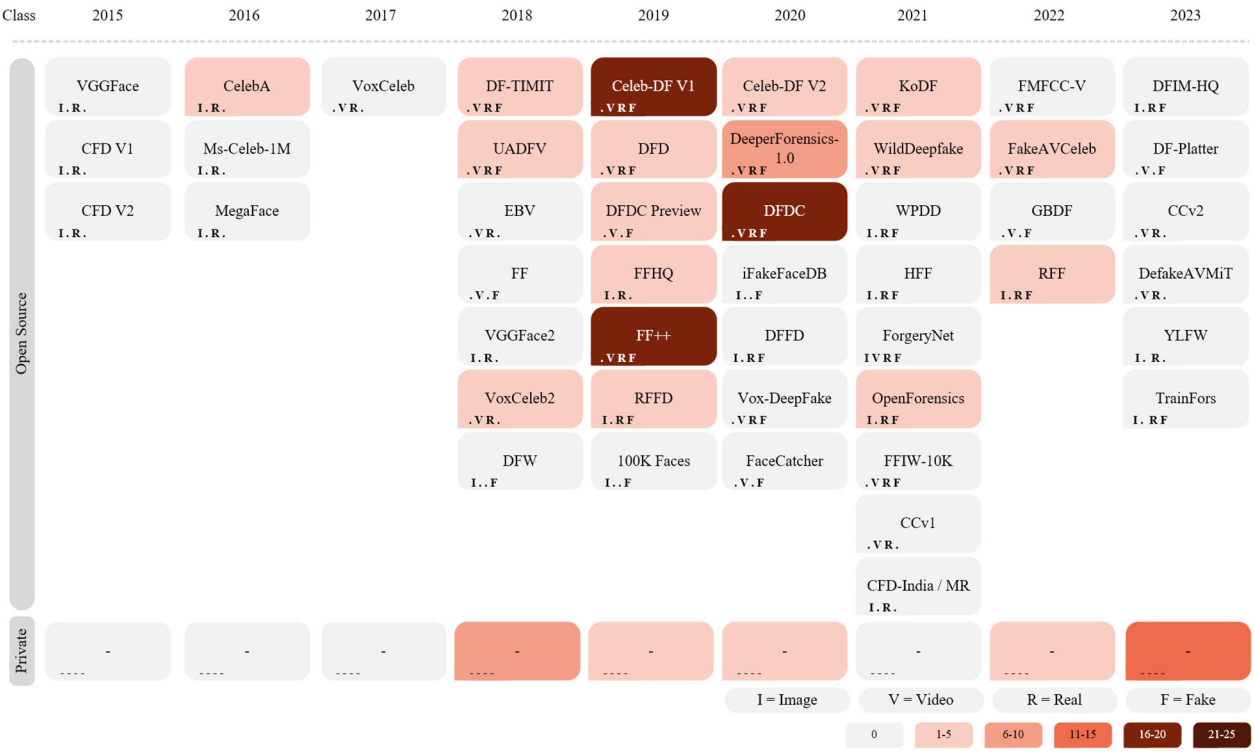

FIGURE 3. Timeline of datasets by class. The darker coloured boxes indicate the number of papers that reference the use of a particular dataset. The papers associated with this figure are based on the deepfake detection methods from Table 3. The specific details of the private datasets are not known at the time of writing this paper.

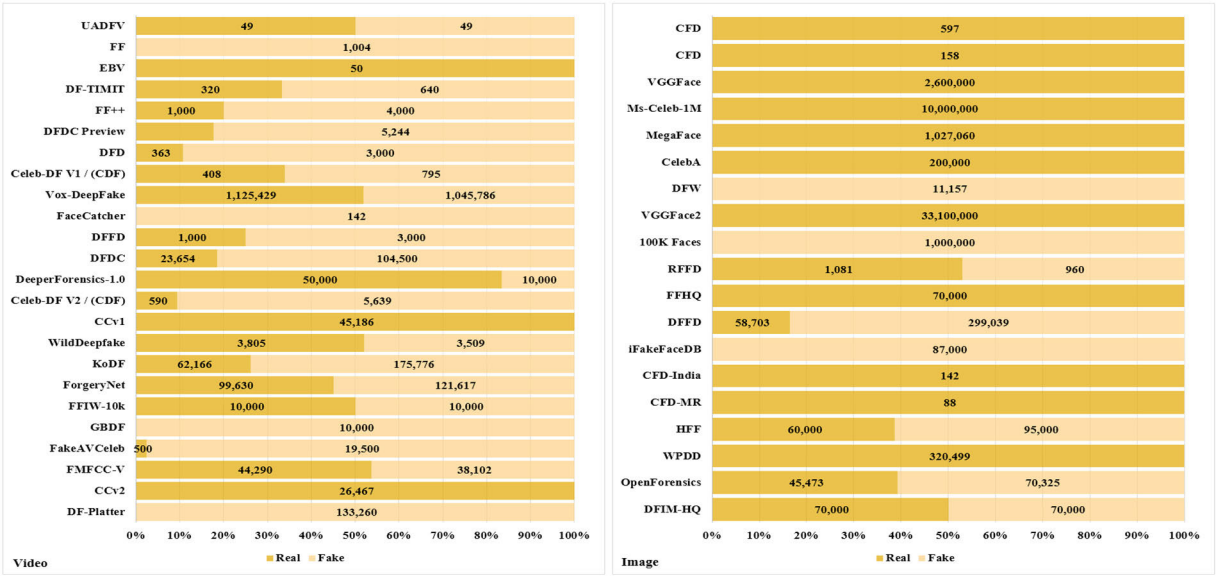

FIGURE 4. Timeline showing the ratio of real and fake content for each common dataset. Video datasets (left) and image datasets (right). The papers associated with this figure are based on the deepfake detection methods from Table 3.

take into account the breakdown of raw, HQ and Low. Furthermore, the generation of manipulated content was based on techniques that have since been superseded. In addition, the under-representation of gender could result in gender bias if used in the training and evaluation of a model. Thus, it can be argued that this dataset presents less of a challenge and may not be suitable as a primary benchmarking dataset to evaluate and detect current deepfakes. Despite this, this dataset provides a valuable reflection of the once-considered novel manipulation techniques.

B. DEEPERFORENSICS-1.0

Released in 2020, the DeeperForensics-1.0 (DF-1.0) [86] dataset contains a series of 60,000 high-quality video clips, split between 50,000 original and 10,000 manipulated videos that attempt to emulate real-world scenarios. It is comprised of 100 paid subjects with consent being granted for their face to be used in its original and manipulated form. In terms of gender split, there are 55% male and 45% female subjects. The age range is between 20 and 45, which the authors believe is consistent with the typical age range used in common deepfake videos. Other attributes, including skin colour and nationality, are also factored into the selection process. In addition to the dataset, a novel face-swapping framework called DeepFake Variational Audo-Encoder was developed as a learning-based approach to streamline the process of face-swapping. The authors highlight that a key factor in this dataset is not just the comprehensive size but also the consideration of environmental factors such as lighting, head pose, camera angle, etc. to represent rich and diverse real-world scenarios. The DeeperForensics-1.0 is not only a detailed and well-structured dataset but one that offers a wide range of identities based on age, gender and race. The high-quality footage of real actors recorded in a controlled environment provides valuable material for the research community. However, the limited subject age could result in poor generalisation in situations where the deepfake does not meet the common use case stated in the study. Additionally, work to enhance the range of deepfake swapping techniques would further complement this study in the future.

C. DEEPFAKE DETECTION CHALLENGE

The Deepfake Detection Challenge (DFDC) [87] dataset released in 2020 is another example of a large-scale dataset for testing deepfake detection methods. Published by Facebook AI, it is based on video footage from 960 paid subjects with prior consent as per the paper usage agreement. Preprocessing was applied to each of the video frames to crop and align the subject's faces before resizing to a resolution of 256×256 pixels. According to the accompanying paper, a total of 5,000 frames from the video footage were extracted and used to train the face-swapping models [87]. Multiple face-swapping techniques were incorporated including DFAE, MM/NN face swap, NTH, FSGAN and StyleGAN. Blending and noise reduction techniques were applied during the post-processing stage to ensure the video content was of a suitable quality. A subset, Deepfake Detection Challenge Preview [88], was published in 2019. An interesting aspect of this dataset was the inclusion of footage from an outdoor perspective, something not often seen in deepfake datasets. Additionally, unlike the DeeperForensics-1.0 [86] dataset, none of the footage was constrained to a single face-on camera. Again, and similar to DeeperForensics-1.0 [86], more advanced creation techniques, including Stable Diffusion and Dall-e, will likely impact generalisation. Furthermore, and based on the figures in Table 3, the highlighted evaluation accuracy for the dataset demonstrates an average score of 83%, which is in line with the results recorded in [89]. Thus, this dataset remains quite challenging, which is very valuable to assess the latest deepfake detectors.

D. CELEB-DF

The Celeb-DF (CDF) V2 [80] dataset released in 2020 is a revised version of the CDF V1 dataset and is comprised of 5,639 deepfake videos from 590 original video sources on the Internet. The subjects are 59 celebrities, with over 88% of those being Caucasians, followed by 6.8% for African Americans and 5.1% for Asians. Data augmentation is used in the pre-processing of the training data to overcome issues with colour mismatch, followed by post-processing techniques to remove unwanted noise using a Kalman smoothing algorithm. As reported in the original study, the evaluation results highlight reduced performance compared with other datasets tested, which the authors attribute to factors including improved visual quality and resolution. Consequently, the tested methods are unlikely to generalise well given that the dataset contains deepfakes generated using newer techniques. Subsequent advancements in the generation of deepfakes will likely result in this dataset being less effective as a benchmarking dataset.

E. WILDDEEPFAKE

The WildDeepfake [26] dataset is comprised of deepfake videos sourced from the Internet in 2021. furthermore, a rigorous process was undertaken to ensure that from the videos found, only those with an original unmanipulated version, were collated. In addition, the study employed three annotators to visibly inspect the videos and confirm if they believed the content to be real or fake. From the final 707 deepfake videos, a further 7,314 face sequences (3,805 real and 3,509 fake) were extracted. As reported in the study, preprocessing using facial landmarks was completed to ensure the faces were aligned. Evaluation results, as observed in the study, demonstrated a decrease in accuracy when the dataset was benchmarked against other datasets using a range of architectures and detection methods. The authors attribute this to the diverse range of settings used in the video sequences, which include interviews, broadcasting and films, and therefore believe this to be a challenging dataset, particularly when combined with other datasets. Interestingly, the study emphasises a great level of trust in the acquisition of these videos, particularly when there is little scope to validate the content source. More importantly, from the sourced videos, consent may not have been granted for their identities to be used and therefore, could result in a breach of user privacy while presenting ethical implications. Furthermore, there may be implications for using trademark material.

F. KOREAN DEEPFAKE DETECTION DATASET

The Korean DeepFake Detection Dataset (KoDF) [90] was established in 2021 to overcome the shortfall of Asian

subjects currently represented in deepfake datasets. Based on 403 subjects, the content boasts a collection of fake videos using six synthesised models and real video content. Additionally, the setting for the video footage is based on the subject talking directly to the camera, which is assumed as the at-most-risk setting for manipulation. Augmentation and other pre-processing techniques were omitted from this dataset to allow the researcher to apply their approach accordingly. However, post-processing was applied to the synthesised videos to sharpen the quality of the content and remove unwanted noise. As reported in the study, the results from the evaluation demonstrate comparative performance against the other datasets tested. However, the study observes a markable increase in performance when additional datasets such as DFDC [87] and FF++ [84] are combined during model training.

G. GENDER BALANCED DEEPFAKE DATASET

Proposed in 2022, the Gender Balanced Deepfake Dataset [91] is an amalgamation of 10,000 real and fake videos sourced from the FF++ [84], Celeb-DF [80] and DeeperForensics-1.0 [86] datasets respectively. Designed to promote gender fairness, the dataset was manually annotated to remove gender bias in model training and evaluation. In addition, the study has removed deepfake videos containing irregular gender swapping. The concluding results, as reported in the study, highlight that improvements in gender bias can be achieved and should be seen not only as an important milestone in gender fairness but also in reducing overall all bias across other attributes such as age and race.

H. DEEPFAKE IMAGE-HIGH-QUALITY

The Deepfake Image-High-Quality (DFIM-HQ) [92] became publicly available in 2023 to provide a challenging dataset for benchmarking detection methods. The dataset consists of various scenarios, poses, degradation and illumination, making it more closely aligned with the type of images found in the wild. Furthermore, the dataset is a combination of 70,000 images from the FFHQ [79] dataset and 70,000 StyleGAN2 images [93]. Consequently, the authors recognise the risk of introducing inherent bias transfer from FFHQ, which could result in a trained model becoming discriminant of certain attributes (age, race and gender). However, the authors propose adversarial debiasing as a technique to overcome this limitation, which is applied to the model training and not the dataset directly. To assess the effectiveness of adversarial debiasing, the authors utilised metrics from AI Fairness 360 [94], which demonstrated reduced bias. It should be noted that deepfake techniques used in DFIM-HQ are limited to those generated from StyleGAN2 and do not take into account the imagery from newer generations of StyleGAN such as StyleGAN3 [95]. Furthermore, other forms of deepfakes derived from face-swapping techniques are also omitted. In addition, the range of permutations by users who have applied image filters in the FFHQ dataset could result in poor generalisation as they do not provide a true reflection of real images.

I. DATASET CONSIDERATIONS

Managing datasets, both in terms of volume and structure, is complex and requires careful planning. A data management process can be used to not only streamline the data but also reduce the risk of data contamination or unexpected changes to the data itself. Each dataset should be analysed to understand how the data is distributed and avoid the risk of under or over-representation, which could result in an unbalanced and biased dataset.

1) ETHICS AND PRIVACY

How deepfake datasets are acquired, stored, and used is important from an ethical and privacy perspective. Unfortunately, the datasets [26] and [80], reviewed in Section VI, are examples of where material has been obtained from the Internet, and the associated papers contain little to no evidence of compliance concerning ethics and subject privacy. Indeed, each subject must not only consent to the usage of their identity but also understand how their identity will be used. Recent news has highlighted a growing trend were high-profile deepfakes are targeting individuals. For example, an audio deepfake of US President Jo Biden was distributed on social media, attempting to disrupt voters ahead of an upcoming election [48], [96]. Similarly, Taylor Swift was subject to a series of deepfake images that were shared online and without her consent [48], [97]. However, high-profile individuals are not the only targets. High school principal Eric Eiswert was the subject of a deepfake recording that was used to spread racial slur. Fortunately, police investigators were able to validate the recording as a deepfake [98]. Thus, a framework was defined to outline the best practices for the creation and usage of publicly available datasets to ensure best practices are followed in a standardised manner [99]. In addition, the side effects of training DL models with vast volumes of personal identities are becoming a topic of interest, with recent research exposing the risk of data leakage caused by a DL model's ability to memorise its training data [100]. Worryingly, using a Stable Diffusion [41] model to generate a new identity, they discovered that the generated individual matched closely one of the identities in the training dataset. Furthermore, it was reported in 2022 that a patient's medical images were found in the LAION-5B [101] dataset without the person's content [102]. This highlights the risk of exposing confidential training data due to adversarial attacks on trained models with reverse engineering techniques poses a serious risk.

In addition to the points raised above, the impact of model bias on real-world applications have far reaching consequences that need to be addressed. For example, model bias is highlighted in several key datasets, including Celeb-DF and DFDC, where trained models are unable to correctly classify images against certain demographic populations. This novel

research identifies that by addressing the imbalance of facial attributes in leading datasets, improved generalisation could also be achieved against unseen data. Furthermore, the annotated labels used in the creation of a deepfake dataset are often omitted from the public release, making it difficult to evaluate the fairness of the data and how it is represented across a global demographic [91]. The consequence of eroding trust in deepfake detection due to the inequality of data fairness, presents an important challenge moving forward. The severity of misclassification caused by systems put in place to safeguard society, has the potential to cause great harm, while allowing threat actors to continue taking advantage.

2) DATASET DIVERSITY

Acquiring adequate data to perform model training while maintaining a fair distribution of key attributes (age, gender, ethnicity, etc.) is a notable challenge. Using augmentation techniques in the spatial domain to increase dataset diversity has been seen as an effective approach in overcoming this limitation. However, the process of generating a fake sample for a given dataset is often dataset specific in their implementation, which can lead to inconsistencies in the distribution from one dataset to another. To counteract this, the authors suggest that combining within and cross-domain augmentation in the latent space to learn more intrinsic features can lead to improved generalisation. Alternatively, it has been proposed to composite multiple samples from real and fake samples. The authors claim that this technique can not only overcome forgery specific bias within the data but also close the gap between dataset distributions.

The Synthetic-20K Dataset attempts to overcome the challenge of demographically underrepresented groups of people using techniques to mirror the complex and intricate details associated with the human face. While this is a positive move forward, the value associated with this dataset is somewhat limited by the StyleGAN2 [93] architecture. Opportunely, the ever-improving Generative AI technology that has emerged over recent years offers substantial promises for the creation of realistic synthetic imagery. The many Diffusion Models (DMs) [44] already available to the public is a clear indicator to this high level of interest. However, the research community has been slow to respond, with little progress being made to address the detection capabilities against DMs. The authors identify that existing detection capabilities will unlikely generalise to content from DMs due to variations found in the higher-frequency space of the frequency domain. Despite this, the presence of some degree of similarity in GANs, suggests that models trained with DM data may lead to improved generalisation. Already usage of various DMs has contributed to overcome the limitation of access to domain specific data with the creation of the DiffusionDB-Face and JourneyDB-Face datasets. The novelty of these datasets includes the comprehensive use of Text-to-Image prompts to generate a rich and diverse collection of imagery. Yet this may lead to additional pre-processing to remove unrealistic imagery. In any case, the presence of model bias through data diversity has the potential to erode trust. As technology continues to push boundaries in quality and realism, how we perceive and differentiate between media in the future will likely be influence our cognitive bias.

3) PRE AND POST-PROCESSING

Providing adequate documentation of all pre- and post-processing techniques used in the creation of a dataset's pipeline is crucial. Understanding the processes used in the creation can avoid unnecessary operations that could impact model performance. Furthermore, improved transparency can help to promote the reproducibility of results, which can lead to more accurate model performance. Le et al. [103] address the need for guidance to inform the researcher not only on how the data was prepared (resizing the input vs. cropping) but also on how this may impact the input pipeline of a given deepfake detector. There are many scenarios where pre- and post-processing are used to improve the quality of the data. For example, the implementation of pre-processing operations through data enhancement to remove unwanted information during feature extraction [70]. However, applying these techniques to data that may have already been subject to these operations could have the effect of removing useful features.

VII. OBSERVATIONS AND FUTURE TRENDS

This section provides reflections following the Challenge themes previously defined, i.e., dataset, architecture and scalability, explainability and evaluation.

A. DATASET

Access to rich and diverse datasets is still considered a challenge, as reported in the papers covered in Section V. This is believed to be a contributing factor as to why so many detection methods achieve poor performance when generalising against data from the wild [62]. Indeed, there is a lack of data representing complex and noisy environments as the background or situational setting of current dataset is generally limited and lacks realistic real-world composure [69]. Pre-processing techniques to enhance image quality before training have been extensively used in this field to encourage improved feature learning and model generalisation. For example, Gaussian blur noise reduction is used to enhance and refine training images to help improve overall generalisation [55]. Alternatively, augmentation techniques to present the data in various composures, including cropping and rotation, can supplement a dataset when the data lacks diversity. Indeed, two augmentation strategies have been applied to an input image using a contrast loss to improve the feature representation [67]. In addition, new samples can be created through a generator using the disentangled artefacts. An Artefacts Cycle Consistency Loss approach is proposed to help reduce the dependency on large datasets and improve overall performance [70]. New and improved architectures are presenting ways to not only optimise a model but also reduce its complexity. Rational augmented CNN has also been seen as

a possible solution to the challenge of limited data, where a model could not only be used for detection but also for the creation of new facial identities and to extend training and evaluation data [54]. However, there is a risk of the image quality degrading from imperfections in the source training data. Many of the SOTA methods proposed in academic literature focus on model performance through evaluating against large-scale datasets. However, limited coverage is given to the training data and performance of the model itself, which can lead to challenges of experiment repeatability. To overcome this limitation, the TrainFors dataset was curated as a benchmarking dataset for model training and evaluation.

Thus, instead of creating additional training data, it has been proposed to mitigate that need by exploiting adversarial learning to reconstruct low-level features to their former state [58]. An alternative approach is to develop lightweight models which can not only achieve stronger generalisation compared with other SOTA models but can do so with less data [75]. Also promising is usage of a rich representation of learned features using a GNN architecture as this can overcome limitations when access to diverse datasets is not available [57]. Finally, the authors in [60] believe their E-Cap Net can generalise to variations in data permutation without requiring additional training data as the model can represent the input image in its entire state and the relationship between each interconnecting part.

B. ARCHITECTURE AND SCALABILITY

It has been highlighted that in some instances, too much emphasis is placed on forgery localisation over the direct task of detection [62]. In addition, tailoring the detection method around the architecture for image segmentation may contribute to some SOTA methods experiencing poor generalisation and limited robustness. Although improved generalisation can be achieved through the ViT architecture due to its ability to learn high-level features, this poses significant challenges from an increase in model complexity and the requirement for large training datasets [56]. Despite this, the authors conclude that, by applying a fine-tuning strategy while freezing the backbone of the network, significant improvements to the model's efficiency can be achieved. Also exploiting a pre-trained network, a lightweight Transformer architecture is designed, where the parameters are frozen while using only 3% of the parameters typically used in SOTA methods [65]. An interesting trend is the usage of architecture-agnostic approaches, where novel detection methods are not constrained to any specific type of image classification model [63] and [64]. In other words, this will help to future-proof a model to adapt to new generations of DL architectures. As convergence speed is a crucial element in model deployment ability, it is proposed to deliver a sparse gradient and compact feature representation by using a Max-Feature-Map (MFM) activation function [60]. Furthermore, improved model efficiency using MFM can also be achieved for other common activation functions, including ReLU and Sigmoid. Using a Prototype Learning Layer to cluster faces based on similarity matching presents an interesting approach. The model's effectiveness at discriminating real and fake content without having been trained on fake content could lead to improve generalisation as the model itself is not forgery domain specific.

Finally, although advancements in generative AI and the increasing range of available Diffusion Models allow the production of realistic imagery, which is on par with StyleGAN2 [93] and StyleGAN3 [95], only a limited coverage of DMs are observed in the papers reviewed. Despite this, the range of generative AI models capable of creating multi-modal content has surged in recent years. The level of realism and sophistication in generating content using Text-to-Image, Text-to-Audio, etc. requires a shift towards a multi-modal approach to detection. One novel approach utilises contrastive learning to capture features from the audio-visual space using a cross-model technique. Real video content is fed to a model trainer for downstream learning, where masked embeddings are learnt from alternate modalities. A second model is subsequently trained on the learnt embeddings as a classifier to determine between real and fake videos. Despite improved generalisation, the technique only works with single person videos. Alternatively, it has been proposed to combine a multi-modal transformer approach to capture spatial and temporal inconsistencies across the audio-visual space. Furthermore, by using dynamic weight fusion, not only is the loss of vital information during the training phase reduced, but also common features are extracted. Evaluation of the DFDC dataset demonstrates significant improvement in performance using this multi-modal approach. It is expected that usage of the underlying architectures from systems such as Dall-E 3 [39], Imagen [40] and Stable Diffusion [41], will open new opportunities in research for the detection of deepfakes.

C. EXPLAINABILITY

Earlier research into deepfake detection using ML often focused on solving this challenge as a binary classification problem, for example, labelling the input source as either real or fake. However, the quality and sophistication of deepfakes that are seen in the wild pose a far greater challenge than before. Thus, the ability to understand how a model is reasoning with its input becomes more and more important. This is core to TruFor, where an integrity score is supplemented by an anomaly and confidence map to provide the user with enough evidence to make an informed decision [62]. Similarly, it is proposed to incorporate a graph Transformer relevancy map using the output activation map [72]. There, the relevancy map is designed to provide improved transparency on the subtle details that the model believes are manipulated while ignoring irrelevant information. Based on this concept of output activation map, a Grad-CAM is used as part of model evaluation as it can provide a useful visual aid to highlight the regions within the source image that the model has deemed important [67].

TABLE 4. Summary of concluding challenge themes.

Challenge Theme	Description
Dataset	**Do publicly available datasets provide a fair representation of diversity (age, gender, ethnicity, etc.) in order to accurately train and evaluate models for deepfake detection?** Access to a wide range of diverse datasets is still considered a fundamental challenge for the research community. In particular, under-represented image and video content, specifically associated with age, gender and ethnicity, are recognised as key contributors to poor model generalisation. To address this challenge, techniques including adversarial debiasing have been used to interpret model bias using adversarial training. Understanding the sensitivity of the model towards certain attributes will enable researchers to balance the distribution of data and mitigate against attribute discrimination. The composure, lighting and location setting are instances of other limitations recognised in existing datasets. However, the DFDC dataset goes to some extent to address these limitations. Extending data using augmentation or virtual reality techniques in the pre-processing stage of model training is one solution used by several of the publications covered in this study. For the most part, this helps to enrich the dataset by introducing modified versions of existing data, yet this will not overcome the limitation whereby under-represented data is absent. For this reason, the process of generating new DL synthetic content using DMs like Stable Diffusion and Dall-E 3 has been proposed. Indeed, as the generation becomes evermore realistic, this could perhaps provide an adequate method for bridging the gap of under-represented data. However, the presence of statistical information usually embedded within real media will be non-existent, resulting in a different attributed noise signature. To conclude, while recognising the challenges and proposed mitigations above, it can be observed that publicly available datasets do not adequately provide a fair representation of diversity. Future collaboration between academia and the research community could present opportunities to modernise datasets using combined resources. **How does the practice of focusing on benchmarking algorithms against commonly used (or standard), but often 'old', datasets, instead of exploring new and innovative datasets, affect research progress?** In recent years, datasets including FF++, Celeb-DF and DFDC have become recognised as the leaders when it comes to benchmarking performance. This is partly due to their size and complexity, which far exceeds lesser-known datasets. However, it is evident from the papers reviewed that many of the methods covered already match, if not exceed, the original performance stated by the authors of the dataset. Yet, these datasets still represent the gold standard for model evaluation in what appears as a competitive process of out-doing each other in a gamified fashion. Moreover, improved performance is often associated with minor incremental increases. These datasets still and will continue to play an important role in benchmarking, however, recognising new and innovative datasets based on improved deepfake techniques must contribute to the evaluation process. **Are suitable measures put in place to handle pre-processing activities in terms of the ethical approach to collecting, processing and storing data?** The majority of the papers reviewed in this study recognise the importance of data collection and management as a pre-processing activity. Yet, no specific framework was observed. Some of the datasets identified in Section 6 require a valid academic account and justification before access will be granted. However, once the dataset is in the possession of the researcher, control is thereon governed by the individual. Limited information to describe how the dataset was collated could present ethical and privacy issues for the identities stored in the data. Providing details on how the identities were sourced would help to mitigate these issues. Furthermore, a breakdown of the pre- and post-processing activities would not only help reduce unnecessary processing but also help the researcher to make an informed evaluation once training and evaluation are complete.
Architecture and Scalability	**Are known weaknesses or limitations sufficiently recognised? Are these addressed using suitable mitigations?** Many of the papers reviewed in this study do recognise architecture weaknesses or limitation as an important aspect to designing an effective deepfake detector. Combining multiple architectures to form a hybrid network is observed as a suitable approach to not only overcome known limitations but to also build on the strengths of different network architectures. The popularity of this technique is evident in Figure 2, with a dramatic increase in research directed towards adapting the CapsNet and ViT architectures. **With research shifting towards Deep Learning, are hybrid architectures able to overcome known design limitations while balancing the trade-off between architecture, computational resources and datasets?** Managing model complexity and efficiency is considered a critical challenge in DL. Fusing one or more architectures to form a hybrid variant is recognised as one approach to overcoming known limitations. Observed throughout this study, hybrid architectures have continued to demonstrate SOTA performance. In addition, model optimisation and resource efficiency are observed in DL architectures where techniques to fine-tune and reduce the number of trainable parameters are considered. For example, increasing the model depth of the ViT architecture may result in negative performance due to the vast number of parameters. Increased research is highlighted for hybrid variants of the CapsNet and ViT architecture.

TABLE 4. *(Continued.)* **Summary of concluding challenge themes.**

	How can inference speed be expressed in relation to model complexity and computational resources? Successful deployability of a deepfake detector in any real-world setting will be influenced by the model's inference speed. Failure to achieve satisfactory performance in the wild due to computational demand will ultimately impact the process of adoption. Continued research into deeper networks will result in greater model complexity and subsequently result in additional computational overheads. However, model complexity and computational resources are recognised by many of the methods reviewed. Architecture agnostic approaches are considered as a way of reducing the dependency on specific architectures during the design, therefore, enabling the model to adapt to improved architecture designs in the future. Collaboration with government departments, for example, the Defense Advanced Research Projects Agency (DARPA), could provide researchers with the opportunity to evaluate algorithm performance. DARPA is already committed to a joint research programme with universities across the globe. Further extending this could provide a valuable testbed for research in the future.
Explainability	**What efforts have researchers made to offer better model interpretability?** Model interpretability is recognised as an important factor during the training and evaluation phase to understand how effective the model is at generalising to its datasets. Applying adversarial debiasing is one approach to understanding how the model will likely generalise to its training data and therefore, mitigating against model bias. Alternatively, implementing XAI into the network design using techniques, including LIME and SHAP, can provide greater model transparency and interpretability. Visualising the layers within the network using Grad-CAM provides a way of understanding the regions of the data that the network is focusing greater attention towards. Providing additional context on the decision process of a model is observed as a way of guiding the user in their decision process. **Can the use of a confidence score or visual aid to highlight possible areas of manipulation help provide trust and explainability?** The black-box nature of DL makes it challenging to interpret how a model came to its decision. A binary classification score of true or false is one tried and tested approach. However, this offers little context about a model's inference ability. Incorporating a Grad-CAM into the design is one method proposed by several of the reviewed papers. The Grad-CAM is used to visually inspect the image data through the network layer and provides a way of highlighting the regions which the model considers most important. Combining a confidence score with visual aids would provide greater context for the user to make an informed decision.
Evaluation	**What performance metrics are considered important for evaluating a model's accuracy?** Accuracy and AUC are widely observed as the primary metrics for model evaluation. However, limited usage of metrics, including precision and recall, could result in failure to identify poor generalisation during model training. **How suitable is current consideration for conducting both intra and inter-dataset evaluations and the importance of demonstrating a model's ability to generalise to unseen data?** Measuring model generalisation is acknowledged as a primary indicator for benchmarking model performance. Furthermore, conducting intra and inter-dataset evaluations is recognised as an instrumental component in this process. From detailed analysis, the studies can demonstrate model effectiveness against unseen data while providing valuable evidence to support future research. However, many of the studies offer little to no context on the justification for dataset selection. In many cases, intra-dataset evaluation is limited to those associated with benchmarking datasets. Recognising the strengths and weaknesses of a dataset while ensuring data diversity is imperative to measuring true generalisation. **Since the evaluation of the model's accuracy is artificial by nature, what consideration has been made for testing against data in the wild?** Representation of data from the wild is acknowledged by the research community as a critical challenge, for now and in the future. Newer generation datasets, including DFIM-HQ and WildDeepfake, have to some extent addressed this limitation with promising results. However, a strategic approach is required to ensure that future evolutions of deepfake algorithms are reflected in newer datasets moving forward. Government collaboration with the research community could present opportunities for the evaluation of deepfake methods in a secure and controlled environment. DARPA is one example where joint research from academic institutes across the globe has facilitated the testing of algorithms using datasets that would otherwise be unavailable to the general public.

The architecture and training datasets need to be factored into the model explainability, as this will influence performance. In [104], the authors establish a list of factors that influence a model's decision by measuring the

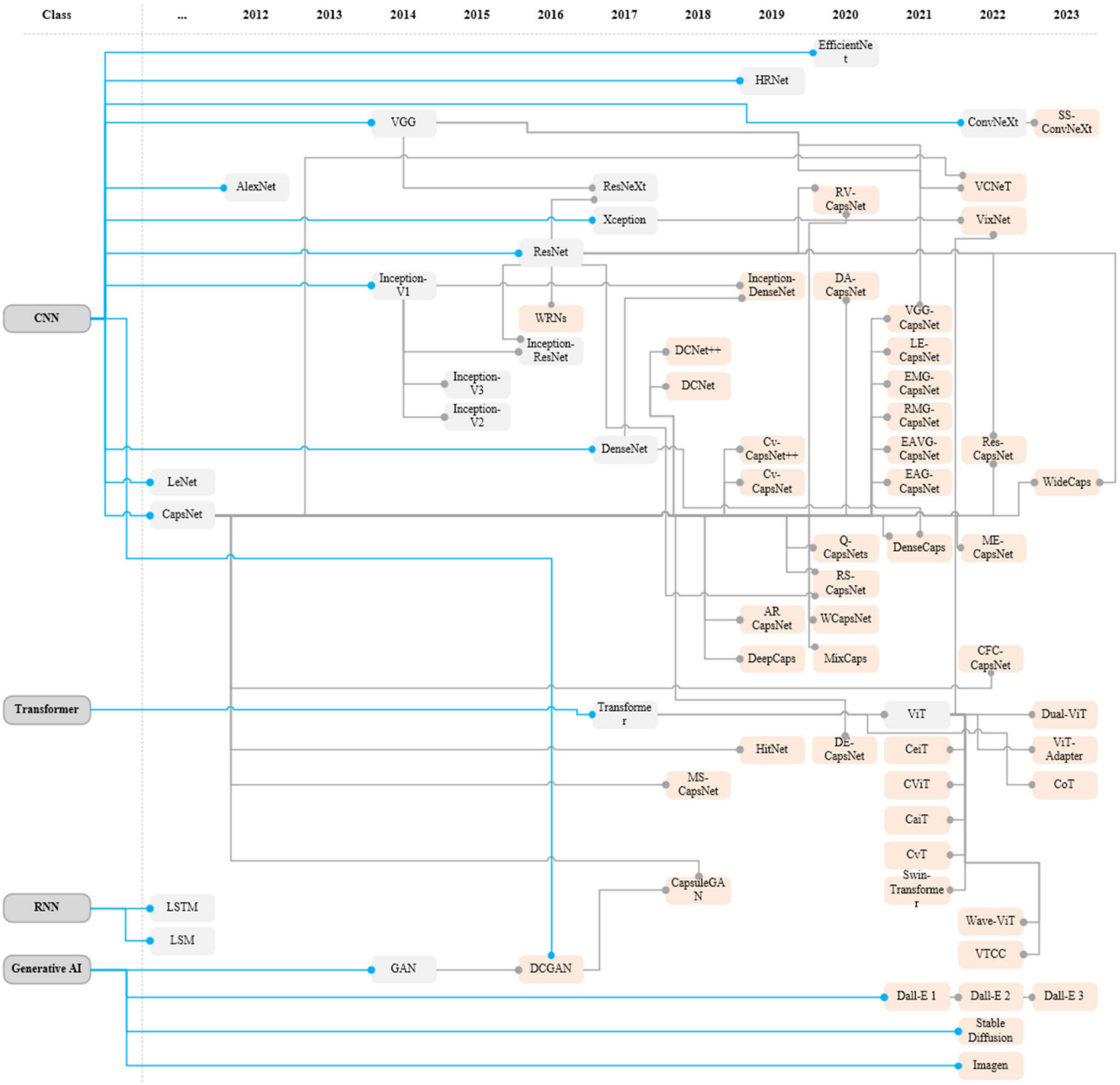

FIGURE 5. Timeline of architectures by class. the diagram highlights some of the main dl architectures and their associated variant architectures over time.

attributes of gender, race and affective (smiling vs. non-smiling) from samples obtained by the StyleGAN2 [93] algorithm and the FFHQ [79] dataset. Tested on various architectures, it was observed that the ViT represented a stronger bias towards females with lighter skin tones using data from the StyleGAN2 dataset, compared with males with darker skin tones from the FFHQ dataset [104]. As reported by [105], rapid growth in the field of Explainable Artificial Intelligence (XAI) has led to a surge in research dedicated to improving AI transparency. Simply put, the need to address concerns over accountability and trustworthiness while ensuring ethical considerations are accounted for will become a necessity in deployable AI in the future. Local Interpretable Model-Agnostic Explanations and Shapley Additive Explanation [106] are two examples of existing techniques designed to provide model interpretability. Incorporating XAI into the design of deepfake detectors would not only enhance model transparency but also aid the process of fine-tuning and optimisation.

D. EVALUATION

An important study describes the necessity of using evaluation metrics to not only understand how a model is making

predictions but also to observe if under- or over-fitting is occurring based on the training data used [51]. Indeed, there are a number of metrics available for DL that can help identify issues with the model or can be used in further fine-tuning. Although the widely accepted metrics include accuracy, AUC, precision, recall and F1-Score, the review in Section V highlights that the majority of the studies only evaluate their models using accuracy and AUC. Furthermore, fewer than twenty percent consider precision, recall and F1-Score as part of their evaluation. Still, accuracy provides a valuable measurement for calculating the number of true positive and true negative matches, which is important for determining how well the model is classifying the given dataset [60]. In addition, AUC can be used to determine how well the model is able to discriminate between the classes that it was trained with. Less common metrics may also be considered. They include a Probability of Detection [49], True Detection [67] and Equal Error Rate Metric (EERM) [56], [59], [60], [74]. EERM is of particular interest as it measures the rate at which the model is likely to misclassify [60]. Using a threshold value, the optimised position is when the False Acceptance Rate and False Rejection Rate are equal. Finally, although over seventy-percent of the papers apply evaluation to both inter and intra-dataset comparison, some studies limit their evaluation to inter-dataset only [54]. One should also note that a multi-modal comparison is performed in one study [49].

VIII. CONCLUSION

This paper presents a comprehensive review of literature relating to the field of deepfake detection. The aim is to provide the reader with the latest research on the architectures, detection methods and datasets currently used in the field while recognising and analysing their strengths and weaknesses. To conclude, Table 4 presents an overview of the original challenge themes defined in Table 1.

APPENDIX

See Figure 5.

REFERENCES

[1] I. Sample. (Jan. 13, 2020). *What Are Deepfakes and How Can You Spot Them*. The Guardian. Accessed: Sep. 22, 2022. [Online]. Available: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

[2] *Facelab*. Accessed: Oct. 18, 2022. [Online]. Available: https://facelab.mobi/

[3] *FaceApp: Face Editor*. Accessed: Oct. 18, 2022. [Online]. Available: https://faceapp.com/

[4] M. Boháček and H. Farid, "Protecting president Zelenskyy against deep fakes," 2022, *arXiv:2206.12043*.

[5] *Publications—Dimensions*. Accessed: Nov. 10, 2023. [Online]. Available: https://app.dimensions.ai/discover/publication

[6] J. F. O'Brien and H. Farid, "Exposing photo manipulation with inconsistent reflections," *ACM Trans. Graph.*, vol. 31, no. 1, pp. 1–11, Feb. 2012, doi: 10.1145/2077341.2077345.

[7] M. K. Johnson and H. Farid, "Exposing digital forgeries in complex lighting environments," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 450–461, Sep. 2007, doi: 10.1109/TIFS.2007.903848.

[8] W. Wu, W. Zhou, W. Zhang, H. Fang, and N. Yu, "Capturing the lighting inconsistency for Deepfake detection," in *Artificial Intelligence and Security* (Lecture Notes in Computer Science), X. Sun, X. Zhang, Z. Xia, and E. Bertino, Eds., Cham, Switzerland: Springer, 2022, pp. 637–647, doi: 10.1007/978-3-031-06788-4_52.

[9] C. Zhu, B. Zhang, Q. Yin, C. Yin, and W. Lu, "Deepfake detection via inter-frame inconsistency recomposition and enhancement," *Pattern Recognit.*, vol. 147, Mar. 2024, Art. no. 110077, doi: 10.1016/j.patcog.2023.110077.

[10] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4980–4989, doi: 10.1109/ICCV.2017.532.

[11] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," 2019, *arXiv:1912.13458*.

[12] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," 2018, *arXiv:1808.08396*.

[13] M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/computers12050091.

[14] A. M. Almars, "Deepfakes detection techniques using deep learning: A survey," *J. Comput. Commun.*, vol. 9, no. 5, pp. 20–35, May 2021, doi: 10.4236/jcc.2021.95003.

[15] S. Tyagi and D. Yadav, "A detailed analysis of image and video forgery detection techniques," *Vis. Comput.*, vol. 39, no. 3, pp. 813–833, Mar. 2023, doi: 10.1007/s00371-021-02347-4.

[16] K. Patil, S. Kale, J. Dhokey, and A. Gulhane, "Deepfake detection using biological features: A survey," 2023, *arXiv:2301.05819*.

[17] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023, doi: 10.1007/s10489-022-03766-z.

[18] M. Zanardelli, F. Guerrini, R. Leonardi, and N. Adami, "Image forgery detection: A survey of recent deep-learning approaches," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 17521–17566, May 2023, doi: 10.1007/s11042-022-13797-w.

[19] L. Stroebel, M. Llewellyn, T. Hartley, T. S. Ip, and M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," *J. Cyber Secur. Technol.*, vol. 7, no. 2, pp. 83–113, Apr. 2023, doi: 10.1080/23742917.2023.2192888.

[20] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Comput. Vis. Image Understand.*, vol. 223, Oct. 2022, Art. no. 103525, doi: 10.1016/j.cviu.2022.103525.

[21] J. W. Seow, M. K. Lim, R. C. W. Phan, and J. K. Liu, "A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, Nov. 2022, doi: 10.1016/j.neucom.2022.09.135.

[22] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: Generation, detection, and applications," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 3, pp. 219–289, Sep. 2022, doi: 10.1007/s13735-022-00241-w.

[23] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious DeepFakes: Survey, battleground, and horizon," *Int. J. Comput. Vis.*, vol. 130, no. 7, pp. 1678–1734, Jul. 2022, doi: 10.1007/s11263-022-01606-8.

[24] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[25] A. Naitali, M. Ridouani, F. Salahdine, and N. Kaabouch, "Deepfake attacks: Generation, detection, datasets, challenges, and research directions," *Computers*, vol. 12, no. 10, p. 216, Oct. 2023, doi: 10.3390/computers12100216.

[26] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," 2021, *arXiv:2101.01456*.

[27] F. Hong, Z. Chen, Y. Lan, L. Pan, and Z. Liu, "EVA3D: Compositional 3D human generation from 2D image collections," 2022, *arXiv:2210.04888*.

[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Ha, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[30] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Artificial Neural Networks and Machine Learning—ICANN 2011* (Lecture Notes in Computer Science), T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds., Berlin, Germany: Springer, 2011, pp. 44–51, doi: 10.1007/978-3-642-21735-7_6.
[31] M. K. Patrick, A. F. Adekoya, A. A. Mighty, and B. Y. Edward, "Capsule networks—A survey," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 1, pp. 1295–1310, Jan. 2022, doi: 10.1016/j.jksuci.2019.09.014.
[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, *arXiv:1409.4842*.
[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016, *arXiv:1610.02357*.
[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015, *arXiv:1512.00567*.
[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017. Accessed: Aug. 2, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html
[36] A. Vaswani, "Attention is all you need," 2017, *arXiv:1706.03762*.
[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[38] R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of generative AI applications," 2023, *arXiv:2306.02781*.
[39] J. Betker et al., "Improving image generation with better captions," *Comput. Sci.*, vol. 2, no. 3, p. 8, 2023. [Online]. Available: https://cdn.openai.com/papers/dall-e-3.pdf
[40] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022, *arXiv:2205.11487*.
[41] Stability AI. *Stable Diffusion 2.0 Release*. Accessed: Jan. 14, 2024. [Online]. Available: https://stability.ai/news/stable-diffusion-v2-release
[42] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Bus. Inf. Syst. Eng.*, vol. 66, no. 1, pp. 111–126, Sep. 2023, doi: 10.1007/s12599-023-00834-7.
[43] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2014, pp. 1–11. Accessed: Nov. 17, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html
[44] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.
[45] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020, *arXiv:2006.11239*.
[46] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," 2020, *arXiv:2006.09011*.
[47] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, *arXiv:2011.13456*.
[48] A. Birrer and N. Just, "What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape," *New Media Soc.*, May 2024, doi: 10.1177/14614448241253138.
[49] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest DeepFake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 943–952, doi: 10.1109/CVPRW59228.2023.00101.
[50] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 6, pp. 1–20, May 2023, doi: 10.1145/3588574.
[51] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," *Sci. Rep.*, vol. 13, no. 1, p. 7422, May 2023, doi: 10.1038/s41598-023-34629-3.
[52] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "DeepFake detection algorithm based on improved vision transformer," *Appl. Intell.*, vol. 53, no. 7, pp. 7512–7527, Apr. 2023, doi: 10.1007/s10489-022-03867-9.
[53] M. Soleimani, A. Nazari, and M. E. Moghaddam, "Deepfake detection of occluded images using a patch-based approach," *Multimedia Syst.*, vol. 29, no. 5, pp. 2669–2687, Oct. 2023, doi: 10.1007/s00530-023-01140-8.
[54] S. R. A. Ahmed and E. Sonuç, "Deepfake detection using rationale-augmented convolutional neural network," *Appl. Nanoscience*, vol. 13, no. 2, pp. 1485–1493, Feb. 2023, doi: 10.1007/s13204-021-02072-3.
[55] T. Lu, Y. Bao, and L. Li, "Deepfake video detection based on improved CapsNet and temporal–spatial features," *Comput., Mater. Continua*, vol. 75, no. 1, pp. 715–740, 2023, doi: 10.32604/cmc.2023.034963.
[56] R. Shao, T. Wu, L. Nie, and Z. Liu, "DeepFake-adapter: Dual-level adapter for DeepFake detection," 2023, *arXiv:2306.00863*.
[57] F. Khalid, A. Javed, Q.-U. Ain, H. Ilyas, and A. Irtaza, "DFGNN: An interpretable and generalized graph neural network for deepfakes detection," *Expert Syst. Appl.*, vol. 222, Jul. 2023, Art. no. 119843, doi: 10.1016/j.eswa.2023.119843.
[58] J. Ke and L. Wang, "DF-UDetector: An effective method towards robust deepfake detection via feature restoration," *Neural Netw.*, vol. 160, pp. 216–226, Mar. 2023, doi: 10.1016/j.neunet.2023.01.001.
[59] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7278–7287. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Dynamic_Graph_Learning_With_Content-Guided_Spatial-Frequency_Relation_Reasoning_for_Deepfake_CVPR_2023_paper.html
[60] H. Ilyas, A. Javed, K. M. Malik, and A. Irtaza, "E-cap net: An efficient-capsule network for shallow and deepfakes forgery detection," *Multimedia Syst.*, vol. 29, no. 4, pp. 2165–2180, Aug. 2023, doi: 10.1007/s00530-023-01092-z.
[61] S. Usmani, S. Kumar, and D. Sadhya, "Efficient deepfake detection using shallow vision transformer," *Multimedia Tools Appl.*, vol. 83, no. 4, pp. 12339–12362, Jun. 2023, doi: 10.1007/s11042-023-15910-z.
[62] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20606–20615, doi: 10.1109/CVPR52729.2023.01974.
[63] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "TALL: Thumbnail layout for deepfake video detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, vol. 27, Oct. 2023, pp. 22658–22668. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Xu_TALL_Thumbnail_Layout_for_Deepfake_Video_Detection_ICCV_2023_paper.html
[64] J. Wang, X. Du, Y. Cheng, Y. Sun, and J. Tang, "Si-Net: Spatial interaction network for deepfake detection," *Multimedia Syst.*, vol. 29, no. 5, pp. 3139–3150, Jul. 2023, doi: 10.1007/s00530-023-01114-w.
[65] M. A. Raza, K. M. Malik, and I. Ul Haq, "HolisticDFD: Infusing spatiotemporal transformer embeddings for deepfake detection," *Inf. Sci.*, vol. 645, Oct. 2023, Art. no. 119352, doi: 10.1016/j.ins.2023.119352.
[66] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1335–1348, 2023, doi: 10.1109/TIFS.2023.3239223.
[67] C. Tian, Z. Luo, G. Shi, and S. Li, "Frequency-aware attentional feature fusion for deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094654.
[68] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "AVoiD-DF: Audio-visual joint learning for detecting deepfake," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2015–2029, 2023, doi: 10.1109/TIFS.2023.3262148.
[69] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang, "Depth map guided triplet network for deepfake face detection," *Neural Netw.*, vol. 159, pp. 34–42, Feb. 2023, doi: 10.1016/j.neunet.2022.11.031.

[70] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1658–1670, Apr. 2023, doi: 10.1109/TCSVT.2022.3217950.

[71] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 12, Jun. 2023, pp. 14548–14556, doi: 10.1609/aaai.v37i12.26701.

[72] A. Khormali and J.-S. Yuan, "Self-supervised graph transformer for deepfake detection," 2023, *arXiv:2307.15019*.

[73] G. Yang, A. Wei, X. Fang, and J. Zhang, "FDS_2D: Rethinking magnitude-phase features for DeepFake detection," *Multimedia Syst.*, vol. 29, no. 4, pp. 2399–2413, Jun. 2023, doi: 10.1007/s00530-023-01118-6.

[74] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Huang_Implicit_Identity_Driven_Deepfake_Face_Swapping_Detection_CVPR_2023_paper.html

[75] S. Mundra, G. J. Aniano Porcile, S. Marvaniya, J. R. Verbus, and H. Farid, "Exposing GAN-generated profile photos from compact embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 884–892, doi: 10.1109/CVPRW59228.2023.00095.

[76] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "MRE-Net: multi-rate excitation network for deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3663–3676, Aug. 2023, doi: 10.1109/TCSVT.2023.3239607.

[77] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Dong_Implicit_Identity_Leakage_The_Stumbling_Block_to_Improving_Deepfake_Detection_CVPR_2023_paper.html

[78] Z. Liu, X. Qi, and P. Torr, "Global texture enhancement for fake face detection in the wild," 2020, *arXiv:2002.00133*.

[79] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2018, *arXiv:1812.04948*.

[80] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," 2019, *arXiv:1909.12962*.

[81] G. B. Huang, M. Ramesh, T. Berg, and E. L. Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep., Oct. 2007.

[82] M. Edwards and X. Xie, "Graph based convolutional neural network," 2016, *arXiv:1609.08965*.

[83] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791, doi: 10.1109/CVPR46437.2021.00084.

[84] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," 2019, *arXiv:1901.08971*.

[85] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," 2018, *arXiv:1803.09179*.

[86] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," 2020, *arXiv:2001.03024*.

[87] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[88] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[89] M. F. Sohan, M. Solaiman, and M. A. Hasan, "A survey on deepfake video detection datasets," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 32, no. 2, p. 1168, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp1168-1176.

[90] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "KoDF: A large-scale Korean DeepFake detection dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10724–10733, doi: 10.1109/ICCV48922.2021.01057.

[91] A. V. Nadimpalli and A. Rattani, "GBDF: Gender balanced DeepFake dataset towards fair DeepFake detection," 2022, *arXiv:2207.10246*.

[92] S. Mathews, S. Trivedi, A. House, S. Povolny, and C. Fralick, "An explainable deepfake detection framework on a novel unconstrained dataset," *Complex Intell. Syst.*, vol. 9, no. 4, pp. 4425–4437, Aug. 2023, doi: 10.1007/s40747-022-00956-7.

[93] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," 2019, *arXiv:1912.04958*.

[94] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Res. Develop.*, vol. 63, nos. 4–5, pp. 4:1–4:15, Jul./Sep. 2019, doi: 10.1147/JRD.2019.2942287.

[95] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," 2021, *arXiv:2106.12423*.

[96] (Jan. 22, 2024). *Fake Biden Robocall Tells Voters to Skip New Hampshire Primary Election*. BBC News. Accessed: Oct. 5, 2024. [Online]. Available: https://www.bbc.com/news/world-us-canada-68064247

[97] (Jan. 26, 2024). *Taylor Swift Deepfakes Spark Calls in Congress for New Legislation*. BBC News. Accessed: Oct. 5, 2024. [Online]. Available: https://www.bbc.com/news/technology-68110476

[98] (Apr. 26, 2024). *Baltimore High School Teacher Arrested Over Deepfake Racist Audio of Principal*. BBC News. Accessed: Oct. 5, 2024. [Online]. Available: https://www.bbc.com/news/world-us-canada-68907895

[99] K. Peng, A. Mathur, and A. Narayanan, "Mitigating dataset harms requires stewardship: Lessons from 1000 papers," 2021, *arXiv:2108.02922*.

[100] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," 2023, *arXiv:2301.13188*.

[101] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," 2022, *arXiv:2210.08402*.

[102] M. Growcoot. *Shocked Artist Finds Private Medical Photos in AI Training Data Set*. PetaPixel. Accessed: Nov. 20, 2023. [Online]. Available: https://petapixel.com/2022/09/26/shocked-artist-finds-private-medical-photos-in-ai-training-data-set/

[103] B. Le, S. Tariq, A. Abuadbba, K. Moore, and S. Woo, "Why do deepfake detectors fail?" 2023, *arXiv:2302.13156*.

[104] M. P. Gangan, A. Kadan, and L. V L, "Exploring fairness in pre-trained visual transformer based natural and GAN generated image detection systems and understanding the impact of image compression in fairness," 2023, *arXiv:2310.12076*.

[105] G. P. Reddy and Y. V. P. Kumar, "Explainable AI (XAI): Explained," in *Proc. IEEE Open Conf. Electr., Electron. Inf. Sci. (eStream)*, Apr. 2023, pp. 1–6, doi: 10.1109/eStream59056.2023.10134984.

[106] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

PETER EDWARDS received the B.Sc. degree (Hons.) in computer science from Brunel University, Uxbridge, U.K., in 2005, and the M.Sc. degree in advanced computing and digital forensics from Edinburgh Napier University, Edinburgh, U.K., in 2020. He is currently pursuing the Ph.D. degree with the School of Computer Science and Mathematics, Kingston University, Kingston, U.K.

His interests include deep learning, media forensics, and data analytics.

JEAN-CHRISTOPHE NEBEL (Senior Member, IEEE) received the M.Sc.(Eng.) degree in electronics and signal processing from the Institute of Chemistry and Industrial Physics, Lyon, France, in 1992, and the Ph.D. degree in parallel programming from the University of St Etienne, France, in 1997.

From 1997 to 2004, he was a Postdoctoral Research Associate/Fellow with the Computing Science Department, The University of Glasgow, U.K. Since 2004, he has been a permanent Academic with Kingston University London. Currently, he is a Full Professor of computer science with the Faculty of Engineering, Computing and the Environment, where he leads research in AI and machine learning applied to a variety of topics, including computer vision, bioinformatics, cybersecurity, renewable energy, and ecology.

Prof. Nebel was awarded the A. H. Reeve Premium by the Council of the Institute of Electrical and Electronics Engineers for a journal article describing his and the co-authors' pioneer work in developing a 3D Dynamic Whole Body Measurement System, in 2004.

DARREL GREENHILL received the Ph.D. degree in computer vision from Royal Holloway, University of London, in 1994.

He is currently an Associate Professor with the School of Computer Science and Mathematics, Kingston University. His research is in the areas of gamification, games education, and computer vision, resulting in more than 50 publications. He is the Course Leader of computer games programming B.Sc., game development (programming) M.Sc. and game development (design) M.A. His teaching interests include game programming, game engine programming, and artificial intelligence for game development.

XING LIANG (Member, IEEE) received the Ph.D. degree in mobile satellite communications from the University of Bradford, in 2007.

She is currently a Senior Lecturer with the School of Computer Science and Mathematics, Kingston University London. More than the years, she has actively contributed to multiple EU and U.K. research projects in AI and telecommunications, including H2020 ENSURESEC, H2020 EUNOMIA, EPSRC CONCORD, IST FIFTH, IST SatNEx, and Dunhill Medical Trust funded ATDA-BSL, and several industrial related research projects. Her current research interests include computer vision, deep learning, generative AI, and quantum machine learning with applications in cyber-physical system security, social media security, healthcare, the IoT, and business intelligence.