2024 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3503-6413-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCB62174.2024.10744498

RuleBoost: A Neuro-Symbolic Framework for Robust Deepfake Detection

Muhammad Anas Raza
Oakland University
Michigan, United States
mraza@oakland.edu

Khalid Mahmood Malik
University of Michigan-Flint
Michigan, United States
drmalik@umich.edu

Ijaz Ul Haq
University of Michigan-Flint
Michigan, United States
ijazhaq@umich.edu

Abstract

The proliferation of user-friendly deepfake creation tools poses a serious challenge, demanding robust and adaptable detection strategies. Existing approaches primarily focus on raw data analysis or identifying learned artifacts or manual data-driven rules resulting in the mis-classification of deepfakes with distorted facial poses. These architectures also neglect the potential power of combining learned visual features with explicit rules.

To address this gap, we introduce RuleBoost, a novel NeuroSymbolic AI based framework that seamlessly fuses extracted visual features with automatically learned rules. Our framework employs a scalable rule-based learning approach to extract learned rules from facial geometry such as distance, area, and angle. The extracted rules integrated with deep visual features show promising results giving state-of-the-art area-under-the-curve of 96.19% and 95.44% on WLDR and FaceForensics++ Datasets respectively, surpassing other deep learning specific methods. To figure out the difference NeuroSymbolic approach makes, we also analyze the samples misclassified by traditional DL-based architectures and correctly classified by RuleBoosted architecture. Based on empirical evidence, we conclude that DL-based architectures struggle to accurately detect real and fake samples when facial artifacts lead to poses that deviate from standard facial positioning, while RuleBoost exhibits improved performance in the same scenario.

1. Introduction

The proliferation of sophisticated generative algorithms, including GANs (Generative Adversarial Networks) and diffusion models, has spawned the unsettling reality of synthetic media, particularly deepfakes. These manipulated images, videos, and audio recordings pose a potent threat, wielded by malicious actors to spread misinformation and sow discord [27, 17, 1]. Open-source tools like "DeepFaceLab" and "FaceSwap" empower anyone, regardless of machine learning expertise, to fabricate deceptively convincing forgeries. This widespread accessibility, coupled with the rapid advancement of generative algorithms, creates a pressing need for robust deepfake detection techniques.

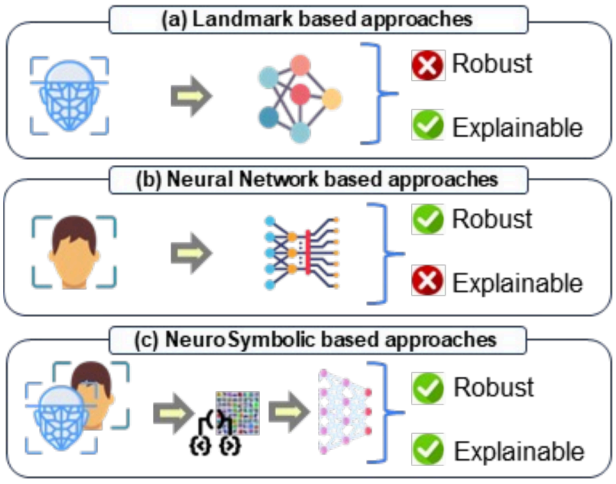

Figure 1. Limitation of the existing deepfake detection techniques. (a) Landmark based approaches are explainable in nature but less robust where extracted landmark are directly fed into various classifiers, (b) Deep Learning or Neural Network based approaches may be robust on given dataset but are neither explainable nor generalizable, and (c) NeuroSymbolic approaches can boost the performance of DL architectures with rules to bring interpretability and improved generalizability(proposed).

The research community has actively devised techniques to counteract the threat of media fabrication and the disinformation spread [26]. Previous efforts employed handcrafted features to identify telltale signs in early deepfakes, such as headpose discrepancies [46], unnatural blinking [13], and warping artifacts [21]. However, the emergence of sophisticated generative models like image-video diffusion models demands more nuanced detection approaches.

As in Figure 1, current deepfake detection algorithms can be broadly classified into three categories: a) landmark-based methods: where landmarks are extracted from

Figure 2. Frames in Red Box Represent Distorted Facial Postures that Appear In Real and Manipulated Facial Frames where Facial Artefacts are Less Visible. e.g. eyes closed, inconsistent artifacts. Deep Learning Models tend to misclassify such samples. The distorted facial data samples are collected from dataset using facial landmarks.

facial regions or various regions and fed to the classifier [37, 22], b) Neural Network-based methods: where facial images are directly fed into the classifiers [48] and, c) Neuro-symbolic approaches: where the advantages of neural networks and reasoning are combined. While promising, existing deepfake detection methods often rely solely on neural networks, failing to capitalize on the complementary strengths of explicit knowledge. Neural networks excel at pattern recognition and capturing complex relationships within data, but they can struggle to encode domain-specific knowledge or handle aspects that are not obvious due to the low volume of the data. Conversely, well-defined rules derived from hand-crafted features can efficiently capture specific visual cues or inconsistencies indicative of deepfakes. By combining these approaches in a neurosymbolic framework, we hypothesize that a deepfake detection system can achieve a more comprehensive and robust representation, leading to significantly improved performance. This integrated approach leverages the strengths of both neural networks and symbolic reasoning, promoting trustworthiness and explainability by explicitly incorporating expert knowledge.

Our analysis reveals a critical limitation of current deep-learning architecture as they struggle to classify deepfakes when presented with images containing unusual or distorted facial poses sampled in Figure 2. This is because deep learning models are trained on massive datasets of images, but these datasets may not encompass the full range of human facial expressions and poses. As a result, the models may not have learned the necessary features to accurately distinguish between genuine facial movements and those manipulated in a deepfake. This is particularly problematic for deepfakes that leverage advanced techniques to create realistic facial expressions on individuals who may never have made them in real life. In such cases, the lack of training data on these extreme poses can lead to misclassifications by the deep learning model.

To address this challenge, we propose RuleBoost, a novel framework that bridges the gap between the strengths of deep learning and the power of rule-based knowledge. RuleBoost seamlessly integrates a set of pre-defined rules specifically designed to identify visual inconsistencies indicative of deepfakes, particularly those associated with unusual facial poses. For extracting rules, we develop a facial vocabulary for deepfake detection which helps us in explainability. The rules resulting from vocabulary can potentially help capture specific patterns, such as unnatural skin texture, inconsistencies in lighting and shadows across facial features, or illogical eye movements, that may be difficult for deep-learning models to detect independently. By incorporating these rule-based checks alongside the deep learning analysis, RuleBoost enhances the overall detection accuracy and robustness. Extensive experimentation across diverse deep learning architectures, including convolutional neural networks (CNNs), Transformer-based architectures, and patch-based methods, demonstrates that RuleBoost consistently outperforms existing methods. This performance improvement is particularly pronounced when presented with deepfakes that leverage unconventional facial poses, highlighting the effectiveness of RuleBoost in addressing this critical limitation. To check the performance of the proposed architectures, we perform experiments on the World Leaders Dataset [4] and FaceForensics++[32]. The proposed boosting architecture performs state-of-the-art Area-Under-The-Curve of 96.19%

on WLDR and 95.44% on FaceForensics++ datasets. We compare the performance of the proposed framework with various models including rules-only, vision-only models as well as models with rule boosting. The overall contributions of this paper are summarized as:

- We proposed a novel NeuroSymbolic framework that combines perception (deep featuers) and rules (knowledge) for robust deepfake detection with extended interpretability.
- We developed a novel set of geometrical features along with a facial vocabulary, which served as the basis for generating rules.
- We conducted extensive experiments to assess the robustness of our proposed framework, including cross-dataset evaluations and an ablation study.

2. Related Work

In this section, we explore landmark-based detection methods, leveraging facial landmarks to expose inconsistencies in deepfakes along with deep learning-based solutions, utilizing neural networks to detect subtle anomalies indicative of manipulated media. Additionally, we investigate rule-based and multimodal approaches, aiming to enhance deepfake detection through the integration of diverse modalities and manual rule formulations. These complementary strategies offer distinct perspectives, potentially leading to more robust and comprehensive detection capabilities. This section also includes a focused subsection on rule extraction methods, which aim to derive explainable rules directly from complex models or through the careful analysis of deepfake artifacts.

2.1. Deepfake Detection

Landmark Based Detection. Facial landmarks offer a potent tool in the fight against deepfakes. These key points, pinpointing features like the eyes, nose, and mouth, reveal subtle inconsistencies and unnatural distortions that frequently betray manipulated videos. Deepfake detection algorithms analyze the relationships between facial landmarks, examining their positions, movements, and relative distances. Disruptions to these expected patterns can signal a deepfake, as human faces adhere to specific proportions and dynamics. By focusing on these vulnerable areas, facial landmark analysis helps expose the artificial nature of deepfakes, providing a valuable defense against these deceptive creations. [43, 36, 47] use facial landmarks for detecting deepfakes.

Deep Learning Based Solutions Deep learning-based solutions are at the forefront of deepfake detection for different modalities offering sophisticated methods for combating these manipulated creations like [15, 31], . These solutions often employ neural networks like Convolutional Neural Networks (CNNs) that excel at analyzing visual data. By training on large datasets of both real and deepfake images/videos, these models implicitly learn anomalies, inconsistencies, and artifacts that betray the artificial nature of deepfakes. [14], unusual facial textures [45], or inconsistencies in lighting and reflections [44].

Rule Based Solutions Several solutions have been proposed for creating manual rules for detecting deepfakes [3, 2]. These rule-based approaches often focus on identifying specific artifacts or inconsistencies that are commonly introduced during the deepfake generation process.

Multimodal Solutions Deep learning's ability to continuously learn and adapt makes it especially powerful against the ever-evolving landscape of deepfake creation techniques. There is also an increased interest in using multimodal solutions for detecting deepfakes [30, 12].

Neural + Symbolic Solutions The current literature lacks an exploration of NeuroSymbolic approaches for detecting deepfakes. Due to the current limitations of deep learning-based architectures, there has been increasing interest in NeuroSymbolic approaches for detecting deepfakes. [9] explore the possibility of using audio and visual modalities along with manually formulated rules for detecting deepfakes. However, this preliminary work does not cover the data-driven knowledge/rule extraction and its integration with deep-learning approach.

2.2. Rule Extraction

Rule-based models (decision trees, etc.) are difficult to train due to their discrete structure. Heuristics [29] and search algorithms are used, but may not find optimal solutions and can be computationally expensive on large datasets. *Bayesian methods* [20] improve structure learning, but scalability and achieving performance comparable to complex models remain challenges. *Ensemble models* (e.g., Random Forests, [5]) outperform rule-based ones, but their decision-making process can be opaque, hindering interpretability [10]. Attempts to bridge this gap exist, but often sacrifice accuracy. *Gradient-based methods* for discrete model training (like STE, [7]) are used in neural network compression, but can have limitations such as requiring gradient information at discrete points. For extracting rules, we employ [42] that employs the *Gradient Grafting* method, which aims to address these issues by utilizing both discrete and continuous model gradients.

3. Proposed Framework

The proposed framework comprises two distinct networks. One network focuses on the extraction and projection of image patches via a backbone network for deep facial feature extraction. Meanwhile, the other network is dedicated to analyzing facial landmarks-based geometry,

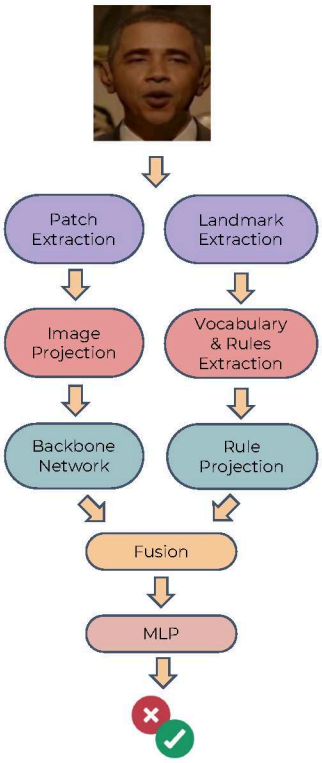

Figure 3. The proposed framework infuses rules and deep learning based features and provides prediction using an MLP head.

which are utilized for generating rules. The image features and rule projections are integrated and finally passed through a classification head to determine the authenticity of a given video. Figure 6 provides a visual representation of the overall framework.

The proposed rule extraction and fusion framework can be employed in other domains as well where relevant features can be extracted. An example of such an application can be audio [16] where we can extract various hand-crafted features and fuse them in DL-based architecture as done in the next steps.

3.1. Problem Formulation

Existing performant deepfake detectors extract deep features from facial crops of images in a binary classification problem. In this paper, we focus on learning a rule knowledge base that improves the performance of deep learning model by including relevant facial information. Let assume a given video X = $\{x_i\}$, where x_i denotes video frames with face, $\mathcal{G}_\theta(X)$, is the proposed framework to make prediction which includes three parts; the image projector and backbone network $\mathcal{F}_{\phi i}$ that maps input video frame to a feature space of $\mathbb{R}^{T_i \times d}$ where T_i and d are the number of image patches and feature dimension. While $\mathcal{F}_{\phi r}$ maps the input image to a rule representation of size $\mathbb{R}^{(T_i+1)\times d}$ where d

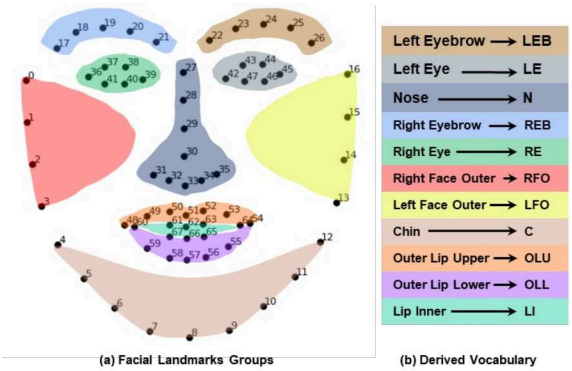

Figure 4. Facial vocabulary generation for rules extraction. (a) grouped landmarks for each facial part and (b) corresponding derived vocabulary for each facial part (same color).

being feature dimension with the help of developed facial vocabulary. Finally, $\mathcal{F}_{\psi ir}$ fuses the image and rule representations into $\mathbb{R}^{(T_i+1)\times d}$ feature vector and fed it into classification head C for prediction.

3.2. Image Feature Projection

We extract image features, project them in a space and do multi-layered processing using a backbone network.

Patch Extraction and Image Projection: We perform patch extraction as a fundamental step in Vision Transformer (ViT) models by dividing an input image x_i into a grid of smaller, fixed-size patches. Each patch is then flattened and fed into the encoder. This allows the network to capture spatial information without relying on convolutional layers, making it computationally efficient and effective for tasks like image classification. The extracted patches are projected using a multi-layer perceptron layer. We also add rotary positional embedding to consider the location of individual image patches.

Backbone Network: To extract learned features from images, we employ a diverse range of architectures designed for effective image representation. These include mixer [38], attention-based models with token learners [33], gMLP [8], External Attention [8], FourierNet [19], Compact Convolutional Transformers [11], and ConvMixer [40]. This selection encompasses techniques that excel in capturing global context, modeling long-range dependencies, and efficiently combining convolutional and self-attention mechanisms. Additionally, we incorporate rule-based elements to complement the feature extraction process.

3.3. Facial Geometry-based Rules Extraction and Validation

We extract learned rules using derived features from facial geometry. Derived features could potentially provide richer information than basic landmarks. Analyzing facial

geometry involves precise calculations of specific information including the distance between the pupils, the size of the mouth, or the slope of the eyebrows etc.

Landmarks and Derived Geometric Feature Extraction: First, we extract facial landmarks, using dlib [18] employing specialized machine learning models to pinpoint 64 landmark points including the nose tip, eye corners, and lip. Based on various deepfake types such as face-swap or Lip-Sync etc., the extracted landmarks are grouped into various facial parts to develop human understandable vocabulary as illustrated in Figure 4. Next, we derived a robust feature-set, by analyzing the relative distances, angles, and areas formed by these landmarks which provide valuable information about the spatial relationships among facial features, helping to differentiate between genuine and manipulated faces. The derived facial geometrical features including distance d, angle An and and area Ar are calculated as:

$$\mathrm{d} = \sqrt{(\mathrm{x}_1 - \mathrm{x}_2)^2 + (\mathrm{y}_1 - \mathrm{y}_2)^2}$$

$$An = tan^{-1}|(m_1 - m_2)/(1 + m_1 m_2)|$$

$$Ar = \sqrt{s\,(s - d_a)\,(s - d_b)\,(s - d_c)}$$

where, x, m, and s are the landmark point, line slope, and semi-perimeter of a triangle to calculate distance d, angle An and area Ar.

Rules Extraction using RRL We employ Rule-based Representation Learner (RRL) [42], a classifier designed to learn interpretable, nonfuzzy rules for representation and classification automatically. The derived geometrical features symbolized with developed vocabulary are given to RRL along with fake and real labels to generates rules. RRL employs Gradient Grafting, a novel training method, to optimize the discrete model directly. The Logical layers in RRL learn complex logical rules based on underlying data representations. Various combination of binarization and logical layers act as feature learner that fed linear layer to perform classification. The RRL itself filtered out unnecessary logical rules using skip-connections in logical layers. Further technical details of RRL can be found in [42]. The learned rules extracted from facial landmarks are also fed to the fully connected network to extract learned deep features of the same dimensionality as the dimension of features extracted from image patches.

3.4. Fusion and Classification Head

The extracted image features and projected rule embeddings are concatenated together in the joint space. We also employ normalization on the concatenated features to have them on same scale. To process images and rules together in a joint space, the image and rule features obtained after fusion are fed to an MLP Head from where we get binary classification output. The final logits are given by a fully connected network. Further, we apply sigmoid function to the logits for obtaining confidence score for deepfake detection.

Model Name	Type	WLDR	FF++
CCT [11]	✠	81.23	77.23
	®	86.33	83.12
CaIT [39]	✠	76.32	84.17
	®	84.72	90.17
gMLP [23]	✠	84.24	87.32
	®	93.39	**95.44**
EANet [8]	✠	90.31	83.33
	®	92.65	87.43
FNet [19]	✠	86.33	92.73
	®	90.32	87.37
ConvMixer [40]	✠	81.37	92.55
	®	83.38	91.48
Token Learner [33]	✠	82.81	82.46
	®	95.84	94.38
MLPMixer [38]	✠	92.35	85.64
	®	**96.19**	93.26

Table 1. Results on WLDR and FaceForensics++ Datasets ✠: models tested without rules, ®: Ruleboosted models. Evaluation matrix is AUC(%).

4. Experiments and Results

This section details the experimental setup, datasets, and results used to analyze the RuleBoost model's performance. Analysis includes comparisons with state-of-the-art techniques and evaluation across diverse datasets.

4.1. Datasets and Metrics

In our experiments, we utilize both the FaceForensics++ (FF++) [32] and World Leaders Dataset (WLDR) [4] datasets to ensure a comprehensive evaluation. For cross-dataset evaluation, we employ Presidential Deepfakes Dataset (PresDD) [34]. The employed datasets offer a diverse range of manipulation techniques, allowing us to test the robustness of our detection models across different scenarios. We selected distorted facial poses from the existing WLDR Dataset. While we did not construct a separate dataset focused solely on distorted poses, the WLDR Dataset does include variation in facial expressions and poses, allowing us to partially assess our method's performance on this challenging subset.

We rely on the Area Under the Curve (AUC) and Accuracy (ACC) metrics for the assessment of our models on full dataset and cross-dataset evaluation, respectively.

Method	DF	F2F	FS	NT
C3D [24]	92.86	88.57	91.79	89.64
XceptionNet-avg [32]	98.93	98.93	99.64	95.00
I3D [6]	92.86	92.86	96.43	90.36
TEI [25]	97.86	97.14	97.50	94.29
FaceNetLSTM [35]	89.00	87.00	90.00	-
DeepRhythm [28]	98.70	98.90	97.80	-
Comotion-35 [41]	95.95	85.35	93.60	88.25
HolisticDFD [31]	98.00	95.00	94.00	96.00
Ruleboost (proposed)	97.00	96.20	95.40	97.80

Table 2. Performance analysis on different subsets of FaceForensics++ c23, DF: Deepfake, F2F: Face2Face, FS: FaceSwap, NT:Neural Texture. Evaluation matrix is ACC(%).

Testing Subset	MT_ml [30]	AVFNet[12]	RB
W - LipSync	83.33	69.32	**84.39**
W - FaceSwap	75.84	73.98	**76.82**
W - Imposter	**91.66**	61.74	84.28
P - full	70.00	78.12	**83.62**

Table 3. Cross-dataset Evaluation: RB: Ruleboost architecture trained on FaceForensics++, W: World Leaders Dataset, P: Presidential Deepfake Detection Dataset. Evaluation matrix is ACC(%).

4.2. Performance Analysis

The results in Table 1 summarizes the performance of several deep learning models on two image classification tasks: WLDR and FF++. The results demonstrate that ruleboosted models generally outperform their standard counterparts. For the WLDR dataset, MLP Mixer® achieved the highest AUC of 96.19%, followed closely by Token Learner® at 95.84%. This significant improvement for Token Learner highlights the effectiveness of combining rule-based systems with deep learning approaches. On the FF++ dataset, gMLP® reigned supreme with an AUC of 95.44%, while CaIT® placed second at 90.17%. It's worth noting that CaIT✠ underperformed relative to other models on this dataset, suggesting that rule-based enhancements can be particularly beneficial for certain model architectures or tasks.

Model	$d + An$	$d + Ar$	$Ar + An$	$d + An + Ar$
SVM	67.40	72.2	80.90	81.30
LR	77.90	78.3	74.50	76.20
DT	86.2	86.6	88.10	89.40
Mixer	95.30	93.20	94.30	**96.19**

Table 4. Ablation Study of Rules with *SVM*: Support Vector Machine, *LR*: Logistic Regression, *DT*: Decision Trees, *NB*: Naive Bayes. *d*: Distance-based rules, *An*: Angle-based rules, *Ar*: Area-based rules. Evaluaiton matrix is ACC(%)

Furthermore, the table reveals interesting trends across the different models. For example, ConvMixer demonstrates a larger performance boost from rule augmentation on the WLDR dataset (81.37% to 83.38%) compared to the FF++ dataset (92.55% to 91.48%). This suggests that the specific type of image data and the model's inherent strengths may influence how much it benefits from rule-based guidance. Overall, the results highlight the potential of combining deep learning with rule-based systems to achieve superior performance on image classification tasks. We also evaluate the performant Ruleboosted architecture on the subset of FaceForensics++ dataset as shown in Table 2. The evaluated methods include both video-based architectures (C3D, I3D, TEI) and those utilizing additional modalities (FaceNetLSTM, DeepRhythm). The proposed Ruleboost method demonstrates competitive accuracy across all four deepfake manipulation types (Deepfake, Face2Face, FaceSwap, Neural Texture).

4.3. Cross Dataset Evaluation

We also perform cross-dataset evaluation as in Table 3 shows the results of a cross-dataset evaluation where the Ruleboost (RB) architecture, trained on FaceForensics++, is tested on different image/video manipulation detection tasks. It's compared against two other models: MT_ml and AVFNet. The results indicate that RB generally outperforms the other models across most of the testing subsets. On "WLD-FaceSwap", it achieves the highest accuracy of 76.82%, and it has a similarly strong performance on "WLD-LipSync" with 84.39% accuracy. While MT_ml gets the top score on "WLD-Imposter", RB still demonstrates a respectable accuracy. Notably, RB delivers the most substantial improvement on the "PDD-full" subset, reaching 83.62% accuracy. These results suggest that the Ruleboost architecture possesses valuable generalization abilities, making it more robust for detecting different types of image and video manipulations, even those unseen during its initial training.

4.4. Ablation Study

To perform an ablation study of the proposed ruleboosting approach, we also evaluate the performance of various models (SVM, Logistic Regression, Decision Trees, and our proposed Mixer model) using different combinations of rules: distance-based rules (d), angle-based rules (An), and area-based rules (Ar). Each column represents a different combination of these rules applied to the models. For this experiments, we used WLDR dataset and the results are shown in Table 4. The results indicate that the performance varies depending on the combination of rules used. For instance, in the case of the SVM model, using all three types of rules ($d + An + Ar$) yields the highest accuracy of 81.30%, while the combination of distance and angle-based

Support	Rule
0.8081	1) ***An***($N_{35}LFO_{15}LFO_{14}$) > 58.14 & ***An***($N_{28}RE_{39}N_{29}$) < 22.09 & ***An***($C_4OLU_{48}C_5$) < 43.28 & ***An***($OLL_{59}C_5C_6$) < 72.91
0.7595	2) ***d***($N_{35}C_8$) > 61.34 & ***An***($LEB_{22}REB_{21}N_{27}$) > 33.64 & ***Ar***($REB_{21}LEB_{22}N_{27}$) < 306.24 & ***Ar***($N_{31}OLU_{48}C_4$) < 524.44
0.8148	1) ***d***($REB_{19}RE_{37}$) > 15.28 & d($LE_{42}N_{27}$) > 101.79 & ***d***($N_{31}RE_{40}$) > 22.23 & ***An***($N_{31}RFO_2RFO_3$) > 69.07 & ***An***($RFO_3N_{31}C_4$) > 16.40 & ***d***(C_6OLL_{58}) < 55.37 & ***Ar***($RE_{39}REB_{21}N_{27}$) < 174.43 & ***An***($N_{29}LE_{42}N_{30}$) < 14.40 & ***An***($RE_{36}REB_{18}RE_{37}$) < 30.614
0.709	2) ***An***($RFO_1RE_{41}N_{31}$) > 82.84 & ***An***($N_{31}RFO_2RFO_3$) < 84.36

Figure 5. Example Performant data-driven rules and their support score.

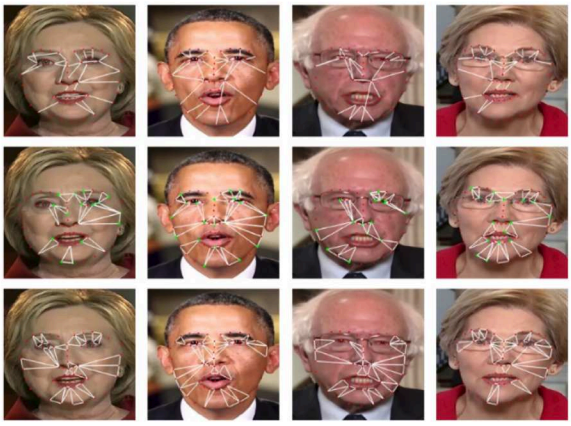

Figure 6. A visual representation of extracted rules of real class over political identities from WLDR dataset. top row: distance-based rules, middle row: angle-based rules, bottom row: area-based rules.

rules ($d+An$) results in the lowest accuracy of 67.40%. Logistic Regression (LR) shows less variation in performance across different rule combinations, with the highest accuracy of 78.3% achieved using distance and area-based rules ($d + Ar$). Decision Trees (DT) exhibit improved performance with more comprehensive rule sets, achieving the highest accuracy 89.40% when all three rules are combined.

The proposed Ruleboosted Mixer model consistently outperforms the LR, DT and SVM-based models across all rule combinations, with the highest accuracy of 96.20% achieved when all three rules are used together. This demonstrates the robustness and effectiveness of the Mixer model in integrating multiple rule types for enhanced performance.

4.5. Discussion

Our approach focuses on augmenting traditional DL models with rule-based insights, which can provide complementary information that pure DL models might overlook. By incorporating these rules, we aim to bolster the models' ability to identify subtle discrepancies and anomalies that are indicative of deepfake manipulation, particularly in scenarios where facial images have been intentionally distorted to evade detection mechanisms. This integration of rule-based systems with DL models represents a hybrid methodology that leverages the strengths of both approaches, potentially leading to more robust and accurate detection systems.

To further explain the generated rules, let take example of rules shown in Figure 5 and their corresponding vocabulary illustrated in Figure 4. The specified rules check the facial geometry features formed by landmarks for specific individual satisfy certain threshold. The landmark vocabulary (color codes) shows a mapping between facial landmark regions (e.g., eyebrows, eyes, nose, lips) and their corresponding landmark indices/vocabulary. For example, the generated rule 1 from Figure 5 involves checking specific angles between facial landmarks (e.g., nose, face-outer region, eyes, lips) such as $An(N_{35}LFO_{15}LFO_{14}) > 58.14$ means N_{35} represents a landmark on the nose, and LFO_{15} and LFO_{14} are landmarks on the left face's outer region. The rule states that the angle between $N_{35}LFO_{15}$ and $LFO_{15}LFO_{14}$ must be greater than 58.14 degrees for real face. Further, $An(N_{30}RE_{39}N_{29})$ states that angle formed by landmarks N_{28}, RE_{39}, and N_{29} must be less than 22.09 degrees and "$An(J_4OLU_{48}J_5)$" specifies the angle formed by landmarks J_4, OLU_{48}, and J_5 must be less than 43.28 degrees, "$An(OLL_{59}J_5J_6)$" specifies the angle formed by landmarks OLL_{59}, J_5, and J_6 must be less than 72.91 degrees. In Figure 6, the generated rules for real class are ploted for better understanding where a specific region in a face for an individual is clearly highlighted to contribute in rules generation. Moreover, the provided rule 2 consists of four conditions related to facial distances, angles, and areas, such as "$d(N_{35}J_8 > 61.342)$" denotes that the distance between landmarks N_{35} and J_8 must be

	CCT [11]	CaIT [39]	gMLP [23]	EANet [8]	FNet [19]	ConvMixer [40]	TL [33]	MLPMixer [38]
✠	54.50	63.20	74.90	69.40	56.80	68.20	67.40	85.40
®	87.40	83.40	86.50	83.30	79.60	89.20	90.40	**94.30**

Table 5. Performance Analysis on Distorted Facial Poses. ✠: models tested without rules, ®: Ruleboosted models

greater than 61.342, $An(LEB_{22}REB_{21}N_{27}) > 33.645$ specifies that the angle formed by landmarks LEB_{22}, REB_{21}, and N_{27} must be greater than 33.64 degrees, $Ar(REB_{21}LEB_{22}N_{27}) < 306.243$ denotes the area formed by landmarks REB_{21}, LEB_{22}, and N_{27} must be less than 306.243 and $Ar(N_{31}OLU_{48}J_4 < 524.445$" indicates the area formed by landmarks N_{31}, OLU_{48}, and J_4 must be less than 524.44. Similarly, the other rules can be used to explain the reason for the model decision for a specific individual. We select rules based on their effectiveness as depicted by RRL [42] as using all rules in the framework would be inefficient. Based on empirical analysis, we only select rules whose support is greater than 0.7 as shown in Figure 5.

Our experimental findings, summarized in Table 5, demonstrate the potential of incorporating rules into deep learning (DL) architectures for the detection of deepfake images, particularly those with facial distortions. The Ruleboosted models, denoted by ®, consistently outperform their rule-free counterparts, represented by ✠, across all the tested architectures. This consistent performance boost underscores the ability of Ruleboost to significantly enhance the predictive accuracy of DL models when tackling the challenging problem of deepfake detection.
A particularly noteworthy observation is the substantial improvement seen with the MLPMixer architecture. When Ruleboost is applied, MLPMixer achieves an impressive overall accuracy of 94.30%, the highest among all evaluated models. This remarkable performance suggests a strong compatibility between the MLPMixer's unique architecture and the integration of data-driven rules. It highlights the potential synergy between this type of DL architecture and the Ruleboost methodology. The superiority of Ruleboosted models across various architectures emphasizes the versatility and effectiveness of the proposed architecture for deepfake detection. By integrating specific rules into the DL models, Ruleboost not only enhances accuracy but also provides a more robust framework for handling the intricacies of distorted facial images.

5. Conclusion and Future Works

In this paper, we underscore the inherent limitations of relying solely on deep learning models or data-driven rules for deepfake detection, particularly in scenarios with atypical facial poses. The introduction of RuleBoost offers a significant step forward, demonstrating the superiority of combining learned visual features with explicit rules. This hybrid approach achieves state-of-the-art performance, highlighting the importance of adaptability and multifaceted strategies in combating the evolving threat of deepfakes. Our analysis further indicates that by addressing the blind spots of deep learning models, RuleBoost offers a more robust and reliable framework for deepfake detection.
In future, we plan to extend RuleBoost to multimodal deepfake detection. By incorporating audio and video analysis alongside our current visual approach, we aim to significantly enhance RuleBoost's robustness across diverse deepfake scenarios. Our research will focus on exploring optimal integration strategies for these modalities, potentially including early fusion, late fusion, and hybrid techniques. Additionally, we may investigate the use of attention mechanisms to dynamically focus RuleBoost on the most crucial aspects of the multimodal data, further refining its detection capabilities.

References

[1] Experts warn deepfakes and ai could threaten election integrity. *NBCNews.com*, Mar 2024.

[2] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020.

[3] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020.

[4] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, page 38, 2019.

[5] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[7] M. Courbariaux, Y. Bengio, and J.-P. David. Binaryconnect: Training deep neural networks with binary weights during propagations. *Advances in neural information processing systems*, 28, 2015.

[8] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5436–5447, 2022.

[9] I. U. Haq, K. M. Malik, and K. Muhammad. Multimodal neurosymbolic approach for explainable deepfake detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[10] S. Hara and K. Hayashi. Making tree ensembles interpretable. *arXiv preprint arXiv:1606.05390*, 2016.

[11] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.

[12] H. Ilyas, A. Javed, and K. M. Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing*, page 110124, 2023.

[13] T. Jung, S. Kim, and K. Kim. Deepvision: deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

[14] T. Jung, S. Kim, and K. Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

[15] A. Khan, A. Javed, K. M. Malik, M. A. Raza, J. Ryan, A. K. J. Saudagar, and H. Malik. Toward realigning automatic speaker verification in the era of covid-19. *Sensors*, 22(7):2638, 2022.

[16] A. Khan and K. M. Malik. Securing voice biometrics: One-shot learning approach for audio deepfake detection. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2023.

[17] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*, 56(Suppl 1):513–566, 2023.

[18] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[19] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

[20] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. 2015.

[21] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[22] P. Liang, G. Liu, Z. Xiong, H. Fan, H. Zhu, and X. Zhang. A facial geometry based detection model for face manipulation using cnn-lstm architecture. *Information Sciences*, 633:370–383, 2023.

[23] H. Liu, Z. Dai, D. So, and Q. V. Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021.

[24] J. Liu, K. Zhu, W. Lu, X. Luo, and X. Zhao. A lightweight 3D convolutional neural network for deepfake detection. *Int. J. Intell. Syst.*, 36(9):4990–5004, Sept. 2021.

[25] Z. Liu, D. Luo, Y. Wang, L. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and T. Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11669–11676, 2020.

[26] 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. *Applied Intelligence*, pages 1–53, 2022.

[27] E. C. A. Professor. Deepfakes are still new, but 2024 could be the year they have an impact on elections. *The Conversation*, Mar 2024.

[28] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4318–4327, 2020.

[29] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[30] M. A. Raza and K. M. Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2023.

[31] M. A. Raza, K. M. Malik, and I. U. Haq. Holisticdfd: Infusing spatiotemporal transformer embeddings for deepfake detection. *Information Sciences*, 645:119352, 2023.

[32] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[33] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34:12786–12797, 2021.

[34] A. Sankaranarayanan, M. Groh, R. Picard, and A. Lippman. The presidential deepfakes dataset. In *workshop 'AIofAI: 1st workshop on adverse impacts and collateral effects of artificial intelligence technologies'. http://ceur-ws. org*, volume 2942, 2021.

[35] S. J. Sohrawardi, A. Chintha, B. Thai, S. Seng, A. Hickerson, R. Ptucha, and M. Wright. Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2613–2615, 2019.

[36] D.-C. Stanciu and B. Ionescu. Deepfake video detection with facial features and long-short term memory deep networks. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2021.

[37] Y. Sun, Z. Zhang, I. Echizen, H. H. Nguyen, C. Qiu, and L. Sun. Face forgery detection based on facial region displacement trajectory series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 633–642, 2023.

[38] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

[39] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

[40] A. Trockman and J. Z. Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

[41] G. Wang, J. Zhou, and Y. Wu. Exposing deep-faked videos by anomalous co-motion pattern detection. *arXiv preprint arXiv:2008.04848*, 2020.

[42] Z. Wang, W. Zhang, N. Liu, and J. Wang. Scalable rule-based representation learning for interpretable classification. *Advances in Neural Information Processing Systems*, 34:30479–30491, 2021.

[43] J. Wu, B. Zhang, W. Luo, J. Fan, Z. Teng, and J. Fan. Leveraging facial landmarks improves generalization ability for deepfake detection. *Available at SSRN 4620608*.

[44] W. Wu, W. Zhou, W. Zhang, H. Fang, and N. Yu. Capturing the lighting inconsistency for deepfake detection. In *International Conference on Artificial Intelligence and Security*, pages 637–647. Springer, 2022.

[45] B. Xu, J. Liu, J. Liang, W. Lu, and Y. Zhang. Deepfake videos detection based on texture features. *Computers, Materials & Continua*, 68(1), 2021.

[46] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. *CoRR*, abs/1811.00661, 2018.

[47] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[48] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang. Istvt: interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1335–1348, 2023.