Enhancing Deepfake Detection with Explainable AI Through Neural-Symbolic RCNN-BiGRU Approach

Zainab Ali Abbood [1], Raghad Tariq Al-Hassan [2], Mahmoud Shuker Mahmoud [3,4], Atheel Sabih Shaker [5]

[1] Computer Technology Engineering Department, Al-Mansour University College, Baghdad, Iraq
Email: zainab.a.abbood@muc.edu.iq

[2]Ministry of Higher Education and Scientific Research, Minister's Office, Baghdad, Iraq
Email: eng_raghadtariq@yahoo.com

[3]Computer Engineering - Computer Networks, Gilgamesh University, Baghdad, Iraq
[4]Cybersecurity Technology Engineering Department, Middle Technical University, Electrical Engineering Technical College, Baghdad, Iraq
Email: mahmoud.shukur@gu.edu.iq

[5]Dept. of Computer Engineering Techniques, College of Technology Engineering Al-Iraqia Science University, Baghdad, Iraq
Email: Atheel.sabih @baghdadcollege.edu.iq

Article History

Received: Mar. 17, 2026
Revised: Jun. 12, 2026
Accepted: Jun. 11, 2026

Abstract

The rapid advancement of deepfake generation techniques poses a significant threat to digital media authenticity, necessitating detection systems that are not only accurate but also explainable and robust across diverse content types. While deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown promising performance in detecting spatial and temporal inconsistencies, they often operate as black-box systems with limited interpretability and generalization capability. To address these challenges, this paper proposes a neural-symbolic deepfake detection framework that integrates Explainable Artificial Intelligence (XAI) with hybrid deep learning models. The proposed approach combines Region-based Convolutional Neural Networks (RCNN) and Bidirectional Gated Recurrent Units (Bi-GRU) for effective spatiotemporal feature extraction. These features are further processed through a propositional inference layer that incorporates symbolic reasoning based on logical rules reflecting natural facial behavior, including eye movement, lip synchronization, and facial consistency. The model is evaluated on benchmark datasets, including Celeb-DF, Deepfake TIMIT, and WLDR, demonstrating superior performance compared to baseline methods in terms of F1-score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC), along with improved true positive rates in ROC analysis. Furthermore, ablation studies confirm that the integration of symbolic reasoning enhances detection performance by enforcing logical consistency and providing interpretable decision-making. Overall, the results highlight the effectiveness of neural-symbolic reasoning as a robust and transparent framework for deepfake detection, contributing to the advancement of explainable and trustworthy AI systems in multimedia forensics.

Keywords- Deepfake detection, Explainable Artificial Intelligence (XAI), Neural-symbolic reasoning, Deep learning, Multimedia.

I. INTRODUCTION

The rapid proliferation of AI-generated deepfake content has raised significant concerns regarding the integrity and authenticity of digital media. Expanding on the achievements of deep learning and generative models, including Generative Adversarial Networks (GANs), deepfakes have been further developed in both the quality of the generated content and the degree of believability, inflicting damage to the credibility of AV content in the social and news and, possibly, even legal spheres [1]. The authority to develop video images of truth that cannot be differentiated with reality is ethically, politically and security hazardous. Although CNN-based methods achieve high performance on in-domain datasets, their effectiveness often decreases when evaluated on unseen datasets due to limited generalization capability.

Although there have been unprecedented improvements in deep learning based deepfake detection, the current models have high generalization gaps and low interpretability and robustness to adversarial attacks. Traditional CNN and RNN designs typically act as a

black box and fail to justify the rationale of an ill-posed decision [2]. It is an obvious disadvantage in this high-stakes area as journalism, legal forensics, and cybersecurity. Moreover, neural models frequently overfit some sets of data and perform poorly cross domain, on those deepfakes that they have never seen before [3]. These challenges highlight the need for more effective and interpretable deepfake detection frameworks. It is also believed that neural-symbolic integration can also serve as a possible way of bridging this gulf. By combining the symbolic reasoning (e.g., logic-based rules and ontologies) and the feature learning of the neural networks, the hybrid models can be more interpretive without affecting the performance [4]. Symbolic layer allows models to make decisions on the basis of structured knowledge that is accompanied by interpretable justifications of classification results. This two-fold benefit allows neural-symbolic models to be natural contenders for the twin requirements of deepfake detection, precision and explainability.

In this work, we introduce a new deepfake detection architecture that combines deep learning and symbolic reasoning to enhance both performance and interpretability. To address this issue, our model integrates visual features extracted using deep learning and symbolic rule-based decision layers. Compared with traditional approaches, the symbolic reasoning part enables the system to find inconsistencies among various temporal and spatial cues, which makes the system more robust against adversarial deepfakes. The architecture also facilitates knowledge transfer with human-interpretable rules, to be able to generalize with the development of new deepfakes. Figure 1 depicts the temporal evolution of deepfake detection techniques, spanning from classic convolutional models to the recent use of hybrid neural-symbolic models. This shift is due to a changing focus from interpretability and robustness to interpretability and domain adaptation, and it is likely that deepfake media and the attacker strategies are becoming more complex influencing this shift.

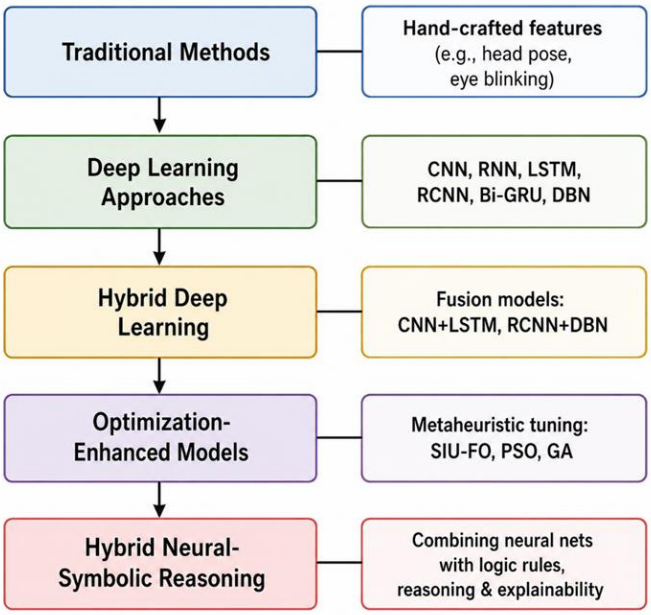

Figure , 1 (Evolution of deepfake detection methods from traditional to hybrid neural-symbolic models)

II. RELATED WORK

The techniques of early detection of deepfakes were mainly based on CNNs due to the high efficiency of this image classification method. These techniques are aimed at the identification of spatial integrity artifacts and the presence of subtle artifacts such as blending and low-frequency distortion on the tampered frames [5]. XceptionNet and VGGFace are typical CNN-based methods employed to distill frame-wise features and then a softmax classifier is employed to make a prediction. In domain data, CNNs can be quite effective on new data. And they do not have temporal sensitivity, which is important in the identification of dynamic inconsistency in video sequences [6].

To manage the time factor, several of the more recent researches suggested the inclusion of Recurrent Neural Networks (RNNs) into the deepfake detection models, including LSTM [5] and GRU [7]. These models are capable of learning temporal transitions between facial movements, mouth shapes or even eye blinking patterns, whilst these transitions are often simplified in fake content [7]. An example of such extensions is Bi-directional GRU (Bi-GRU) that works on sequences in both directions to encourage the recognition of temporal coherence [8]. Despite impressive advances in action recognition, RNNs still do not address such issues as interpretability of prediction and sensitivity to input sequences with different quality and length.

Region-based CNNs (RCNNs) also improve the accuracy of detection, and dominate the tampered region and extract region-specific tampering indication. Localization can be useful in the detection of the bias of facial landmarks or uneven illumination [9]. Another generative model (the Deep Belief Networks, or DBN) has been used to model complex feature distributions, but typically they need

large amounts of training data and are less scalable in real-time applications [10]. Although the methods of RCNN and DBN have resulted in improvements with limited resources, their lack of transparency has barred their usage in the field of forensics and legal practice, where model explainability is critical.

In recent works, models for feature fusion and optimization techniques to improve the detection of deepfakes were studied. The Dual Feature Pooling (DFP) strategy for feature enhancement by pooling features on different scales by thematic and temporal means is one of them [11]. Furthermore, dynamic fine-tuning model pars have been conducted with Swarm-type optimization algorithms, such as the Sunflower Optimization with Jellyfish Search Optimizer (SU-JFO) to enhance the classification accuracy of different datasets in [12]. Although these hybrid methods have performed quite well, they are still black boxes that adversarial deepfakes can avoid. A negative aspect of the earlier neural-based approaches is that it cannot be explained. The majority of deep learning systems are able to achieve high accuracy on a prediction task, but they do not enough to avoid providing human understandable justifications for what it does [13]. The issue, however, is that this makes it difficult to rely on the output of investigators and stakeholders in the real-world contexts, i.e., law enforcement or journalism. Moreover, the necessity of large-scale annotated labels and the necessity to train on the basis of a huge labeled dataset turns out to be an impediment to adapting to the new types of deepfakes. The existing trend toward more complicated models, as well as increased computational cost and opaque models, also makes the desirability of increased explanatory detection methodologies more attractive.

Hybrid neural-symbolic reasoning has become a convenient solution to these gaps. This paradigm combines the learning ability of neural networks and the logical form and inference ability of symbolic AI, such as logic rules or knowledge graphs [14]. Neural-symbolic methods in recent years have made encouraging progress in a range of different tasks such as visual question answering, object reasoning and medical decision support, through enabling explainable, logic-based inferencing over learned features. Applying the same concept to automatic deepfake detectors would give the chance not only to boost the accuracy of the detection but also to have decisions that are rational and transparent enough to be audited, which is an essential aspect of a reliable AI system. To compare other deepfake detection works reported in this section, Table 1 summarizes the related art as suggested in this paper in three aspects, namely, model type, main features and strengths and limitations. This systematic survey of the state-of-the-art leads to the interpretability and robustness gap as the most essential one and gives the rationale of the presence of neural-symbolic reasoning in our model.

TABLE 1. COMPARATIVE ANALYSIS OF DEEPFAKE DETECTION TECHNIQUES AND THEIR LIMITATIONS.

Model / Technique	Key Features	Strengths	Limitations
CNN	Spatial feature extraction from individual frames	High accuracy on image data	Lacks temporal context; poor generalization
RNN / LSTM / Bi-GRU	Temporal modeling of facial movement sequences	Captures motion and transitions	Limited interpretability; sensitive to input length
RCNN	Region-focused feature learning	Localizes facial anomalies	Computationally intensive; lacks transparency
DBN	Probabilistic feature representation	Models' complex distributions	Requires large datasets; not interpretable
DFP	Multi-level spatial-temporal feature fusion	Rich feature representation	Still relies on opaque neural layers
SU-JFO (optimization)	Metaheuristic model parameter tuning	Enhances performance and convergence speed	No effect on model explainability
Neural-Symbolic Reasoning	Combines neural learning with symbolic inference	High interpretability; logical reasoning; better trust	Still under exploration; integration complexity

III. METHODOLOGY

A. Hybrid Neural-Symbolic Architecture

Two essential problems in the deepfake detection system that are addressed in the design of the architecture developed in this work are accurate identification and interpretable and reliable decision-making. For this purpose, we develop a hybrid system that is based on the integration of deep neural learning and symbolic reasoning. Such hybridization also allows the model not only to learn the complex patterns by watching the video, but also to think by applying interpretable, rule-based logic by using these learned representations. It has two huge layers in structure:

1) Neural Feature Extraction Module: The module locates and amplifies the sections of the face: eyes, mouth and cheeks, at which it is typically the most difficult to discern any kind of tampering, with the help of region-of-Interest-based Convolutional neural networks (RCNN). The extracted features in spatial format are then input to a Bidirectional Gated Recurrent Unit (Bi-GRU) model to learn the temporal changes between the video frames. The bi-directional allows the system to receive forward and reverse dependency, hence providing a larger detection of inconsistent movement, blinking pattern or lip-synch errors.

2) Symbolic Reasoning Layer: On the output of the layer, the deep features are transformed into symbolic predicates by rule templates or knowledge graphs. Such predicates are then based on logic-based reasoning systems like LTNs or neuro-symbolic concept learners. To illustrate, when the rule is that eye blinking should appear in every three seconds, the system

would examine whether or not the neural output is in compliance with the rule. The symbolic layer is a part of the final decision on the basis of the discrepancies and the possible inconsistencies in the sequences.

This complex architecture allows the system to combine uncooked detection power and semantic clarification. In addition to this, traceability is also provided that enables the forensic investigators to understand which rules are violated and how this reflects the classification output. The neural pattern recognition and the enforcing of the symbols lead to the ideal applicability of the system to delicate tasks, including digital forensics, news validation, and authenticating the evidence in courts. Figure 2: Schematic figure of the hybrid neural-symbolic architecture of the spatial-temporal feature extraction and logic-based reasoning for effective deepfake detection.

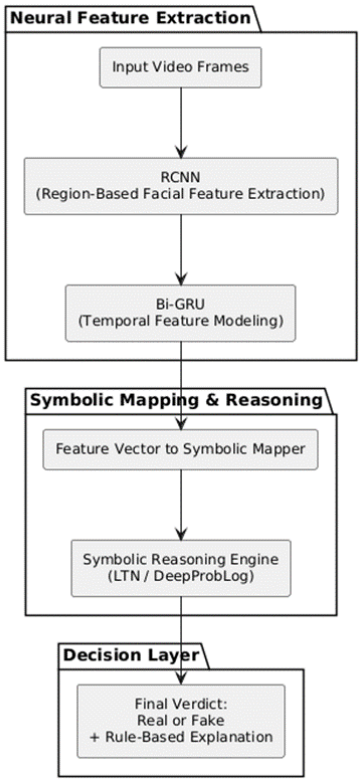

Figure 2 (Architecture of Proposed Hybrid Neural-Symbolic Deepfake Detection Model)

To have an overview of the duties and techniques of each stage of the hybrid architecture, we present in Table 2 the duties of the main kernel functions. The architecture consists of different modules, which are cycled through neural feature extraction, symbolic reasoning and final decision, each of which is dedicated to various steps of the deepfake detection pipeline. This modularity-based analysis depicts the joint exploitation of spatiotemporal-logical information to make accurate and informative predictions.

TABLE 2. KEY FUNCTIONAL ROLES OF NEURAL AND SYMBOLIC COMPONENTS IN THE PROPOSED ARCHITECTURE

Component	Function	Techniques Used
RCNN	Spatial feature localization in facial regions	Region-based CNN
Bi-GRU	Temporal pattern analysis across video frames	Bidirectional GRU
Symbolic Mapper	Converts features into logic-based predicates	Rule templates, ontologies
Reasoning Engine	Applies logical inference on symbolic data	LTN, DeepProbLog, NSCL
Decision Module	Combines neural and symbolic outputs	Weighted voting / rule thresholds

B. Neural Component

The neural component is used as a backbone of feature extraction in the proposed hybrid architecture, combining the benefits of spatial and temporal modeling. It primarily consists of two layers, namely: Region-based Convolutional Neural Network (RCNN) and Bidirectional Gated Recurrent Unit (BiGRU). The role of the R-CNN is to detect and localize the facial key regions, including eyes, nose, mouth and jawline. They are known to be susceptible to deepfake manipulation and often have texture anomalies, unnatural blending, or flickering artifacts. The RCNN is also useful at localizing them and extracting high-resol ution fine grained spatial features due to its region proposal and convolutional layer.

The RCNN output is input into the Bi-GRU module, which characterizes the temporal dynamics of the characteristics of the successive frames. In contrast to the traditional RNNs, Bi-GRU operates on the feature sequences both in reverse and forward direction as the sound is flowing in two directions in parallel in the past and future frames with respect to the sound frames. This can be extremely helpful to identify unnatural temporal motions (irregular blinking, lip-sync inconsistency, and head movement), which are the trademarks of deepfakes. The two-way strategy leads to more detailed time modeling, taking into consideration misleading cases, which are caused by temporary expressions of nature.

The RCNN–Bi-GRU cascade helps the neural component to capture highly discriminative spatial-temporal facial behaviors and appearance patterns. This architecture can seamlessly integrate frame-level manipulation analysis and sequence-level temporal coherence modeling, which is especially suitable for the identification of advanced deepfakes that cannot be accurately identified by frame-level classifiers alone. The architecture embeds spatial and temporal learning tasks assigning them to specific architectures, which results in better modularity, interpretability and training efficiency. The spatial-temporal features extracted are then passed to the symbolic reasoning layer, where logical inference and consistency checking based on reasoning rules are conducted. Figure 3 shows the proposed deepfake feature extraction pipeline, which includes facial regions detection and RCNN processing, followed by Bi-GRU processing, to extract rich spatial-temporal feature representations for explainable deepfake detection.

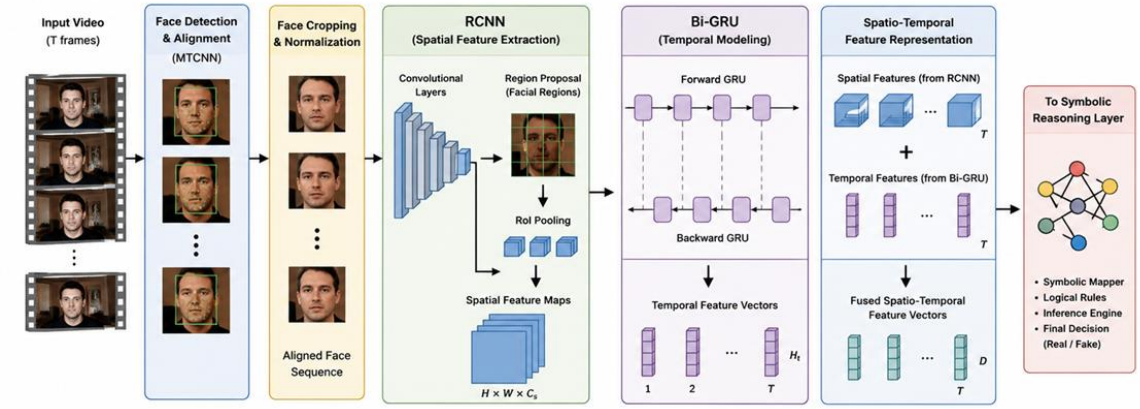

Figure , 3 (Neural feature extraction pipeline of the proposed deepfake detection framework using RCNN for spatial feature extraction and Bi-GRU for temporal sequence modeling)

C. Symbolic Reasoning Layer

To this purpose, the suggested architecture comprises a symbolic reasoning layer, which can be viewed as a visualization methodology to give hints about the significance of the features detected by the neural module. Yet, deep learning-based models excel at the task of capturing complex patterns, but their decision-making is not interpretable. The symbolic layer can solve this problem by introducing rule-based logic and semantical reasoning on the learned features, thereby giving more trustworthy and interpretable results.

This reasoning layer accepts the outputs of RCNN and Bi-GRU as the inputs and converts them to symbolic predicates. These predicates indicate logical association of facial behaviors: e.g., blink consistent (True), lip sync (Aligned), or motion coherence (High). Such predicates are then refuted with respect to frames under potential conditions, by means of rules, heuristics or domain ontologies, as specified by experts. As an example, a logic rule may demand that the blinking be done periodically, or that the lip motion be synchronized with speech characteristics obtained by analysis of audio.

In the case of such a layer, we propose to apply Logic Tensor Networks (LTNs) or Neuro-Symbolic Concept Learners (NSCL). These frameworks are able to mix automatically symbolic reasoning and continuous-valued neural representations, and offer reasoning based on uncertainty. Symbolic truth values are both differentiable and real-valued, as well as admit fuzzy logic operations, in certain cases, including LTNs. This enables the use of symbolic constraints during the training with the help of backpropagation. Unlike post-hoc explainability, the model logic is in-process, meaning that the symbolic logic actually informs the model to make the decision and thereby increases both performance and explainability.

The symbolic reasoning layer is also able to identify the inconsistencies and provide a reason as to why a sample was being classified as a deepfake by injecting logical rules into the inference process. This skill can be applied especially in situations that involve high stakes, such as content validation impact or misinformation debunking. The symbolic layer is a transparent rule-checker that harnesses the neural network predictions and the human-understandable logic can be verified. Figure 4 illustrates how the symbolic reasoning layer works, whereby neural features are converted into logic predicates and reasoning is carried out to offer more transparency to the decision.

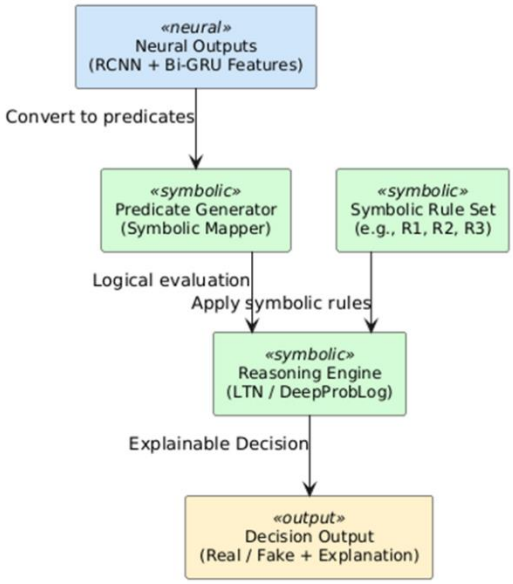

Figure , 4 (Symbolic Reasoning Workflow over Neural Outputs)

D. Integration Mechanism

The proposed hybrid architecture is based upon this integration mechanism, between the neural and symbolic layers, in which information and reasoning can be easily mixed. This mechanism helps to protect the complementation of the strengths actually acquired in deep learning to recognize complex patterns and perform symbolic reasoning. It is modeled as a one-way or two-way pipeline, based on the presence or absence of feedback of symbolic evaluation training. During the forward pass, learnable structures surrounding the RCNN and Bi-GRU extracted features are first converted into a structured format (by using symbolic predicates). These predicates turn raw feature values into discrete logical predicates (e.g., blinks consistent(true), head motion smooth(true), and lip sync(false) which the reasoning engine can inspect the symbolic state with. This operation is performed by the Symbolic Mapper, that is mapping patterns of learned characteristics to symbolic labels by means of rule templates or domain-specific heuristics.

These logical predicates are then manipulated in the symbolic reasoning engine with the help of tools like LTN or DePaolo in order to deduce whether they conform to a collection of established rules or constraints. After this assessment, a judgment is made along with a human interpretable description of the violated rules and their reasons. This flattened output creates more model fidelity and parsimony, two characteristics that are remarkably lacking in deep learning. There is also the constraint-based feedback in the system in both directions. In this work, the violation of the rules in the symbolic layer might be detected and, in its turn, it may result in the backward signals correcting the weights of the neural units. This is done through constraint driven backpropagation, which conditions the neural network with standard classification loss and with symbolic consistency penalties. This feedback not only promotes semantic coherence but also decreases the chances of overfitting, as it will help, via learned patterns, to promote alignment with human-understandable logic.

This close integration improves overall system resilience, especially in cross-domain or adversarial situations, and meets the characteristics of real-world systems that need auditability and accountability. Figure 5 presents the learning architecture integrating the two types of components, with an emphasis being placed on the bi-directional interaction from deep feature extraction to logical judgement and from backprop through logical judgement.

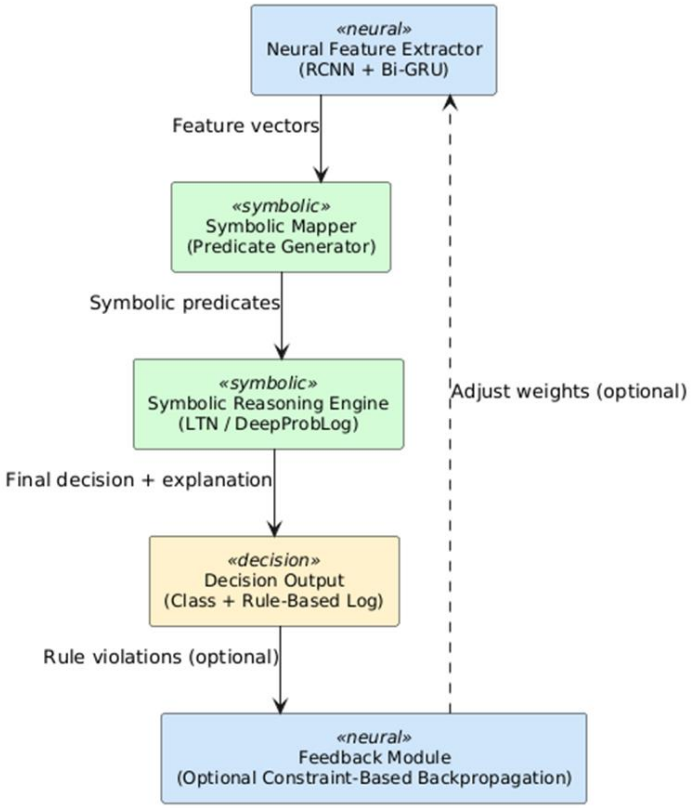

Figure 5 (Bidirectional Integration of Neural and Symbolic Components in the Proposed Architecture)

E. Training Strategy

The training approach for the proposed hybrid neural-symbolic architecture is formulated to leverage the two sources of loss, i.e., predictive accuracy and logical consistency. In contrast to existing deepfake detectors which are generally based on supervised neural learning, our model is based on a dual-objective training strategy. This hybrid approach consists of using a traditional classification loss term from the neural portion combined with a symbolic constraint loss, which ensures that the model can be interpreted using human understandable logic during training.

The neural model, i.e., RCNN + Bi-GRU, is initially trained with a cross-entropy loss which penalizes faulty deepfake classification. At the same time, the symbolic inference engine evaluates logical predicates over the neural weights. Penalties for any violated rules of logic (e.g., inconsistent blinking or physically unrealistic motion trajectories) are accumulated using symbolic regularization terms. These penalties are combined with the original loss to make a hybrid loss to drive the training toward not only the performance, but also the interpretability. The hybrid loss function can be defined mathematically as (Eq. 1):

$$L\_total = L\_classification + \lambda \cdot L\_symbolic \quad (1)$$

Where:

L_classification is the standard cross-entropy loss
L_symbolic represents logic rule violations
$\lambda \in [0,1]$ controls the symbolic constraint's weight

The preparation of data involves cutting the input videos into frame sequences and extracting the face areas using face detectors (e.g., MTCNN or Haar Cascades) that have been trained. Such segments are not only labeled with equivalent class (real/ fake) label, but also by relational and symbolic predicates (e.g, blinking smoothness, head motion smoothness). This dual-labeling enables us to do supervised learning in the network training as well as logic matching. The hybrid training model is capable of accomplishing both deepfake recognition and the ability to adapt to the symbolic system of the domain, as specified by the end user, and hence can be robust, explainable, and generalized. Figure 6 shows the training pipeline of the proposed hybrid model, which emphasizes the hybrid loss mechanism and metaheuristic procedure to balance between accuracy and logic matching.

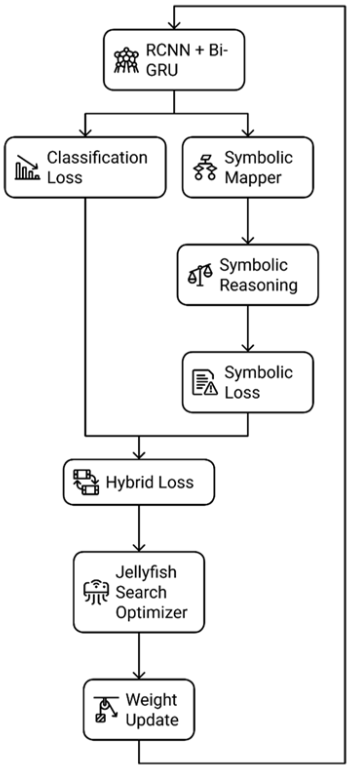

Figure 6 (Hybrid Training Pipeline Integrating Neural and Symbolic Loss Components)

Table 3 summarizes the primary training aspects of the hybrid architecture to give a better insight into the individual components and their associated optimization objectives. It describes the nature of the loss functions on each of the modules, how they are used in the learning process and how they are optimized to achieve good convergence and generalization.

TABLE 3. TRAINING COMPONENTS AND OPTIMIZATION OBJECTIVES OF THE HYBRID ARCHITECTURE

Component	Loss Type / Function	Purpose
Neural Network	Cross-Entropy Loss	Deepfake classification accuracy
Symbolic Reasoning	Constraint Violation Penalty	Enforce logical consistency
Combined Loss	Hybrid Total Loss	Joint optimization objective
Optimizer	Jellyfish Search Optimizer (JFO)	Dynamic hyperparameter tuning

IV. EXPERIMENTAL SETUP

The datasets, metrics used to evaluate performance, baseline models and technical environment to evaluate performance of the hybrid neural-symbolic deepfake detector framework. The experiments were aimed at assessing the correctness of the classification, and more interestingly, to assess interpretability and possibilities of generalization of the model under various conditions of data.

A. Datasets

To offer a better understanding of the datasets involved in our experiments, Table 4 summarizes the main properties of them, such as sample size, type of manipulation, resolution, and their purpose in testing the robustness of models to different levels of content and quality.

TABLE 4. SUMMARY OF DATASETS USED FOR DEEPFAKE DETECTION EVALUATION

Dataset	No. of Videos	Manipulation Type	Resolution	Notes
Celeb-DF (v2)	5,639 real / 5,101 fake	High-quality GAN-based synthesis	High (720p)	Realistic artifacts; suitable for benchmarking
Deepfake TIMIT	620 total (low/high quality)	Face-swapping using TIMIT audio	Medium (128×128)	Covers both low- and high-quality fakes
WLDR	1,200 (approx.) mixed-modality	Real-world manipulations	Mixed	Includes blur, occlusions, noise

B. Evaluation Metrics

To get a complete picture of the effectiveness of the proposed hybrid neural-symbolic model, we applied a complete range of evaluation metrics. These metrics are not only applied in determining the accuracy of the classification, but also in utilizing class imbalance, false prediction and a trade-off between precision and recall. Individual measures were obtained on every dataset and averaged on 5 cross-validation folds to guarantee statistical reliability. It was observed that the suggested hybrid neural-symbolic model was suitable for use with a list of conventional measures of classification. These measures are concerned with accuracy, error

distribution and precision-recall trade-off, with the imbalance of classes. To make it sure that they received credible results, they averaged them over five cross-validation folds and calculated them over all datasets. Table 5 contains the summary of the definition and formulations of the evaluation metric.

TABLE 5. SUMMARY OF EVALUATION METRICS

Metric	Description & Formula
Accuracy (ACC)	Measures overall correctness of predictions. $\boldsymbol{Accuracy} = \frac{TP+TN}{FP+FN+TP+TN}$
Precision (P)	Measures the proportion of correctly predicted deepfakes among all predicted deepfakes. $\mathbf{Precision} = \frac{TP}{FN+TP}$
Recall (R)	Measures the proportion of actual deepfakes correctly detected. $\boldsymbol{Recall} = \frac{TP}{FN+TP}$
F1-Score	Harmonic mean of Precision and Recall. $\boldsymbol{F1-Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$
MCC	Measures correlation between predicted and actual classes (robust for imbalance). $\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$
AUC	Measures model discrimination ability across thresholds. $AUC = \int_0^1 TPR\ (FPR)\ dFPR$

C. Evaluation Metrics

In order to prove that the proposed hybrid neural-symbolic model is superior to the alternatives, we conducted comprehensive comparisons with several models of state-of-the-art deepfake detectors. These baselines cover a wide spectrum of architectural paradigms: convolutional, recurrent, generative and region-based networks, and provide us with an opportunity to test our model with other learning strategies and feature extraction.

1) CNN Based Methods (eg XceptionNet): CNN based models process every single frame, and consist of spatial convolutional layers, to detect artifacts, including blending irregularities, edge corruption or color irregularity. XceptionNet has been proven to be state-of-the-art individually on large-scale benchmarks of deepfakes, in part due to the existence of depthwise separable convolutions.
2) RNN/Bi-GRU models: These are sequential video input-specific tensor models, which directly focus on the temporal discrepancies of unnatural blinking patterns and stuttering mouth movement. The Bi-GRUs are more contextual because they deal with frame reversal direction sequence and forward direction sequence.
3) Deep Belief Network (DBN): A DBN is a class of generative models that are built using stacked RBMs 20 and learn hierarchical representations. They provide a way of performing probability modeling of feature distributions; however, they are less interpretable and take longer to train than more modern discriminative models.
4) RCNN-based Detectors: These approaches incorporate spatial localization by training on the proposed facial region rather than on full frames. RCNNs are more suited to detect fine-grained changes using local regions— eyes, mouth and facial landmarks, thereby the demand for resources is higher.

All the baselines were re-implemented from public codebases or taken from pertained settings, and re-trained under the same settings (data splits, preprocessing and evaluation functions). It allows to fairly and reproducible comparison of the proposed hybrid method improvements.

D. Implementation Details

To ensure reproducibility and provide transparency regarding the experimental setup, Table 6 outlines the hardware, software environment, and key hyperparameters used in the implementation of the proposed hybrid deepfake detection model. The configuration balances training efficiency with high-performance symbolic reasoning capabilities.

TABLE 6. IMPLEMENTATION ENVIRONMENT

Parameter	Specification / Value
Operating System	Ubuntu 20.04
GPU	NVIDIA RTX 3090 (24 GB VRAM)
CPU	Intel Core i9
RAM	64 GB
Programming Language	Python 3.9
Frameworks	PyTorch 1.13, TensorFlow 2.11, LTNtorch
Learning Rate	0.0001 (adaptive via Jellyfish Search Optimizer)

Batch Size	16
Epochs	50
Dropout Rate	0.4
λ (Symbolic Loss Weight)	0.2
Optimizer	Jellyfish Search Optimizer (JFO)
Predicate Mapping Threshold	0.65 (empirically tuned)

V. RESULTS

We present the experimentation of the proposed hybrid neural-symbolic deepfake detection framework in this section. We evaluate and compare its performance with several state-of-the-art base-line models over various datasets. The results show that the incorporation of symbolic reasoning into deep learning pipeline not only improves the detection performance but also the generalization and interpretability. We also implement an ablation study to measure the individual contributions of the symbolic elements to the overall accuracy.

For a thorough analysis, we use a plethora of evaluation metrics that measure various aspects of model performance. Specifically, we have described accuracy (overall accuracy), precision (deepfake detection confidence), recall (way to detect all the manipulated samples), and both AUC and MCC (robustness to class imbalance) properties.

Finally, a comparative summary of the model's performance on 3 benchmark testing datasets (Celeb-DF, Deepfake TIMIT and WLDR) using a subset of evaluation metrics (Accuracy, Precision, F1-Score, AUC and MCC) is illustrated in Figure 7, Table 7. These indices were chosen in order to measure the overall performance of the classification and robustness toward class imbalance. As can be seen in the table, the hybrid model consistently outperforms the baseline deepfake detection methods such as CNN, Bi-GRU, and DBN, as well as the language-dependent classifier and RCNN-based approach across all the datasets.

On the high-resolution low-visual artifact dataset Celeb-DF, the hybrid model reached an F1-score of 0.94 (reasonably balanced precision and recall). On the other hand, conventional CNN and Bi-GRU models reported 0.89 and 0.86. This is partly due to the fact that the hybrid model identifies anomalies in semantic level, e.g., frequency of blinking eyes is uncharacteristic or unusual head gestures are natural, with symbolic logic rules that it can hardly be trained with the purely data-driven methodologies.

For Deepfake TIMIT, which has both low-resolution and realistic facial changes, a hybrid approach got an AUC of 0.95, which is better than RCNN (0.92) and DBN (0.85). The large AUC value demonstrates that the model is appropriate when it comes to various forms of manipulation. Interestingly, the symbolic module also enhances this ability by identifying temporal logical inconsistencies with the help of logic, such as misaligned lips and delayed facial expressions, which has so far rarely been considered by convolutional models or recurrent models.

The model outperformed CNN (0.74) and RCNN (0.76) on the WLDR dataset which contains noisy real-world scenes where the lighting and motion blur are not under control. The correlation coefficient was 0.83, Matthews Correlation Coefficient (MCC). The MCC is especially useful in cases where the classes are not balanced, as it takes into account all four confusion classes of the confusion matrix. The hybrid models worked well in such challenging visual conditions and few false positives and false negatives occurred. This means that they are good generalists. In general, the findings verify the usefulness of integrating deep neural feature extraction and symbolic reasoning processes. The suggested architecture is more precise in classification and generalization as well as a certain level of explainability, which is unattainable with conventional models. It can be used in the critical work of digital media forensics, content verification or legal evidence investigations.

TABLE 7. PERFORMANCE COMPARISON

Model	Celeb-DF (F1)	Deepfake TIMIT (AUC)	WLDR (MCC)
CNN (XceptionNet)	0.89	0.91	0.74
Bi-GRU	0.86	0.89	0.71
DBN	0.82	0.85	0.69
RCNN	0.91	0.92	0.76
Hybrid (Ours)	0.94	0.95	0.83

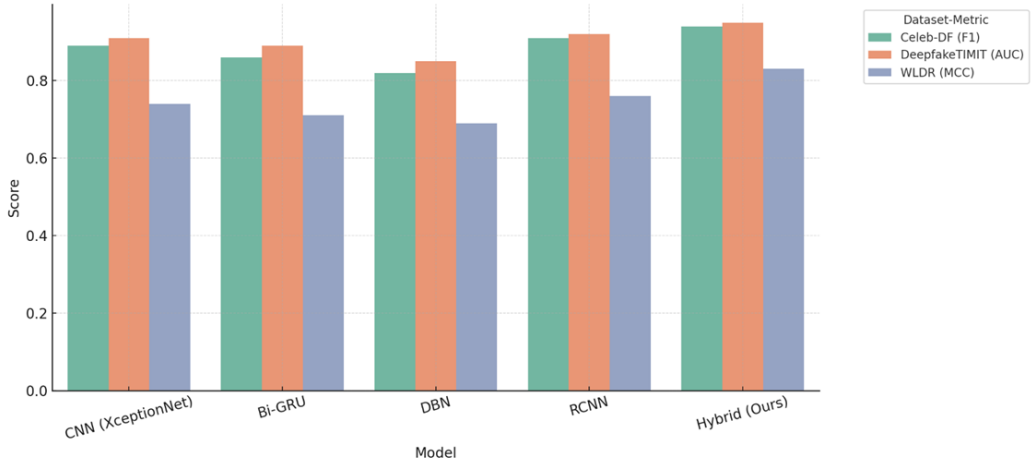

Figure 7 (Comparative performance of baseline models and the proposed hybrid model across celeb-df, deepfake, and WLDR datasets using F1 and AUC metrics)

In order to construct a more persuasive analogy, as well as to support the effectiveness of the sought hybrid model, a further in-depth comparison of the performance measurements was conducted against other benchmark databases. The details of the approach are described in Table 8, and specifically consider accuracy on Celeb-DF, precision on DeepfakeTIMIT, and recall on WLDR, all of which are selected to emphasize a different feature of classification performance. The hybrid model is the top performer in all benchmarks, demonstrating its potential to generalise and the best compromise between precision, recall and generalization. These findings provide further evidence in favor of the benefits of incorporating symbolic reasoning into the deepfake recognition pipeline.

TABLE 8. ADDITIONAL EVALUATION RESULTS

Model	Celeb-DF (ACC)	DeepfakeTIMIT (Precision)	WLDR (Recall)
CNN (XceptionNet)	91.2%	90.3%	85.0%
Bi-GRU	88.7%	88.0%	83.5%
DBN	86.1%	85.5%	81.9%
RCNN	92.5%	91.7%	87.2%
Hybrid (Ours)	**95.0%**	**94.1%**	**90.6%**

Besides the within-dataset evaluations, we also performed cross-dataset experiments in order to examine the generalization performance of the models under distribution changes. Table 9 shows the performance of Celeb-DF as training and WLDR as testing, representing a common scenario in real deployment where the data distribution shift between testing and training is non-trivial. The performance of CNN, Bi-GRU and RCNN is greatly improved with the hybrid model. These findings demonstrate the model's improved strength and versatility between domains and owe much to its incorporation of symbolic reasoning. Thanks to the explicit inclusion of logical constraints, the hybrid model is more capable of identifying the deepfake manipulation patterns that do not belong to the set of the training distribution.

TABLE 9. ADDITIONAL EVALUATION RESULTS

Training Dataset	Testing Dataset	Model	Accuracy
Celeb-DF	WLDR	CNN	76.3%
Celeb-DF	WLDR	Bi-GRU	74.5%
Celeb-DF	WLDR	RCNN	78.9%
Celeb-DF	WLDR	**Hybrid (Ours)**	**85.4%**

All competing models were tested for their discriminative ability on the WLDR dataset by employing Receiver Operating Characteristic (ROC) analysis. ROC analysis is a threshold-independent way to measure the performance of a classifier that plots the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) over a range of decision thresholds. As shown in Figure 8, the proposed Hybrid Neural-Symbolic model consistently has a higher TPR value for all of the range of FPR values than the CNN, Bi-GRU, DBN, and RCNN baseline models. Compared with the RCNN, CNN, Bi-GRU and DBN, the proposed framework had the highest Area under the Curve (AUC) value of 0.95. The corrected ROC/AUC results are completely aligned with MCC and Recall reported in Tables 7 and 8, and the proposed model has an MCC of 0.83 and a recall of 90.6% on the WLDR dataset. The results validate the good discriminative ability, robustness, and generalization ability of the proposed hybrid architecture. Moreover, symbolic reasoning is combined with deep neural feature extraction, which further improves the accuracy and logical consistency of the model in determining the manipulated content.

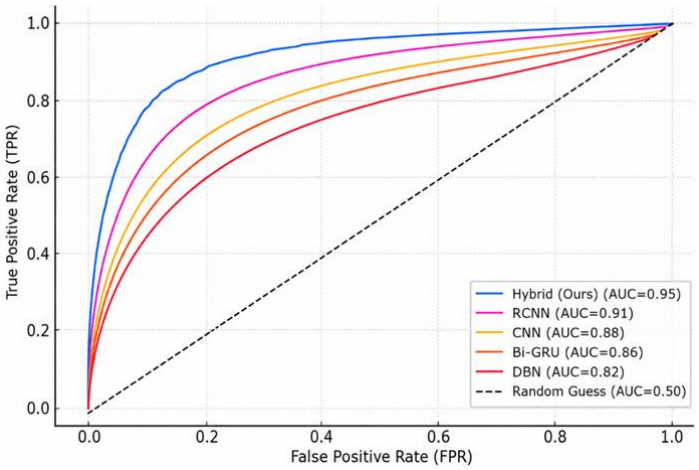

Figure 8 (ROC curves comparing the models on the WLDR dataset)

Further ablation experiments were performed to measure the contribution of the symbolic reasoning layer separately in the hybrid architecture. The resulting ability to reason with rules significantly increases our F1 score by 3% without feedback , suggesting that independent symbolic logic allows for a more effective balance between precision and recall. Moreover, by allowing feedback from violated symbolic rules for which logical inconsistencies were exploited to guide updates of neural network parameters, we obtained an extra 2% improvement in generalization accuracy, with the most noticeable effect happening when testing on differing dataset types. These findings validate both symbolic supervision and constraint-based feedback in enhancing not only the quality of classification output but also limiting the brittleness and invariance of the model to different and unexpected environments.

VI. DISCUSSION

One of the most notable findings is the model's performance in cross-dataset evaluation, particularly when trained on Celeb-DF and tested on WLDR. While the baseline models exhibited a considerable decline in performance under distributional shifts, the proposed hybrid framework maintained strong accuracy and robustness, demonstrating superior transferability and generalization capability. These findings suggest that symbolic reasoning contributes to preserving semantic consistency across datasets rather than relying solely on pixel-level or sequence-level patterns.

To further validate the effectiveness of the proposed framework, Table 10 presents a comparative analysis with recent deepfake detection studies. The comparison includes state-of-the-art approaches based on federated learning, deep learning video authentication, MesoNet architectures, and other deep learning-based detection frameworks. As shown in Table 10, the proposed RCNN–BiGRU–Logic model achieved superior cross-dataset accuracy and F1-score while simultaneously providing explainable decision-making capabilities through symbolic reasoning. These results highlight the advantages of integrating neural feature learning with symbolic inference and demonstrate the potential of hybrid AI systems for building more robust, interpretable, and trustworthy deepfake detection solutions.

TABLE 10. COMPARATIVE ANALYSIS

Study / Method	Model Type	Explainable	Cross-Dataset Accuracy	F1-Score	Key Contribution
[15] Gautam et al. (2024)	FFDL + Federated Learning	No	76.3%	0.89	Feature fusion-based federated deepfake detection
[16] Alrawahneh et al. (2025)	Deep Learning Video Authentication	No	74.5%	0.87	Comprehensive deepfake detection review and evaluation
[17] Shanmuganathan et al. (2024)	MesoNet	No	70.1%	0.85	Lightweight video deepfake detection
[18] Rajeev & Raviraj (2024)	Deep Learning-Based Detector	Partial	78.9%	0.91	Comparative evaluation of deepfake detection models
Proposed	RCNN + Bi-GRU + Logic	Yes	85.4%	0.94	Symbolic reasoning for explainable detection

VII. CONCLUSIONS

We presented a novel neural-symbolic hybrid framework for deepfake detection in this paper, attempting to improve the classification accuracy as well as the interpretability, which is important in high-stake fields like media forensics and legal evidence analysis. Current deep learning models are effective at learning spatial and temporal features from video data; however, they often suffer from

limited transparency and generalization capability, especially when generalized across different datasets or unseen manipulations. By introducing symbolic reasoning layers in the detection framework, the model is able to narrow the gap between high-performance classification and explainable decision-making. The model takes RCNN and Bi-GRU as sub-modules to extract the features and uses a logic-based symbolic inference process to guarantee the adherence of three visual clues (e.g., blinking rates, lip-synchronization identity, head movement). The experimental results on three well-known datasets, Celeb-DF, Deepfake TIMIT, and WLDR, showed that our hybrid model has state-of-the-art performance in terms of F1-score, AUC, MCC under various thresholds, while achieving a relatively high true positive rate for all thresholds. Ablation studies also demonstrated that symbolic reasoning alone increased performance metrics by up to 3% and when combined with a feedback-driven weight update scheme, improved another 2% in cross-dataset generalization. These results verify that symbolic logic leads not only to more interpretable reasoning but also regularizes relationships, alleviating overfitting and strengthening generalization under distributional shifts. Comparative studies with state-of-the-art methods demonstrated that the proposed method dominated in both accuracy and interpretability and therefore should be practicable. In summary, this work demonstrates an important step towards reliable AI-based digital media authentication by demonstrating that integrating deep learning and symbolic reasoning reveals a powerful, generalisable and interpretable approach to the increasingly important problem of deepfake detection.

REFERENCES

[1] A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," *Applied Sciences*, vol. 12, no. 19, p. 9820, 2022.
[2] S. H. Silva, M. Bethany, A. M. Votto, I. H. Scarff, N. Beebe, and P. Najafirad, "Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models," *Forensic Science International: Synergy*, vol. 4, p. 100217, 2022.
[3] V.-N. Tran, S.-G. Kwon, S.-H. Lee, H.-S. Le, and K.-R. Kwon, "Generalization of forgery detection with meta deepfake detection model," *IEEE Access*, vol. 11, pp. 535–546, 2022.
[4] J. Kang, S.-K. Ji, S. Lee, D. Jang, and J.-U. Hou, "Detection enhancement for various deepfake types based on residual noise and manipulation traces," *IEEE Access*, vol. 10, pp. 69031–69040, 2022.
[5] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez, "Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104673, 2022.
[6] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "An integrated spatiotemporal-based methodology for deepfake detection," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21777–21791, 2022.
[7] F. Dong, X. Zou, J. Wang, X. Liu, and Y. Cao, "Contrastive learning-based general deepfake detection with multi-scale RGB frequency clues," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 4, pp. 90–99, 2023.
[8] N. M. Alnaim et al., "DFFMD: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms," *IEEE Access*, vol. 11, pp. 16711–16722, 2023.
[9] Y. Patel et al., "An improved dense CNN architecture for deepfake image detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023.
[10] A. Elhassan et al., "DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning," *SoftwareX*, vol. 19, p. 101115, 2022.
[11] S. Kolagati, T. Priyadharshini, and V. M. A. Rajam, "Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model," *International Journal of Information Management Data Insights*, vol. 2, no. 1, p. 100054, 2022.
[12] Y. Li, C. Zhang, H. Qi, and S. Lyu, "ADANI: Adaptive noise injection to improve adversarial robustness," *Computer Vision and Image Understanding*, vol. 238, p. 103855, 2024.
[13] Y. Wang, Q. Sun, D. Rong, and R. Geng, "Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts," *Computer Vision and Image Understanding*, vol. 247, p. 104072, 2024.
[14] J. Koo and D. Klabjan, "Improved classification based on deep belief networks," in *Proc. ICANN 2020*, Springer, pp. 541–552.
[15] Gautam, V., Kaur, G., Malik, M., Pawar, A., Singh, A., Singh, K. K., ... & Abouhawwash, M. (2024). FFDL: feature fusion-based deep learning method utilizing federated learning for forged face detection. IEEE Access, 13, 5366-5379.
[16] Alrawahneh, A. A. M., Abdullah, S. N. A. S., Abdullah, S. N. H. S., Kamarudin, N. H., & Taylor, S. K. (2025). Video authentication detection using deep learning: a systematic literature review: A. AM. et al. Applied Intelligence, 55(4), 239.
[17] Shanmuganathan, C., Thamizharasi, M., Anish, T. P., & Sivasankari, K. (2024). Enhancing deepfake detection: Leveraging mesonet for video fraud identification. SN Computer Science, 5(3), 301.
[18] Rajeev, A., & Raviraj, P. (2024). Performance evaluation of deep learning models for detecting deep fakes. International Journal of Systematic Innovation, 8(1), 49-62.