An Ensemble Approach to Facial Deepfake Detection Using Self-Supervised Features

Yan Martins B. Gurevitz Cunha
yangurevitz@telemidia.puc-rio.br
Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro

José Matheus C. Boaro
boaro@telemidia.puc-rio.br
Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro

Daniel de Sousa Moraes
danielmoraes@telemidia.puc-rio.br
Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro

Pedro Cutrim dos Santos
thiagocutrim98@gmail.com
Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro

Polyana Bezerra da Costa
polyanabcosta@gmail.com
Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro

Antonio José Grandson Busson
antonio.busson@btgpactual.com
BTG Pactual

Julio Cesar Duarte
duarte@ime.eb.br
Computing Engineering Dept. – Military Institute of Engineering

Sérgio Colcher
colcher@inf.puc-rio.br
Telemidia Lab. – Pontifical Catholic University of Rio de Janeiro

ABSTRACT

Substantial efforts have been dedicated to developing methods for detecting deepfake content, especially with the creation of large and diverse datasets with both higher image quality and demographic features. In this scenario, CNN-based approaches showed good initial success, later improved by their combination with Vision Transformers. More recently, Foundation Models (FMs) have emerged, improving performance across many visual tasks, including deepfake detection, and combining self-supervised features generated by FMs with CNN-based classifiers has resulted in significant performance gains. However, taking advantage of multiple maps of self-supervised features is not as straightforward as just adding more channels to the classifier. Therefore, this work explores ensemble techniques to effectively utilize these diverse self-supervised feature maps for realistic facial deepfake detection. Our experiments indicate that combining the output results of different classifiers, each one utilizing a single map of self-supervised features, leads to significant performance improvements, and several committee approaches consistently outperform individual classifiers, demonstrating the potential of these methods in enhancing deepfake detection accuracy.

KEYWORDS

deep fake detection, self-supervised, vision transformers, deep learning, foundation models

1 INTRODUCTION

In recent years, there has been an increased focus on the effects of deepfake multimedia content on public discourse and personal lives. This has manifested in forms such as explicit fake images of celebrities [22] or politically sensitive material, facilitating the spread of misinformation and identity theft and possibly leading to threats of violence [14]. This scenario has renewed societal discussions on the nature of online content and led to legislative debates in countries such as the United States on reducing the creation and distribution of this kind of media [3].

Concurrent with the advancement of deepfake generators, significant progress has also been made in efforts to improve the detection and containment of this kind of content. This starts with the creation of large-scale datasets for deepfake detection in video and image formats [10, 21], enabling the development of methods to better differentiate between real and fake content. In this context, CNN-based models showed initial success, and their combination with Vision Transformers (ViT) managed to achieve state-of-art performance in recent years [13].

Also, the recent emergence of Foundation Models (FMs) opened the door for new approaches to deepfake detection. In this light, studies have demonstrated that using self-supervised features generated by pre-trained FMs in combination with CNN-based models can result in significant performance improvements [12] . In this approach, the authors incorporated feature maps generated by FMs as extra channels in CNN input, showing that, when used individually, each map would improve the performance of its base model. However, it has also been shown by Gomes et al. [12] that simultaneously applying multiple attention maps to the same model by simply adding more channels to the input does not necessarily improve performance, leaving open the following question: *Is there a way to better combine multiple attention maps to achieve superior deepfake detection performance?*

To pursue this question, we conducted experiments with DINOv2 FM [18] and the CNN-based XceptionNet [6], chosen for their established use and strong performance in this task. Initially, we evaluated each self-supervised feature map generated by DINOv2 individually, resulting in three distinct models. Next, we tried multiple committee approaches to combine these models' predictions. For our experiments, we used both the Deepfake Detection Challenge (DFDC) [10] and Face Forensics [21] datasets. In both

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

datasets, our proposed committees performed better than using a single attention map, while the performance gain varied depending on the specific committee approach employed.

The remainder of this paper is organized as follows. Section section 2 covers related works on deepfake generation and detection, as well as research on committee approaches for deep learning models. Section 3 describes the employed architectures and details the different committee approaches proposed for this task. Section 4 discusses how we conducted our experimental methodology, including specifics on data, setup, and results analysis. Finally, Section 5 presents our final thoughts and outlines possible directions for further research in this field.

2 RELATED WORK

In this section, we go over other important research that relates to the topics covered in this work. Subsection 2.1 discusses recent advances in deepfake detection, focusing on methodologies aimed at countering the generation of synthetic media. Subsection 2.2 explores diverse works related to the combination of ensemble approaches and CNN models for image processing, enhancing model robustness, accuracy, and generalization across diverse datasets and tasks.

2.1 Methods for Facial Deepfake Detection

Recent years have seen a large increase in efforts dedicated to detecting realistic deepfakes and differentiating them from genuine facial media, with video and images being the dominant targets of these works [23]. This progress has been enabled by the production of a multitude of recent datasets for the task, such as the Deepfake Detection Challenge (DFDC) [10], Celeb-DF [16], FaceForensics, and FaceForensics++ [21] datasets. These datasets contain large volumes of visual data with a variety of deepfake techniques and an increasing, though still insufficient, concern for demographic fairness [27], allowing their use for techniques that deal with both image and video.

Many recent works have been based on CNN classifiers, typically employing either EfficientNet [25] or XceptionNet [6], often combining other types of architectures. Tjon et al. [26] used EfficientNet B4 as an encoder in combination with Y-Net [17], achieving very good performance on the DFDC dataset for both image and video.

Due to its prominence, EfficientNet has been considered one of the best classifiers for this task, with entire studies dedicated to further analyzing its overall performance, such as the one conducted by Pokroy and Egorov [19]. In their work, the authors experimented with different versions of the architecture, which vary according to the dimensions of the input data and the number of trainable parameters. They trained each variant, from EfficientNet B0 to B7, for twenty epochs on the DFDC dataset, concluding that larger networks do not necessarily achieve superior performance.

Meanwhile, XceptionNet has gained notoriety in deepfake detection due to its state-of-the-art performance on the FaceForensics dataset [21] and its high performance on the DFDC dataset when combined with self-supervised features generated by FMs [12] . In this scenario, it achieved performance superior to that of the EfficientNet.

With the recent emergence of transformers on computer vision tasks, many researchers have also attempted to apply this technique to deepfake detection. Heo et al. [13] combined an EfficientNet B7 pre-trained on the DFDC dataset with a vision transformer by merging the embeddings extracted by these models and passing them on to the transformer encoder. This approach achieved state-of-the-art performance on the DFDC dataset for video format.

Similarly, both Coccomini et al. [8] and Wang et al. [28] proposed architectures that combine vision transformers and CNN-based classifiers. The former established an ensemble of two branches of Efficient ViT, one to deal with smaller features and the other to handle larger ones. Meanwhile, the latter proposed a multiscale architecture to identify regions synthesized by generative models. Both works achieved high performance on multiple datasets, further establishing vision transformers as a dominant approach for the task.

At the same time, FMs have started gaining ground recently, leading to the possibility that they could potentially overtake vision transformers in deepfake detection. In this light, Zhao et al. [31] proposed a self-supervised approach employing Contrastive Learning to detect deep fake videos through features obtained from lip movements, with two encoders for audio and video. This work reached close to state-of-the-art performance on the FaceForensics++ dataset, showcasing the potential of self-supervised approaches. Meanwhile, considering image analysis, self-supervised features have demonstrated significant improvements in the performance of CNN-based classifiers [12] .

Reiss et al. [20] have shown that combining textual information and other contextual sources with audio-visual data input can result in performance improvements, especially for certain types of attacks where data about the target is publicly available.

Lastly, Lanzino et al. [15] demonstrated that using Binary Neural Networks [9] can achieve performance close to the state-of-the-art on the recent COCOFake dataset [2], while keeping efficient in terms of computational cost.

2.2 Committee Approaches for Image Classification

Ali et al. [1] proposed a simple ensemble approach combining VGG19-UNet and DeeplabV3+ architectures for melanoma detection. Their experiments on the ISIC 2018 dataset, consisting of 2,594 dermoscopy images, demonstrated promising results with an overall accuracy of 93.6%, an average Jaccard Index of 0.815, and a dice coefficient of 0.887 on the test dataset. They highlighted the efficacy of ensemble techniques over individual architectures, especially in challenging cases like low contrast, ink, and dark corner artifacts, emphasizing its robustness and potential for broader imaging applications.

In their research on plant leaf recognition and disease detection, Chompookham and Surinta [7] addressed the complexities of computer vision challenges by proposing an ensemble CNN approach combining MobileNetV1, MobileNetV2, NASNetMobile, DenseNet121, and Xception models to enhance recognition accuracy. Ensemble techniques such as weighted averages were applied to combine predictions from multiple CNN models, showing superior performance over individual models across all datasets.

Consequently, their approach achieved accuracies of 99.93% and 99.47% on the tomato and corn leaf disease datasets, respectively, demonstrating that ensemble methods enhance the performance of CNN architectures.

Bonettini et al. [4] addressed the need for robust detection of manipulated faces in video sequences by exploring ensemble approaches using CNNs. Their study employed EfficientNetB4 as the base model, enhancing it with attention layers and siamese training techniques to improve detection accuracy on video datasets. The evaluation of the FaceForensics++ and DFDC datasets, yielding AUC values of 0.9444 and 0.8800, respectively, indicated superior performance when combining different CNN models compared to baseline methods. This highlights the efficacy of attention-based modifications in enhancing detection accuracy while providing insights into useful discriminative features for effective detection. Additionally, it addresses practical constraints by ensuring computational efficiency while achieving processing speeds suitable for real-world applications in limited hardware scenarios.

Also in the realm of deepfake detection, Giatsoglou et al. [11] investigated several ensemble architectures designed to enhance robustness and generalization across different types of facial manipulations, including deepfakes generated by technologies like FaceSwap and NeuralTextures. Using the FaceForensics++ dataset for training and evaluation, the research employs EfficientNet-B0 as the base classifier due to its balance of performance and resource efficiency, while using simple ensembles like binary detection, multi-class attribution, one-manipulation-vs-real, and one-manipulation-vs-rest. The results indicated that while ensembles can outperform individual models under certain conditions, their generalization across newer and more diverse datasets remains a challenge. This highlighted the need for future work to improve the ability of ensembles to detect increasingly sophisticated and varied manipulations in digital media.

Finally, Chaudhary et al. [5] proposed the integration of CNNs and Random Forest (RF) for the efficient classification of tomato diseases. By employing CNNs for feature extraction and the RF ensemble for accurate classification across four categories, the model achieved an overall accuracy of 97.03%. The performance of the hybrid CNN-RF approach surpassed traditional CNN-based models, which typically achieve accuracies between 88% to 92%, emphasizing the synergistic benefits of combining deep feature extraction capabilities with the interpretability and generalization power of RF.

While prior studies have explored different deepfake detection methods and ensemble approaches involving CNN models, this work integrates multiple sets of self-supervised features from FMs to enhance detection accuracy. Unlike traditional methods that often rely on single-model architectures or simple ensemble techniques, our approach considers the diverse feature representations of advanced FMs, resulting in a more robust and generalized detection system.

3 METHOD

Figure 1 illustrates our proposed integration of committee approaches for combining different self-supervised features, as well as the overall data flow of the classification process. It starts with extracting a frame from the original video and applying a face detector that results in an RGB patch $x \in \mathbb{R}^{(\mathrm{H} \times \mathrm{W} \times 3)}$ of a face. Next, the self-supervised facial model generates three feature maps $x_{am0}, x_{am1}, x_{am2} \in \mathbb{R}^{(\mathrm{H} \times \mathrm{W} \times 1)}$ from the given patch. Subsequently, the original RGB patch x is individually concatenated with x_{am0}, x_{am1} and x_{am2}, resulting in three tensors $x_{t0}, x_{t1}, x_{t2} \in \mathbb{R}^{(\mathrm{H} \times \mathrm{W} \times 4)}$, each of which is fed into a different classifier model trained to use that specific set of self-supervised features. The probabilistic output results of each classifier are finally sent to a committee, which uses them to determine whether an image should be classified as real or fake.

In the remainder of this section, we briefly discuss our chosen architectures, introducing the self-supervised facial feature extractor and the CNN-based classifier in Subsection 3.1. Furthermore, in Subsection 3.2, we elaborate on the different committee approaches we tested, describing their distinct methodologies for combining the results from the classifiers.

3.1 Models

3.1.1 *Self-Supervised Facial Feature Extractor.*

We used a model from the DINO family to extract the self-attention activation maps from images, specifically opting for the newer DINOv2 [18]. This updated model significantly outperforms its predecessor through three main improvements: significantly larger and more diverse training dataset known as LVD-142M, with 142 million images, enhanced training algorithms and implementation techniques using PyTorch2[1] and xFormers[2] for better stability and efficiency, and advanced knowledge distillation process for compressing large models into smaller ones without substantial accuracy loss. These enhancements contribute to DINOv2's superior understanding, segmentation capabilities, and performance across several tasks, maintaining high efficiency even with reduced model sizes.

The authors provided on their GitHub Repository[3] the weights of the pre-trained models both in the ViT-Base architecture, with 86M parameters and in ViT-Small (ViT-S), with 21M parameters. For our study, we selected the pre-trained ViT-S/14 model based on its good performance and efficiency in both time and computational resources to extract the self-attention maps from our dataset.

Similarly to previous works [12], we employed transfer learning from a pre-trained model to generate three different attention heads for self-supervised facial features from each facial image in the dataset. By applying a multi-crop strategy, we generated different views of the input image, which are subsequently processed by the networks that comprise the DINO model, generating probability distributions by normalizing the networks' output with a softmax function. These probabilities are then mapped onto the image, producing the attention maps. Figure 2 illustrates the self-attention activation maps extracted by DINOv2, showcasing examples of both correctly and incorrectly classified instances by our best committee classifier.

3.1.2 *CNN-based Classifier.*

In deepfake detection, considering self-supervised features has been shown to improve the performance

[1] https://pytorch.org/get-started/pytorch-2.0/
[2] https://github.com/facebookresearch/xformers
[3] https://github.com/facebookresearch/dinov2

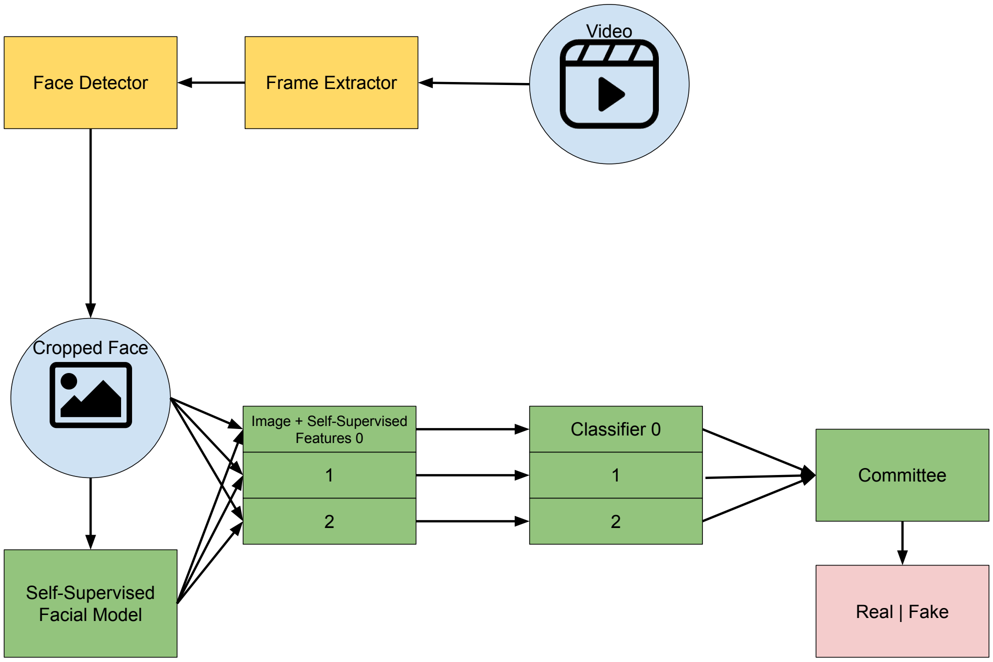

Figure 1: Proposed method for deepfake classification.

of multiple CNN-based classifiers [12] , resulting in various architectures as viable candidates for such tasks, including Inception-Resnet [24], EfficientNet B4 [25] and XceptionNet [6]. We chose to focus, for this work, on the latter due to its robust performance when combined with self-supervised features [12] and its prevalence among other works in deepfake detection [4, 21], allowing for more direct comparisons with existing research findings.

3.2 Ensemble Techniques

To combine the results of the three classifiers, we tested five different committee approaches denoted as *c0*, *c1*, *c2*, *c3* and *c4*. Each approach prioritizes different aspects in terms of performance from our classifier models.

c0 employs a simple majority vote, disregarding the probabilistic result outputs of each model and considering only their final classifications. The goal of this committee is to deal with situations where individual models may make mistakes, which can happen even with our highest-performing classifier. Furthermore, it also serves as a baseline for our approach.

c1 is a weighted vote approach where we aggregate the probabilistic scores assigned by each classifier to each class, selecting the class with the highest total score. This approach differentiates itself from *c0* in cases where a minority classifier is considerably more confident when compared to the majority classifiers. It works under the principle that a confident model is more likely to be correct. This assumption is validated by our finding, where aggregating individual results from our three classifiers, classifications with confidence over 0.85 were approximately 30% more accurate.

Following the same logic, *c2* is a confidence-based committee where if a single classifier has confidence above a given threshold, its classification is considered final. If no classifier meets this threshold or multiple classifiers do, the committee defaults to the weighted voting approach similar to *c1*. For this committee, we found our best results using a threshold of 0.85. This approach was expected to perform closely to *c1*, differentiating itself only in scenarios where a single high-confident classifier is not enough to surpass the combined confidence of the other classifiers.

c3 is another variation based on the confidence premise. It works similarly to *c1*, but it enhances the weight of votes from classifiers that exhibit confidence above a specified threshold to improve their numerical advantage. We again achieved our best results with a threshold of 0.85, multiplying the votes from confident classifiers by a factor of 1.5. This approach amplifies the influence of high-confident classifiers while still considering the collective outputs of all classifiers through a weighted voting scheme.

Lastly, *c4* is an MLP-based stacked committee that aims to learn more advanced patterns from the individual classifications. The model receives the probabilities of an image being fake, according to each classifier individually, and outputs its own probability estimate for the image being fake. In our experiments, the optimal architecture consisted of three dense layers with respective sizes of 16, 8, and 1, with the first two layers using the ReLU activation and the last one using a sigmoid to output the final probability.

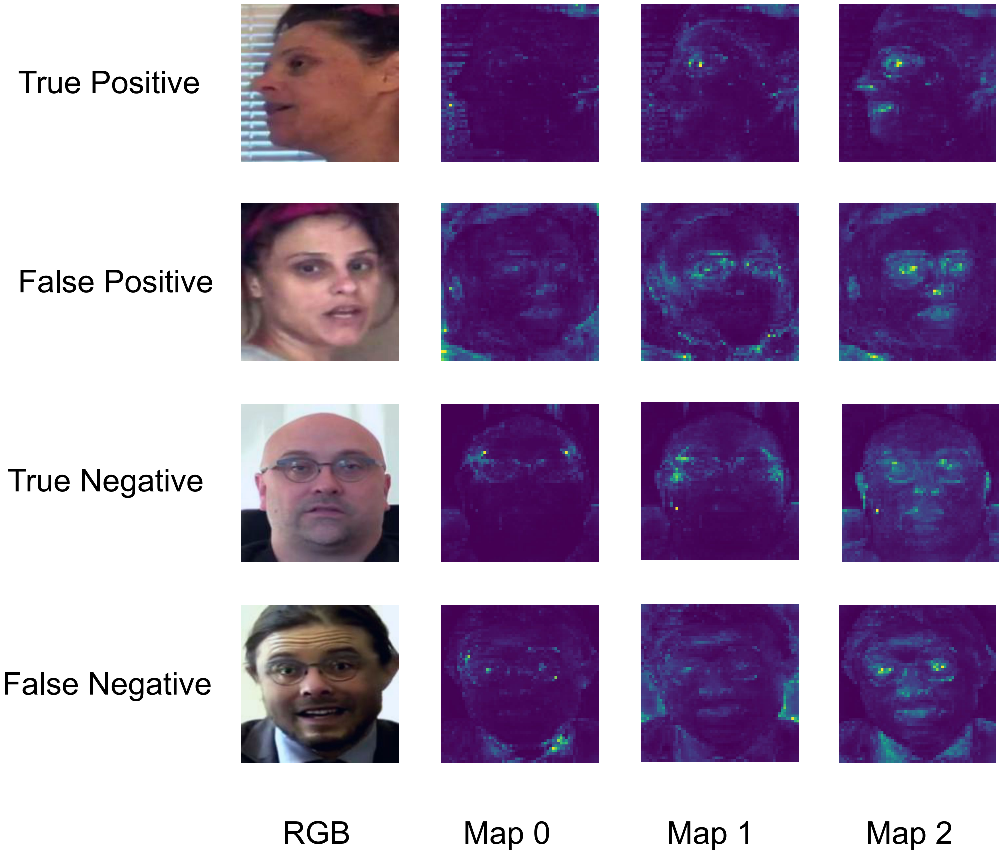

Figure 2: Examples of the attention heads extracted using DINOv2. A "positive" classification means that our best committee classified the image as fake, while "negative" means that it was labeled as real.

This approach employs the MLP's ability to learn intricate relationships among its inputs (our individual classifier outputs), potentially improving overall classification accuracy.

4 EXPERIMENTS

In this section, we present our experiments designed to measure the impact of different committee approaches described in Section 3, in terms of performance, aiming to learn if significant improvements can be achieved by committees over using a single attention map. The first subsection covers the datasets used in our experiments, for both training and testing. The second subsection describes the setup used to conduct the experimental evaluation. Lastly, the third subsection elaborates on the results, comparing them to our baselines and previous works in the field of Deepfake Detection.

4.1 Datasets

We mainly conducted experiments on the Deepfake Detection Challenge (DFDC) dataset [10], using a large subset as our training data and a smaller one for testing purposes. As pointed out by other works [12] , this dataset provides a diverse set of lighting conditions, resolutions, image qualities, and demographic attributes, with the latter being particularly important for developing fairer models for this kind of task [29].

For the sake of a more direct comparison to previous works, we followed a similar process of preparing an image dataset from each video in the DFDC dataset as described in Gomes et al. [12] . We extracted approximately 10 frames on average from each video and used the Multitask Cascaded Convolutional Network (MTCNN) face detector [30], with identical parameters, to identify and crop faces from each frame. This process was applied across more than 124,000 videos in the dataset, resulting in 1,086,737 images for training and 144,316 for validation.

In addition to using the DFDC dataset, we also employed the Low-Quality (LQ) version of the Face Forensics [21] dataset to further validate our approach. This choice came from the past use of the XceptionNet [6] on this dataset, leading to direct comparisons among its standard version, the ones using self-supervised attention [12], and our proposed committee approaches combining these attention maps. To this end, we followed the same methods of frame extraction and cropping provided by the dataset authors.

4.2 Setup

In our experiments, following the approach of Gomes et al. [12], we trained four models of our chosen architecture: a baseline model trained only on 3-channel RGB facial image inputs, and three additional models, each using one of the three attention maps presented in Section 3.1.1 as a fourth channel. The goal was to determine a performance baseline and evaluate how each attention map individually impacted performance, serving as a basis for comparison with our different committee approaches. The training was conducted exclusively on images extracted from the DFDC dataset [10], following the process described in Subsection 4.1, while the Face Forensics dataset [21] was reserved for testing purposes.

For training and evaluation, we kept the default input sizes of our chosen architecture. We also used an Adam optimizer with a learning rate of 1e-4 and the categorical cross-entropy loss function. Our computational setup included a system with 48 GB RAM, 850 GB of storage capacity, and an NVIDIA RTX 2080 with 11 GB VRAM.

4.3 Results

We evaluated our proposed committees using both the validation subset of the DFDC dataset and the Face Forensics dataset, as described in Subsection 4.1. For validation, we used AUC and F1-Score as our primary metrics due to their ability to provide deeper insights into model performance, especially important in deepfake detection [29]. Conversely, for the FaceForensics segment, we judged that simple accuracy was sufficient as it would lead to a more direct and clear comparison with the results reported by Rossler et al. [21]. Comparing our results to those obtained without the use of self-attention, as well as those using each attention map individually, each of our proposed approaches showed improvements in performance to varying degrees.

Table 1: Performance impact of different approaches on the validation subset of the DFDC dataset, showing how committees improve performance over the use of individual attention maps.

Approach	**AUC (%)**	**Diff. (%)**	**F1-Score (%)**	**Diff. (%)**
Baseline	91.91	- - -	84.36	- - -
Map 0	92.03	+0.12	84.39	+0.03
Map 1	92.10	+0.19	84.68	+0.32
Map 2	91.99	+0.08	84.37	+0.01
c0	92.16	+0.25	84.71	+0.35
c1	92.47	+0.56	85.04	+0.68
c2	92.52	+0.61	85.07	+0.71
c3	92.56	+0.65	85.12	+0.76
c4	**93.22**	**+1.31**	**85.63**	**+1.27**

Table 1 shows the results from the DFDC dataset, highlighting the performance impact of the different committee approaches. **c0** represented our simplest committee, with performance matching its simplicity when compared to more complex approaches. Despite its straightforward methodology approach, it still achieved slightly superior performance over the best individual classifier, with an AUC of 92.16%, a 0.25% improvement over the baseline, and a marginal 0.06% increase over Map 1. However, this modest gain suggests a need to explore other approaches to achieve more substantial performance improvements.

The *c1* approach showed significant improvements over *c0*, achieving an AUC of 92.47%, which is 0.56% higher than the baseline and 0.31% higher than *c0*. Additionally, it presented an F1-Score of 85.04%, representing a gain of 0.68% over the baseline, 0.36% over Map 1, and 0.33% over *c0*. This superior performance over *c0* indicates the presence of instances where two models with low confidence might misclassify an image through a simple majority approach, highlighting the better results of confidence-based strategies.

In the same rationale, the *c2* and *c3* approaches exhibited very similar performance, with AUCs of 92.52% and 92.56%, and F1-Scores of 85.07% and 85.12%, respectively. Both results confirm our assumptions about prioritizing classifications with high confidence, as both committees showed improvements over our previous approaches. Furthermore, the superiority of *c3* over *c2* indicates that relying on high confidence alone is not enough to decide a final classification, especially in cases when other models are equally confident but collectively more accurate.

Finally, the *c4* committee stood out as the best-performing approach by a significant margin. It achieved an AUC of 93.22%, surpassing the baseline by 1.31%, Map 1 by 1.12%, and *c3* by 0.66%. Additionally, it achieved an F1 Score of 85.63%, with incremental gains of 1.27%, 0.95%, and 0.51% over the baseline, Map 1 and *c3*, respectively. These considerable gains highlight the presence of intricate patterns in the relationships among the three models, meaning that simple voting mechanisms might overlook scenarios that require more complex decisions for accurate classification.

Conversely, Table 2 shows the results from our experiments on the LQ version of the Face Forensics dataset, comparing them to the baseline performance of the XceptionNet reported in a previous study [21]. While the best individual classifier (Map 1) showed a significant improvement of 1.62% over the considered baseline, achieving an accuracy of 82.62%, it was again surpassed by all committee approaches. The hierarchy of performance among the committees remained the same, with *c0* being significantly behind at an accuracy of 82.71%, *c1*, *c2*, and *c3* showing similar results of 83.24%, 83.31%, and 83.42%, respectively. Once more, *c4* presented the most significant gain margin, achieving an accuracy of 84.73%, which is 3.73% higher than the baseline model. This last value corresponded to an AUC of 82.54% and an F1-Score of 80.27%.

5 CONCLUSION

The growing prevalence of deepfake creation steers research into advancing detection techniques, a critical challenge in digital media. Through exploring strategies integrating sophisticated models and ensemble learning, improvements in detection accuracy can be achieved, enhancing our ability to differentiate between genuine and manipulated content. These efforts reflect a continuous step to mitigate the risks associated with deceptive media in online environments.

Table 2: Performance impact of the different approaches on the LQ version of the Face Forensics dataset.

Approach	Accuracy (%)	Diff. (%)
Baseline	81.00	- - -
Map 0	82.21	+1.21
Map 1	82.62	+1.62
Map 2	81.87	+0.87
c0	82.71	+1.71
c1	83.24	+2.24
c2	83.31	+2.31
c3	83.42	+2.42
c4	**84.73**	**+3.73**

In this work, we showed how combining multiple maps of self-supervised features generated by an FM, through the use of different committee approaches, has a positive performance impact on realistic facial deepfake detection by experimenting with three classifiers, trained on different feature maps and combined through various committee techniques, on two prominent datasets for the task. These findings significantly contribute by improving results found with approaches that relied on a single feature map [12], demonstrating that integrating multiple maps can enhance performance in deepfake detection. This especially solves situations where the optimal feature map to be used is uncertain.

The results show that each committee approach outperformed using a single attention map, indicating significant disparities among different approaches. Techniques based on classifier confidence showed improved predictions compared to simple majority voting approaches. Finally, an MLP-based approach, trained to identify more complex patterns in prediction relationships, achieved the highest performance.

The success of these approaches opens a door for larger committees that can combine more classifiers and utilize other self-supervised feature maps. This indicates the potential to extract richer information from each frame, enhancing detection performance. Furthermore, our findings suggest that similar techniques could be adapted for video formats, which would not require frame extraction. This is enabled by advancements in FMs such as DINOv2 [18], which have already shown significant performance gains for video content in recent years.

ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0475, and also financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) and by CNPQ.

REFERENCES

[1] Redha Ali, Russell C. Hardie, Barath Narayanan Narayanan, and Supun De Silva. 2019. Deep Learning Ensemble Methods for Skin Lesion Analysis towards Melanoma Detection. In *2019 IEEE National Aerospace and Electronics Conference (NAECON)*. IEEE, Dayton, OH, USA, 311–316. https://doi.org/10.1109/NAECON46414.2019.9058245

[2] Roberto Amoroso, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. 2024. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. arXiv:2304.00500 [cs.CV] https://arxiv.org/abs/2304.00500

[3] Ben Beaumont-Thomas. 2024. Taylor Swift deepfake pornography sparks renewed calls for US legislation. https://www.theguardian.com/music/2024/jan/26/taylor-swift-deepfake-pornography-sparks-renewed-calls-for-us-legislation.

[4] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video Face Manipulation Detection Through Ensemble of CNNs. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Milan, Italy, 5012–5019. https://doi.org/10.1109/ICPR48806.2021.9412711

[5] Preeti Chaudhary, Aditya Verma, Vinay Kukreja, and Rishabh Sharma. 2024. Integrating Deep Learning and Ensemble Methods for Robust Tomato Disease Detection: A Hybrid CNN-RF Model Analysis. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. IEEE, Noida, India, 1–4. https://doi.org/10.1109/ICRITO61523.2024.10522213

[6] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Honolulu, HI, USA, 1251–1258. https://doi.org/10.1109/CVPR.2017.195

[7] Thipwimon Chompookham and OJIEL Surinta. 2021. Ensemble methods with deep convolutional neural networks for plant leaf recognition. *ICIC Express Letters* 15, 6 (2021), 553–565.

[8] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In *Image Analysis and Processing – ICIAP 2022*, Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari (Eds.). Springer International Publishing, Cham, 219–229. https://doi.org/10.1007/978-3-031-06433-3_19

[9] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. arXiv:1602.02830 [cs.LG] https://arxiv.org/abs/1602.02830

[10] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge Dataset. arXiv:2006.07397 [cs.CV]

[11] Nikolaos Giatsoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2023. Investigation of ensemble methods for the detection of deepfake face manipulations. arXiv:2304.07395 [cs.CV] https://arxiv.org/abs/2304.07395

[12] Bruno Rocha Gomes, Antonio J. G. Busson, José Boaro, and Sérgio Colcher. 2023. Realistic Facial Deep Fakes Detection Through Self-Supervised Features Generated by a Self-Distilled Vision Transformer. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web (WebMedia '23)*. Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3617023.3617047

[13] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. 2021. Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353* abs/2104.01353 (2021), 7 pages. https://doi.org/10.48550/arXiv.2104.01353

[14] Brittaney Kiefer. 2023. This Brand's Social Experiment Uses AI to Expose the Dark Side of 'Sharenting'. https://www.adweek.com/brand-marketing/this-brands-social-experiment-uses-ai-to-expose-the-dark-side-of-sharenting/.

[15] Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. 2024. Faster Than Lies: Real-time Deepfake Detection using Binary Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Seattle, WA, USA, 3771–3780.

[16] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 3207–3216.

[17] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G. Elmore, and Linda Shapiro. 2018. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II* (Granada, Spain). Springer-Verlag, Berlin, Heidelberg, 893–901.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal* 1 (2024), 1–31. https://doi.org/10.48550/arxiv.2304.07193

[19] Artem A Pokroy and Alexey D Egorov. 2021. EfficientNets for deepfake detection: Comparison of pretrained models. In *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. IEEE, St. Petersburg, Moscow, Russia, 598–600. https://doi.org/10.1109/ElConRus51938.2021.9396092

[20] Tal Reiss, Bar Cavia, and Yedid Hoshen. 2023. Detecting Deepfakes Without Seeing Any. *ArXiv* abs/2311.01458 (2023), 16 pages. https://api.semanticscholar.org/CorpusID:264935112

[21] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Seoul, Korea (South), 1–11. https://doi.org/10.1109/ICCV.2019.00009

[22] Rhianna Schmunk. 2024. Explicit fake images of Taylor Swift prove laws haven't kept pace with tech, experts say. https://www.cbc.ca/news/canada/taylor-swift-ai-images-highlight-need-for-better-legislation-1.7096094.

[23] Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip, and Mohiuddin Ahmed. 2023. A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology* 7, 2 (2023), 83–113. https://doi.org/10.1080/23742917.2023.2192888 arXiv:https://doi.org/10.1080/23742917.2023.2192888

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*. AAAI Press, San Francisco, California, USA, 4278–4284. https://doi.org/10.48550/arXiv.1602.07261

[25] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, CA, USA, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

[26] Eric Tjon, Melody Moh, and Teng-Sheng Moh. 2021. Eff-YNet: A Dual Task Network for DeepFake Detection and Segmentation. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, Seoul, Korea (South), 1–8. https://doi.org/10.1109/IMCOM51814.2021.9377373

[27] Loc Trinh and Yan Liu. 2021. An Examination of Fairness of AI Models for Deepfake Detection. arXiv:2105.00558 [cs.CV]

[28] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. 2022. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval* (Newark, NJ, USA) *(ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 615–623. https://doi.org/10.1145/3512527.3531415

[29] Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. 2023. A Comprehensive Analysis of AI Biases in DeepFake Detection With Massively Annotated Databases. arXiv:2208.05845 [cs.CV]

[30] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

[31] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. 2022. Self-supervised Transformer for Deepfake Detection. arXiv:2203.01265 [cs.CV] https://arxiv.org/abs/2203.01265