International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Exploring AI/ML Techniques for Deepfake Detection: A Comprehensive Review

Dr. Sofiya Mujawar[1], Harshal Madhukar Kumavat[2], Kuldip Sunil Kate[3], Atharva Prashant Thakare[4], Yash Sanjay Parihar[5]

[1]Asst Professor, [2,3,4,5]B. Tech Student
[1,2,3,4,5]SOET, DYPU, Ambi Pune

Abstract

The literature review on AI/ ML- grounded deepfake discovery delves into the evolving geography of ways designed to identify and alleviate the pitfalls posed by deepfake media. Deepfakes, which influence advanced AI to produce largely realistic fake vids and images, have raised significant enterprises regarding sequestration, security, and the integrity of digital content. The review totally categorizes discovery styles into deep literacy- grounded ways, classical machine learning approaches, statistical styles, and blockchain- grounded results. Deep literacy ways, particularly those employing Generative Adversarial Networks (GANs), have surfaced as the most effective in detecting deepfakes. These styles use expansive datasets and sophisticated neural network infrastructures to descry subtle inconsistencies and vestiges that are reflective of manipulation. Classical machine literacy styles, while generally less effective than deep literacy, remain important for point birth and original discovery stages. The review underscores the critical part of different datasets in training and assessing discovery algorithms. Datasets that encompass a wide array of deepfake exemplifications, including colorful manipulation types and quality situations, are pivotal for developing robust discovery systems. Performance criteria similar as delicacy, perfection, recall, and F1- score are employed to measure the effectiveness of these algorithms.

Keywords: Convolutional neural networks, DeepFake, Face2Face, fake face detection, fake face image forensics, multi-channel constrained convolution, transfer learning, video or image manipulation, digital media forensics.

1. Introduction

The notable advances in artificial neural network (ANN) grounded technologies play an essential part in tampering with multimedia content. For illustration, AI- enabled software tools like FaceApp, and FakeApp have been used for realistic- looking face switching in images and vids. This switching medium allows anyone to alter the frontal look, haircut, gender, age, and other particular attributes. The propagation of these fake vids causes numerous anxieties and has come notorious under the hood, Deepfake. The term Deepfake is deduced from Deep literacy(DL) and Fake, and it describes specific print-realistic videotape or image contents created with DL support. [1]. This word was named after an anonymous Reddit stoner in late 2017, who applied deep literacy styles for replacing a person. The associate editor coordinating the review of this handwriting and approving it for publication was Zahid Akhtar. face in pornographic vids using another person's face and created print-realistic fake vids.[3]. To induce similar fake vids, two neural networks (I) a generative network and(II) a discriminational network with a Face exchange fashion were used.[5]. The generative network creates fake images using an encoder and a decoder. The discriminational network denes the authenticity of the recently generated images. The combination of these two networks is called Generative Adversarial Networks(GANs), proposed by Ian Goodfellow. Grounded on a monthly report in Deepfake, DL experimenters made several affiliated improvements in generative modeling.[7]. For illustration, computer vision experimenters proposed a system known as Face2Face for facialre-enactment. This system transfers facial expressions from one person to a real digital icon in real- time. In 2017, experimenters from UC Berkeley presented Cycle GAN to transfigure images and vids into different styles. Another group of scholars from the University of Washington proposed a system to attend the lip movement in videotape with a speech from another source[6]. Eventually, in November 2017[1], the term Deepfake surfaced for participating porn vids, in which celebrities faces were shifted with the original bones

In January 2018, a Deepfake creation service was launched by colorful websites grounded on some private guarantors. After a month, several websites, including Gfycat, Pornhub, and Twitter, banned these services. still, considering the pitfalls and implicit pitfalls in sequestration vulnerabilities, the study of Deepfake surfaced superfast. Rossleretal introduced a vast videotape dataset to train the media forensic and Deepfake discovery tools called Face Forensic in March2018[3]. After a month, experimenters at Stanford University published a system, Deep videotape pictures that enables print-realistic re-animation of portrayal vids. UC Berkeley experimenters developed another approach for transferring a person's body movements to check the facial expressions and move

another person in the videotape [1]. NVIDIA introduced a style- grounded creator armature for GANs for synthetic image generation. According to report, Google hunt machine could and multiple web runners that contain Deepfake affiliated vids We set up the following fresh information from this report[4].

The top 10 pornographic platforms posted 1,790 Deepfake vids, without concerning pornhub.com, which has removed Deepfakes quests.

- Adult runners post 6,174 Deepfake vids with fake videotape content.
- 3 New platforms were devoted to distributing Deepfake pornography.
- In 2018, 902 papers were published in arXiv, including the keyword GAN either in titles or objectifications.
- 25 Papers published on this subject, including non- peer reviews, are delved , and DARPA funded 12 of them piecemeal from Deepfake pornography, there are numerous other vicious or illegal uses of Deepfake, similar as spreading misinformation, creating political insecurity, or colorful cybercrimes.

To address similar pitfalls, the eld of Deepfake discovery has attracted considerable attention from academics and experts during the last many times, performing in numerous Deepfake discovery ways. There are also some sweats on surveying named literature fastening on either discovery styles or performance analysis. still, a further com prehensive overview of this exploration area will be bene cial in serving the community of experimenters and interpreters by furnishing epitomized information about Deepfake in all aspects, including available datasets, which are noticeably missing in former checks[7]. Toward that end, we present a methodical literature review(SLR) on Deepfake discovery in this paper. We aim to describe and dissect common grounds and the diversity of approaches in current practices on

Deep fake discovery. Our benefactions are epitomized as follows. We perform a comprehensive check on being literature in the Deepfake sphere. We report current tools, ways, and datasets for Deepfake discovery- related exploration by posing some exploration questions.

- We introduce a taxonomy that classifies Deepfake discovery ways in four orders with an overview of different orders and affiliated features, which is new and the first of its kind.
- We conduct an in- depth analysis of the primary studies experimental substantiation. Also, we estimate the performance of colorful Deepfake discovery styles using different dimension criteria
- We punctuate a many compliances and deliver some guidelines on Deepfake discovery that might help unborn exploration and practices in this diapason.

2. Process of Literature Review

There are two corner literature checks proposed by Budgen et al. and Zlatko Stapic et al. in the field of software engineering. We borrow their approaches in our SLR and classify the review process into three main stages in order to identify, estimate, and understand colourfull inquiries related to particular exploration questions.

Planning the Review: The purposes of this stage are to(a) identify the need,(b) develop criteria and procedures, and(c) estimate the criteria and procedures related to this SLR.

Conducting the Review: Grounded on the guiding princi ples proposed in this stage includes six obligatory phases.

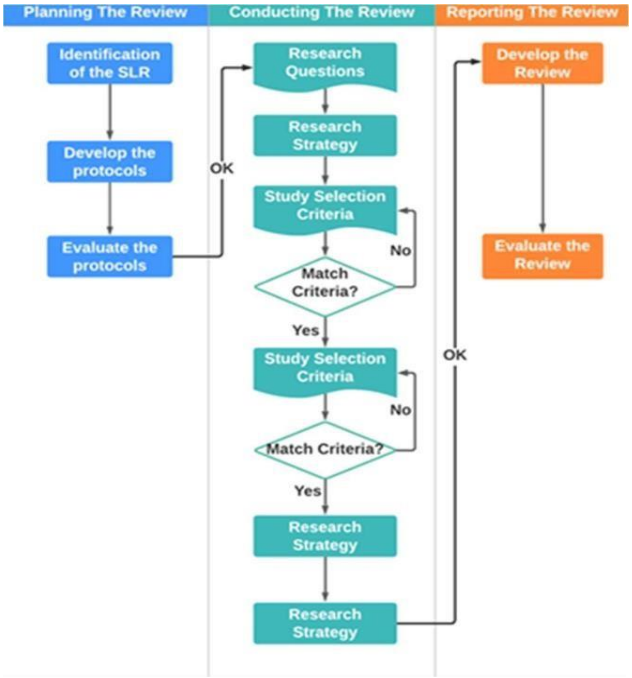


Figure: The process of the SLR

A. Research Questions(RQs): The purpose of the RQ phase is to identify applicable studies that need to be considered in the current review. We determine a set of RQs(described latterly) in the environment of the Deepfake sphere.

B. Search strategy(SS): A predefined hunt strategy aims to find as numerous as primary studies related to our exploration questions. We try to establish an unprejudiced hunt strategy to descry as much of the applicable literature as possible.
C. Study Selection Criteria(SSC):There are challenges in the literature selection process, including the language of the study, knowledge of the authors, institutions, journals, or time of publication etc. Before catching on selection criteria, we follow careful consideration to insure fairness in opting primary studies that give significant substantiation about exploration questions.
D. Quality Assessment Criteria(QAC):The thing of assessing each primarily named study's quality is to insure that the studying are applicable and unprejudiced. We develop a set of quality criteria for assessing individual studies.
E. Data birth and monitoring(DEM): We precisely determine how the information needed from named studies would be attained and record their pieces of substantiation.
F. Data Synthesis(DS): Data conflation aims to organize and epitomize issues attained from the named studies. We follow a set of procedures to synthesize information more.

Reporting the Review: After completing the review of all the studies, we report the issues in a suitable form to the distribution channel and target followership[3].

Data coffers

We intended to collect as numerous workshop as possible that are applicable to our exploration questions. During collecting Deepfake discovery studies, we tried to include all the combinations of related hunt expressions or keywords to avoid any bias[2]. The crucial idea of using Boolean language for combining those searching terms with AND or OR. The hunt words can be outlined primarily(Deepfake OR FaceSwap OR Video manipulation OR Fake face/ image/ videotape) AND(discovery OR descry) OR(Facial Manipulation OR Digital Media Forensics). rather of counting on one or two sources, welooked into several depositories to insure a proper hunt. still, there are numerous digital depositories are available for ending the exploration papers[4].We named 10 popular depositories from them by considering their applicability and vacuity as listed below:

- Web of Science
- IEEE Xplore
- Digital Library
- ACM Digital
- Library Science Direct(ELSEVIER)
- SpringerLink
- Google Scholar
- Semantic Scholar
- Database Systems and Logic Programming(DBLP)

The depositories include journals, conferences, and libraries. We limit our hunt duration from January 2020 to December 2024.

Models Used

1. **Deep Learning Models:** In computer vision, deep literacy models have been used extensively due to their point birth and selection medium, as they can directly prize or learn features from the data. In Deepfake discovery studies, we set up the following deep literacy- grounded models have been used convolutional neural network(CNN) model(e.g., XceptionNet, GoogleNet, VGG, ResNet, Ef cent Net, HRNet, InceptionResNetV2, MobileNet, Incep tionV3, DenseNet, SuppressNet, StatsNet), intermittent Neural Network(RNN) model(e.g., LSTM, FaceNet), Bidirectional RNN model, Long-term intermittent 25502 TABLE . Distribution of used models[2]. Convolutional Neural Network(RCNN) model, Faster RCNN model, Hierarchical Memory Network(HMN) model,Multi-task Protruded CNNsMTCNN) model and Deep Ensemble literacy(DEL)

2. **Machine Learning Model**: This fashion creates a point vector by denying the right features using colorful state- of- art point selection algorithms. It also feeds this vector as input to train a classifier to classify whether the vids or images are manipulated by Deep fake or not. Support Vector Machine(SVM), Logistic Retrogression(LR), Multilayer Perceptron(MLP), Adaptive Boosting(AdaBoost), eXtreme Gradient Boosting(XGBoost), and K- Means clustering(k- MN), Random Forest(RF), Decision Tree(DT), Discriminant Analysis(DA), Naive Bayes(NB) and Multiple Instance Learning(MIL) are used as machine literacy- grounded models. Traditional machine literacy(ML) algorithms are necessary in comprehending the sense for any decision that could be expressed in mortal terms. similar styles are suitable for the Deepfake sphere as there's a better grasp of the data and processes. In addition, tuning hyperactive- parameters and changing model designs are much more manageable. The tree- grounded ML approaches, for illustration, Decision Tree, Random Forest, Extremely Randomized Trees etc..

3. **Statistical Model**: The statistical models are grounded on the use of the information-theoretic study for confirmation. In these models, the shortest paths are calculated between original and Deepfake vids images distributions. For illustration, in a significance is measured for mean regularizedcross-correlation scores between the original and the Deepfake vids, classifying them as fake or real. The frequently- applied statistical models are Anticipation- Maximization(EM), Total Variational(television) distance, Kullback- Leibler(KL) divergence, Jensen Shannon(JS) divergence, etc.

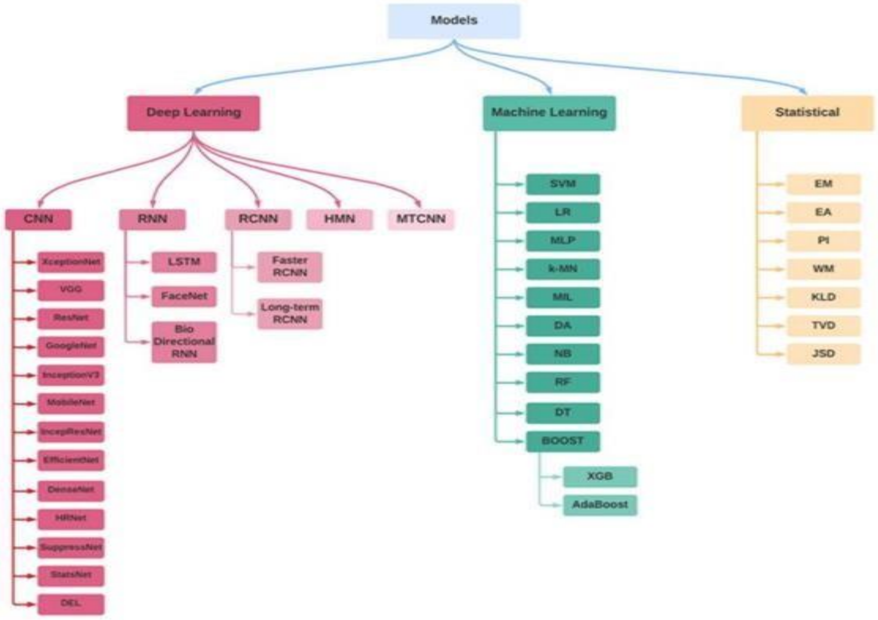
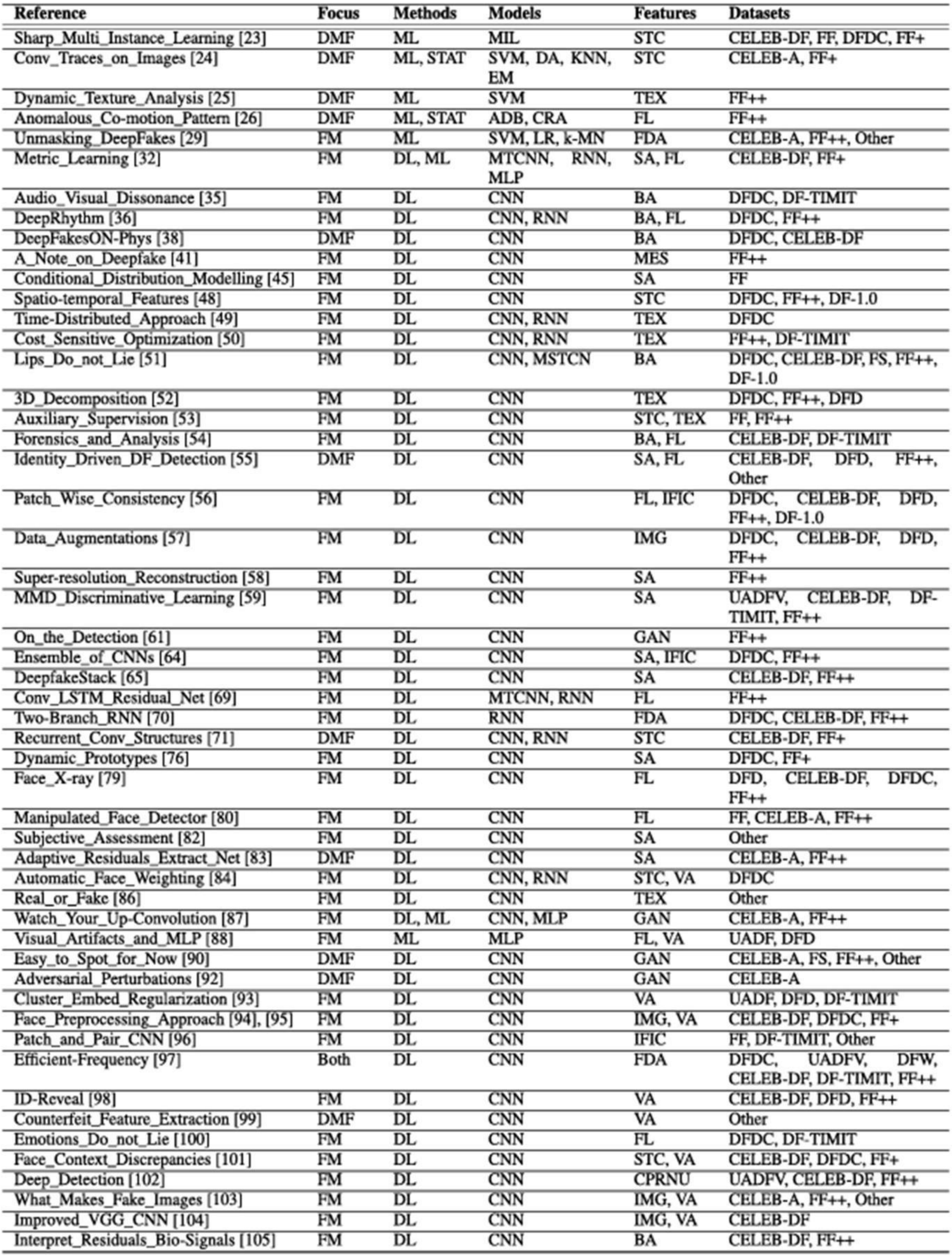


Figure: The list of Deepfake detection models.

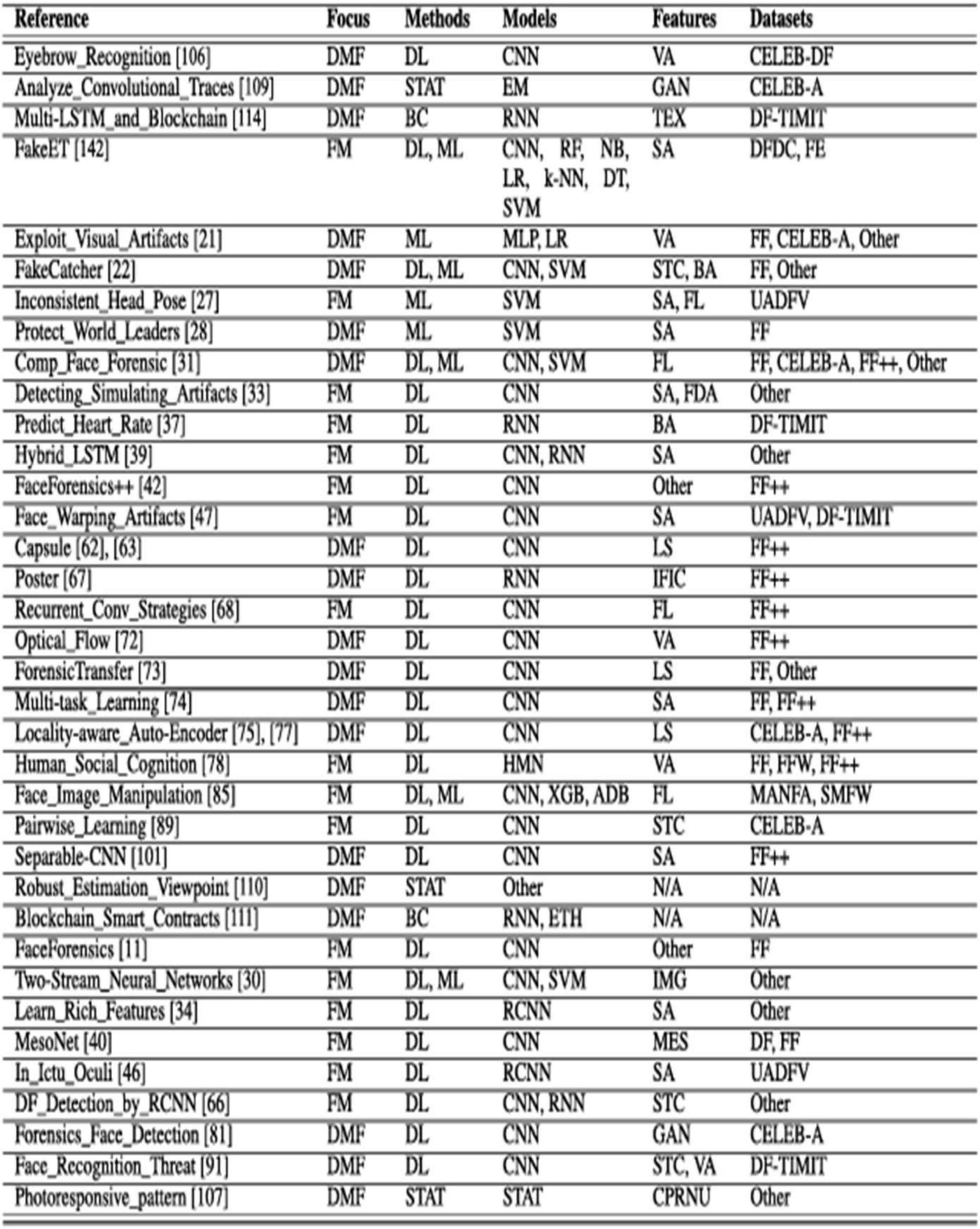
The field of deepfake detection employs a variety of models categorized into three major branches: Deep Learning, Machine Learning, and Statistical models. Deep Learning techniques, known for their ability to identify complex patterns in data, are further subdivided into several architectures. Convolutional Neural Networks (CNNs) include powerful models such as XceptionNet, VGG, ResNet, GoogleNet, Inception V2, MobileNet, InceptionResNet, EfficientNet, DenseNet, HRNet, SuppressNet, Statshe, and DEL, all of which are commonly used for image and video analysis in deepfake detection. Additionally, Recurrent Neural Networks (RNNs), which are well-suited for sequential data, include models like Long Short-Term Memory (LSTM) and Bi-Directional RNN. Region-based CNNs (RCNNs), such as Faster RCNN, FaceNet, and Long-term RCNN, focus on identifying manipulated regions within images or video frames. Other notable deep learning models include Holistic Memorization Networks (HMN) and Multi-task Cascaded Convolutional Networks (MTCNN), each contributing unique strengths in face detection and memory-based learning.

The Machine Learning branch includes more traditional yet effective models that work well when computational resources are limited or when simpler problems are being addressed. This category features classifiers and algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Multilayer Perceptron (MLP), k-Nearest Neighbors (k-NN), Multiple Instance Learning (MIL), Discriminant Analysis (DA), Naive Bayes (NB), Random Forest (RF), and Decision Trees (DT). Ensemble learning methods, including Boosting techniques like AdaBoost, XGBoost, and general BOOST, are also crucial in improving performance by combining the strengths of multiple base learners. Unlike deep learning models, these techniques are typically easier to interpret and train, and are particularly useful when working with structured datasets or limited training data.The third classification focuses on Statistical models, which provide foundational analytical tools for deepfake detection through statistical inference and distribution analysis. Methods like Expectation-Maximization (EM) and Evolutionary Algorithms (EA) are widely used in optimization and clustering tasks. Other statistical metrics such as the Kullback–Leibler Divergence (KLD), Jensen–Shannon Divergence (JSD), Total Variation Distance (TVD), and Wasserstein Metric (WM) help measure the difference between probability distributions, making them valuable for identifying anomalies introduced by synthetic media. The Proximity Index (PI) is also used to detect inconsistencies in data representation. These models are particularly effective in scenarios where manipulation introduces subtle changes in the statistical properties of media content. Together, these three branches form a comprehensive taxonomy of approaches used to tackle the complex challenge of detecting deepfakes

Beyond their individual capabilities, these models often complement one another in hybrid or ensemble frameworks for enhanced deepfake detection performance. For instance, deep learning models can be used to extract high-level features from images or videos, which are then fed into machine learning classifiers like SVM or Random Forest for final decision-making. Such combinations leverage the strengths of both approaches—deep learning's powerful feature extraction with machine learning's interpretability and lower computational cost. Similarly, statistical methods may be used alongside deep learning to validate the authenticity of media by analyzing feature distribution shifts. In recent research, ensemble systems that integrate multiple deep learning models or blend machine learning with statistical anomaly detection techniques have shown improved robustness against evolving deepfake generation tactics. This layered strategy helps counter the sophisticated nature of modern synthetic media and offers a more reliable defense against manipulation across diverse datasets and use cases. These hybrid approaches are particularly useful in real-world applications where deepfakes are increasingly used for misinformation, identity fraud, and digital impersonation.



Reference	Focus	Methods	Models	Features	Datasets
Sharp_Multi_Instance_Learning [23]	DMF	ML	MIL	STC	CELEB-DF, FF, DFDC, FF+
Conv_Traces_on_Images [24]	DMF	ML, STAT	SVM, DA, KNN, EM	STC	CELEB-A, FF+
Dynamic_Texture_Analysis [25]	DMF	ML	SVM	TEX	FF++
Anomalous_Co-motion_Pattern [26]	DMF	ML, STAT	ADB, CRA	FL	FF++
Unmasking_DeepFakes [29]	FM	ML	SVM, LR, k-MN	FDA	CELEB-A, FF++, Other
Metric_Learning [32]	FM	DL, ML	MTCNN, RNN, MLP	SA, FL	CELEB-DF, FF+
Audio_Visual_Dissonance [35]	FM	DL	CNN	BA	DFDC, DF-TIMIT
DeepRhythm [36]	FM	DL	CNN, RNN	BA, FL	DFDC, FF++
DeepFakesON-Phys [38]	DMF	DL	CNN	BA	DFDC, CELEB-DF
A_Note_on_Deepfake [41]	FM	DL	CNN	MES	FF++
Conditional_Distribution_Modelling [45]	FM	DL	CNN	SA	FF
Spatio-temporal_Features [48]	FM	DL	CNN	STC	DFDC, FF++, DF-1.0
Time-Distributed_Approach [49]	FM	DL	CNN, RNN	TEX	DFDC
Cost_Sensitive_Optimization [50]	FM	DL	CNN, RNN	TEX	FF++, DF-TIMIT
Lips_Do_not_Lie [51]	FM	DL	CNN, MSTCN	BA	DFDC, CELEB-DF, FS, FF++, DF-1.0
3D_Decomposition [52]	FM	DL	CNN	TEX	DFDC, FF++, DFD
Auxiliary_Supervision [53]	FM	DL	CNN	STC, TEX	FF, FF++
Forensics_and_Analysis [54]	FM	DL	CNN	BA, FL	CELEB-DF, DF-TIMIT
Identity_Driven_DF_Detection [55]	DMF	DL	CNN	SA, FL	CELEB-DF, DFD, FF++, Other
Patch_Wise_Consistency [56]	FM	DL	CNN	FL, IFIC	DFDC, CELEB-DF, DFD, FF++, DF-1.0
Data_Augmentations [57]	FM	DL	CNN	IMG	DFDC, CELEB-DF, DFD, FF++
Super-resolution_Reconstruction [58]	FM	DL	CNN	SA	FF++
MMD_Discriminative_Learning [59]	FM	DL	CNN	SA	UADFV, CELEB-DF, DF-TIMIT, FF++
On_the_Detection [61]	FM	DL	CNN	GAN	FF++
Ensemble_of_CNNs [64]	FM	DL	CNN	SA, IFIC	DFDC, FF++
DeepfakeStack [65]	FM	DL	CNN	SA	CELEB-DF, FF++
Conv_LSTM_Residual_Net [69]	FM	DL	MTCNN, RNN	FL	FF++
Two-Branch_RNN [70]	FM	DL	RNN	FDA	DFDC, CELEB-DF, FF++
Recurrent_Conv_Structures [71]	DMF	DL	CNN, RNN	STC	CELEB-DF, FF+
Dynamic_Prototypes [76]	FM	DL	CNN	SA	DFDC, FF+
Face_X-ray [79]	FM	DL	CNN	FL	DFD, CELEB-DF, DFDC, FF++
Manipulated_Face_Detector [80]	FM	DL	CNN	FL	FF, CELEB-A, FF++
Subjective_Assessment [82]	FM	DL	CNN	SA	Other
Adaptive_Residuals_Extract_Net [83]	DMF	DL	CNN	SA	CELEB-A, FF++
Automatic_Face_Weighting [84]	FM	DL	CNN, RNN	STC, VA	DFDC
Real_or_Fake [86]	FM	DL	CNN	TEX	Other
Watch_Your_Up-Convolution [87]	FM	DL, ML	CNN, MLP	GAN	CELEB-A, FF++
Visual_Artifacts_and_MLP [88]	FM	ML	MLP	FL, VA	UADF, DFD
Easy_to_Spot_for_Now [90]	DMF	DL	CNN	GAN	CELEB-A, FS, FF++, Other
Adversarial_Perturbations [92]	DMF	DL	CNN	GAN	CELEB-A
Cluster_Embed_Regularization [93]	FM	DL	CNN	VA	UADF, DFD, DF-TIMIT
Face_Preprocessing_Approach [94], [95]	FM	DL	CNN	IMG, VA	CELEB-DF, DFDC, FF+
Patch_and_Pair_CNN [96]	FM	DL	CNN	IFIC	FF, DF-TIMIT, Other
Efficient-Frequency [97]	Both	DL	CNN	FDA	DFDC, UADFV, DFW, CELEB-DF, DF-TIMIT, FF++
ID-Reveal [98]	FM	DL	CNN	VA	CELEB-DF, DFD, FF++
Counterfeit_Feature_Extraction [99]	DMF	DL	CNN	VA	Other
Emotions_Do_not_Lie [100]	FM	DL	CNN	FL	DFDC, DF-TIMIT
Face_Context_Discrepancies [101]	FM	DL	CNN	STC, VA	CELEB-DF, DFDC, FF+
Deep_Detection [102]	FM	DL	CNN	CPRNU	UADFV, CELEB-DF, FF++
What_Makes_Fake_Images [103]	FM	DL	CNN	IMG, VA	CELEB-A, FF++, Other
Improved_VGG_CNN [104]	FM	DL	CNN	IMG, VA	CELEB-DF
Interpret_Residuals_Bio-Signals [105]	FM	DL	CNN	BA	CELEB-DF, FF++



Reference	Focus	Methods	Models	Features	Datasets
Eyebrow_Recognition [106]	DMF	DL	CNN	VA	CELEB-DF
Analyze_Convolutional_Traces [109]	DMF	STAT	EM	GAN	CELEB-A
Multi-LSTM_and_Blockchain [114]	DMF	BC	RNN	TEX	DF-TIMIT
FakeET [142]	FM	DL, ML	CNN, RF, NB, LR, k-NN, DT, SVM	SA	DFDC, FE
Exploit_Visual_Artifacts [21]	DMF	ML	MLP, LR	VA	FF, CELEB-A, Other
FakeCatcher [22]	DMF	DL, ML	CNN, SVM	STC, BA	FF, Other
Inconsistent_Head_Pose [27]	FM	ML	SVM	SA, FL	UADFV
Protect_World_Leaders [28]	DMF	ML	SVM	SA	FF
Comp_Face_Forensic [31]	DMF	DL, ML	CNN, SVM	FL	FF, CELEB-A, FF++, Other
Detecting_Simulating_Artifacts [33]	FM	DL	CNN	SA, FDA	Other
Predict_Heart_Rate [37]	FM	DL	RNN	BA	DF-TIMIT
Hybrid_LSTM [39]	FM	DL	CNN, RNN	SA	Other
FaceForensics++ [42]	FM	DL	CNN	Other	FF++
Face_Warping_Artifacts [47]	FM	DL	CNN	SA	UADFV, DF-TIMIT
Capsule [62], [63]	DMF	DL	CNN	LS	FF++
Poster [67]	DMF	DL	RNN	IFIC	FF++
Recurrent_Conv_Strategies [68]	FM	DL	CNN	FL	FF++
Optical_Flow [72]	DMF	DL	CNN	VA	FF++
ForensicTransfer [73]	DMF	DL	CNN	LS	FF, Other
Multi-task_Learning [74]	DMF	DL	CNN	SA	FF, FF++
Locality-aware_Auto-Encoder [75], [77]	DMF	DL	CNN	LS	CELEB-A, FF++
Human_Social_Cognition [78]	FM	DL	HMN	VA	FF, FFW, FF++
Face_Image_Manipulation [85]	FM	DL, ML	CNN, XGB, ADB	FL	MANFA, SMFW
Pairwise_Learning [89]	FM	DL	CNN	STC	CELEB-A
Separable-CNN [101]	DMF	DL	CNN	SA	FF++
Robust_Estimation_Viewpoint [110]	DMF	STAT	Other	N/A	N/A
Blockchain_Smart_Contracts [111]	DMF	BC	RNN, ETH	N/A	N/A
FaceForensics [11]	FM	DL	CNN	Other	FF
Two-Stream_Neural_Networks [30]	FM	DL, ML	CNN, SVM	IMG	Other
Learn_Rich_Features [34]	FM	DL	RCNN	SA	Other
MesoNet [40]	FM	DL	CNN	MES	DF, FF
In_Ictu_Oculi [46]	FM	DL	RCNN	SA	UADFV
DF_Detection_by_RCNN [66]	FM	DL	CNN, RNN	STC	Other
Forensics_Face_Detection [81]	DMF	DL	CNN	GAN	CELEB-A
Face_Recognition_Threat [91]	DMF	DL	CNN	STC, VA	DF-TIMIT
Photoresponsive_pattern [107]	DMF	STAT	STAT	CPRNU	Other

Table displays the full performances of sensor groups that are set up from these primary studies, where CNN has the most divisions. Grounded on this Table[6], we further apply a subcategorization on CNN models and set up that the following 3 CNN models(i) XeptionNet and ResNet take 17 and(iii) VGGwith12, independently. either, LSTM model take 13 of RNN. In addition to this, the most popular machine literacy model is SVM with 12 and k- MN with 4. The detail distribution in colorful models is presented in Figure that shows the proportion of used models(e.g., DL, ML, Statistical) in colorful studies for detecting Deepfake. either, it provides the answer for SRQ- 2.3. The reviewed papers show that the deep neural network(DNN) models are successful in Deepfake discovery, where CNN- grounded models demonstrate further effectiveness among all the DNN models. At a regard. Focus indicates the indication for the discovery(DMF Digital Media Forensics[7], FM Face Manipulation, Both DMF and FM), styles indicates system delicacy bloody(ML Machine Learning, DL Deep Learning, STAT Statistical system, BC Blockchain), Models represents types of model(DL(CNN Convolutional Neural Network, RNN intermittent Neural Network, RCNN Regional 25504 Various techniques and models are employed in DeepFake detection, including deep learning, machine learning, statistical analysis, and blockchain-based approaches. In the deep learning domain, models such as Convolutional Neural Networks (CNN), Multi-task Cascaded CNN (MTCNN), Protruded CNN, and Multi-scale Temporal CNN (MSCNN) are widely used. Machine learning techniques include Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), K-Means Clustering (k-MN), XGBoost (XGB), AdaBoost (ADB), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbour (KNN), and Discriminant Analysis (DA). Statistical methods like Expectation Maximization (EM) and Correlation Analysis (CRA) are also utilized. In addition, Ethereum-based blockchain (ETH) solutions have been explored for ensuring data integrity and traceability.

Detection features are diverse and include Special Vestiges (SA), Visual Vestiges (VA), Biological Vestiges (BA), Face Landmarks (FL), Spatio-temporal Consistency (STC), Texture (TEX), Frequency Domain Analysis (FDA), Latent Space points (LS), GAN-generated points, Mesoscopic Features (MES), Intra-frame Inconsistency (IFIC), Contrastive and Photo-responsive PRNU pattern (CPRNU), Image Metadata (IMG), as well as techniques involving data Augmentation and Steganalysis. Other unique or uncommon features are also sometimes used depending on the detection model.

Several benchmark datasets support research and evaluation in this domain. These include FaceForensics (FF), Deepfake Detection (DFD), DeepFake Forensics V1 (CELEB-A), DeepFake Forensics V2 (Notoriety-DF), Deepfake Detection Challenge (DFDC), Deepfake-TIMIT (DF-TIMIT), DeeperForensics-1.0 (DF-1.0), Wild Deepfake (WDF), SwapMe and FaceSwap (SMFW), DeepFakes (DFS), Fake Faces in the Wild (FFD), FakeET (FE), FaceShifter (FS), generic Deepfake datasets (DF), Shifted Face Detection (SFD), Inconsistent Head Poses (UADFV), Tampered Face dataset (MANFA), and other custom datasets developed by various authors.

Eventually, we epitomize all at a regard using Table that specifies the features, styles and models, datasets used throughout the studies and also focuses on specific manipulation discovery ways with having a reference to each of the primary studies. For better perceptivity, we epitomize our crucial consummations in Figure. As demonstrated in Figure, we classify overall approaches concerning different rudiments similar as input data, features, system orders, and type of ways. A path between two rudiments denotes the affiliated factors used in the companion paper for any system. As presented in the Figure, utmost papers apply image or videotape as the input data, whereas numerous papers use both image and videotape as the input. Special Vestiges and Texture and Spatio-temporal thickness are the generally used features in colorful papers. About 75 of the styles used the DL- grounded ways as the discovery system order. Only a many papers used Blockchain and Statistical approaches for detecting similar Deepfake[2]. In detecting Deepfake, colorful underpinning ways are available, similar as natural Signals, Phoneme- Viseme Mis matches, facial expression and movements(i.e., 2D and 3D facial corner positions, head disguise, and facial action units), etc. We combine them under two central marquee las of the styles that include Facial Manipulation and Digital Media Forensics. As shown in Figure utmost of the DL- grounded styles exploit Facial Manipulation for the Deepfake discovery[7]. still, Machine Learning grounded styles nearly inversely use both ways. Common to both Blockchain and Statistical approaches, they apply only Digital Media Forensics as part of the discovery fashion. We resolve the models into two groups

I. Deep literacy- grounded models
II. Non-Deep Literacy grounded models

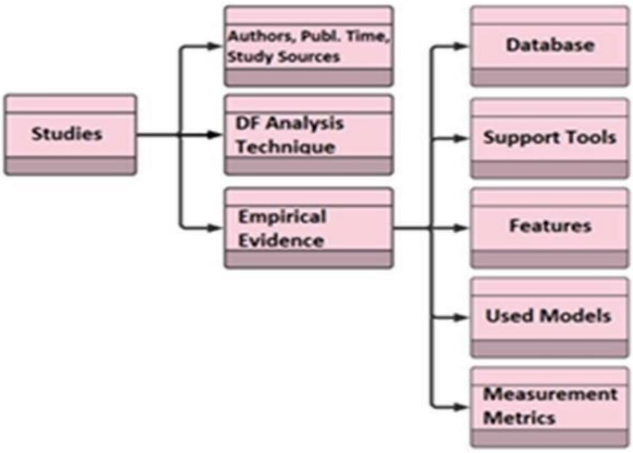
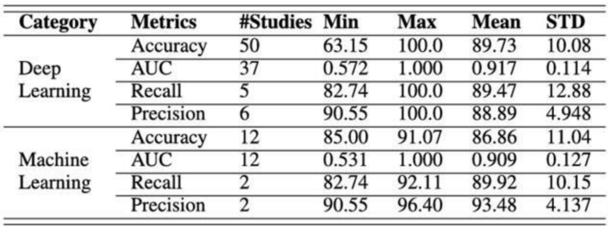


Figure: The information of the extracted data

We discourage mine the mean delicacy, AUC, recall, and perfection. Next, we apply a relative analysis of these two models performance and gain an average result. Grounded on the evaluation of these models using performance measures(delicacy, AUC, recall, and perfection), we observe, in general, deep literacy-grounded models outperformed non deep literacy- grounded models. As the results are reported in Figure, the delicacy and perfection performance in deep literacy models are significantly better thannon-deep literacy models. still, in the case of AUC and recall, the performance is enough analogous. The overall results demonstrate the superiority of deep literacy- grounded models overnon-deep literacy- grounded models.[8]**..**

We substantially consider eight different publication sources from honored conferences, shops, journals, and libraries. We observe that further papers were published as archived papers in the sphere of Deepfake, whereas a many papers were issued in the journal.

Compared with the traditional machine learning approaches, we note that applying deep learning algorithms to detect Deepfake from SRQ-2.3 has become a hot subject. We also nd that most studies follow a traditional CNN approach to classify Deepfake in the deep-learning environment[2].Still, researchers have not yet gured out how to determine Deep fake authorship.



Category	Metrics	#Studies	Min	Max	Mean	STD
Deep Learning	Accuracy	50	63.15	100.0	89.73	10.08
	AUC	37	0.572	1.000	0.917	0.114
	Recall	5	82.74	100.0	89.47	12.88
	Precision	6	90.55	100.0	88.89	4.948
Machine Learning	Accuracy	12	85.00	91.07	86.86	11.04
	AUC	12	0.531	1.000	0.909	0.127
	Recall	2	82.74	92.11	89.92	10.15
	Precision	2	90.55	96.40	93.48	4.137

Figure: Performance of various detection methods

Based on the outcomes of RQ-4,it is observed that the deep learning-based models achieve better performance than the non-deep learning models in Deepfake detection. Therefore, deep learning-based approaches are advised when detecting Deepfake.

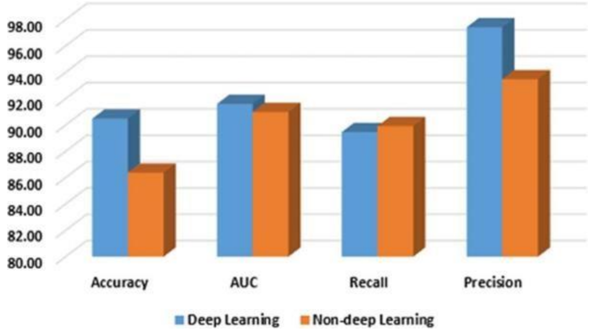


Figure: The comparison of the results among deep learning and non-deep learning based models.

3. Conclusion

This LR presents colorful state- of- the- art styles for detecting Deepfake published in 112 studies from the morning of 2020 to the end of 2024. We present introductory ways and bandy different discovery models effectiveness in this work. We epitomize the overall study as follows:

- The deep literacy- grounded styles are extensively used in detecting Deepfake. In the trials, the FF dataset occupies the largest proportion.
- The deep literacy(substantially CNN) models hold a significant chance of all the models.
- The most extensively used performance standard is discovery delicacy.
- The experimental results demonstrate that deep literacy ways are effective in detecting Deepfake. Further, it can be stated that, in general, the deep literacy models outperform thenon-deep literacy models.

With the rapid-fire progress in underpinning multimedia technology and the proliferation of tools and operations, Deepfake discovery still faces numerous challenges. We hope this SLR provides a precious resource for the exploration community in developing effective discovery styles and countermeasures.

References

1. FaceApp. Accessed: Jan. 4, 2021. [Online]. Available: https://www.faceapp.com/
2. FakeApp. Accessed: Jan. 4, 2021. [Online]. Available: https://www.fakeapp.org/

3. Oberoi, Exploring DeepFakes. Accessed: Jan. 4, 2021. [Online]. Available: https://goberoi.com/exploring-deepfakes-20c9947c22d9
4. J. Hui, How Deep Learning Fakes Videos (Deepfake) and How to Detect it. Accessed: Jan. 4, 2021. [Online]. Available: https://medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9
5. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 2, Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.
6. G. Patrini, F. Cavalli, and H. Ajder, The State of Deepfakes: Reality Under Attack, Deeptrace B.V., Amsterdam, The Netherlands, Annu. Rep. v.2.3., 2018. [Online]. Available: https://s3.eu-west-2.amazonaws.com/rep2018/2018-the-state-of-deepfakes.pdf
7. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2387–2395. doi: 10.1109/CVPR.2016.262.
8. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Oct. 2017, pp. 2242–2251. doi: 10.1109/ICCV.2017.244.
9. S. Suwajanakorn, S. M. Seitz, and I. K. Shlizerman, “Synthesizing Obama: Learning lip sync from audio,” ACM Trans. Graph., vol. 36, no. 4, p. 95, 2017.
10. L. Matsakis, Artificial Intelligence is Now Fighting Fake Porn. Accessed: Jan. 4, 2021. [Online]. Available: https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/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.
11. H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, “Deep video portraits,” ACM Trans. Graph., vol. 37, no. 4, pp. 1–14, Aug. 2018. doi: 10.1145/3197517.3201283.
12. C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, “Everybody dance now,” 2018. arXiv:1808.07371.
13. T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long
14. Beach, CA, USA, Jun. 2019, pp. 4396–4405. doi: 10.1109/CVPR.2019.00453.
15. D. Budgen and P. Brereton, “Performing systematic literature reviews in software engineering,” in Proc. 28th Int. Conf. Softw. Eng., New York, NY, USA, May 2006, pp. 1051–1052. doi: 10.1145/1134285.1134500.
16. Z. Stapic, E. G. Lopez, A. G. Cabot, L. M. Ortega, and V. Strahonja, “Performing systematic literature review in software engineering,” in Proc. 23rd Central Eur. Conf. Inf. Intell. Syst. (CECIIS), Varazdin, Croatia, Sep. 2012, pp. 441–447.
17. B. Kitchenham, Procedures for Performing Systematic Reviews, Software Engineering Group; Nat. ICT Aust., Keele; Eversleigh, Keele Univ., Keele, U.K., Tech. Rep. TR/SE-0401; NICTA Tech. Rep. 0400011T.1, 2004.
18. B. Kitchenham and S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Software Engineering Group; Keele Univ., Durham University Joint, Durham, U.K., Tech. Rep. EBSE-2007-01, 2007.