

Towards Fair & Robust Deepfake Detection: Bias Measurement & Mitigation Across Diverse Datasets

Shrishti Yadav¹, Shilpi Gupta², Pragya Mishra³, Devesh Katiyar⁴, Gaurav Goel⁵

^{1 2 3} Students, ^{4 5} Assistant Professors Department of Computer Science

Dr. Shakuntala Misra National Rehabilitation University, Lucknow, Uttar Pradesh, India

Abstract - The technology of deepfake is getting better fast and it is a big problem for people to know what is real and what is not. Deepfake uses learning to make fake pictures and videos that look very real. Many people have made systems to find deepfakes. These systems do not work well when they are used in the real world. This is because the systems are biased and that means they are not fair to everyone.

This paper is about the problems with deepfake detection and how to make it fair and work well for everyone. We looked at where the biases come from and how to measure if a system is fair. We also looked at ways to make the systems less biased. What we found is that even the good systems do not work well with different data.

So we made a plan to make deepfake detection fair and work well. The plan has four steps:

- test the system with different datasets
- make sure the system is trained with data from many different people
- make the system explain how it makes decisions
- have rules to make sure the system is used fairly

We think this plan will help make deepfake detection better and more fair for everyone.

Key Words: deepfake detection, fairness, bias mitigation, cross-dataset generalization, explainability, AI governance.

I. INTRODUCTION

The field of intelligence has changed a lot in the ten years. This is because of technologies like Generative Adversarial Networks and diffusion-based models.

Generative Adversarial Networks make it easy to create images and videos and audio recordings. These are called deepfakes.

People first used Generative Adversarial Networks to make effects in movies and create characters. Now Generative Adversarial Networks are being used for bad things.

Generative Adversarial Networks are being used to spread information and create fake political content.

They are even being used to steal peoples identities. This is really bad, for the field of intelligence and Generative Adversarial Networks.

The problems caused by deepfakes are very serious. Fake videos of people can change the way people think about politics and fake audio recordings can be used to trick people into giving away their money. In journalism and law it is hard to tell what is real and what is fake which makes it difficult to trust evidence. Because of this researchers are working hard to create systems that can automatically detect deepfakes.

Even though there has been a lot of progress there is still a problem with deepfake detection systems: they are not fair to all people. Most systems are designed to be accurate on average. They do not work well for people with darker skin, women or people from different cultures. This is not a technical problem it is also a moral and legal issue.

Imagine a situation where a deepfake detection system used by a media company always says that real videos of people from minority groups are fake. This would mean that their videos are removed and their voices are not heard while fake videos of people are allowed to stay. This would make people lose trust in systems that use intelligence to decide what content is allowed and it would raise questions about whether these systems are fair.

Another problem with deepfake detection systems is that they are not robust. Models that are trained on one set of data often do not work well when they are tested on data. This means that these models may be learning things that're specific to one set of data rather than learning what makes a deepfake a deepfake. Until we can solve these two problems. Fairness and robustness. We cannot trust deepfake detection systems to make decisions.

This paper tries to solve these problems by looking at the state of deepfake detection analyzing how fair these systems are and proposing a framework for creating systems that are fairer and more robust. We do three things: we identify the sources of bias in deepfake

detection systems we look at evidence from recent studies and we propose a framework for creating systems that are fair and robust. We call this framework governance-aware. It has four phases.

- We look at how detection systems work and where they can go wrong.
- We analyze how fair these systems are and how they can be improved.
- We propose a framework, for creating systems that're fairer and more robust.
- We test this framework. See how it works in practice.

2. BACKGROUND & RELATED WORK

2.1 Evolution of Deepfake Generation

The term deepfake comes from combining learning and fake media. At first deepfakes used encoder-decoder architectures and face-swapping algorithms. Then Goodfellow and others introduced GANs, which was a change. GANs helped create faces. After that architectures like StyleGAN, FaceSwap and Face2Face made it easier to create synthetic media. Recently text-to-video diffusion models have made it possible to create video sequences from text.

2.2 Deepfake Detection Approaches

Early detection systems looked for things like facial features, unnatural blinking and unusual color patterns. Then deep learning came along. Detection systems started using convolutional neural networks and transformer-based architectures. These systems can spot problems in videos. Models like EfficientNet-B4 and Vision Transformers are really good at detecting deepfakes. They do not work equally well for everyone and can be fooled by different types of synthetic media.

2.3 Fairness in Machine Learning

The machine learning community has been working on fairness for a while. They have created metrics like Demographic Parity, Equalized Odds and Individual Fairness. These metrics were first used in areas like credit scoring and crime prediction. Now they are being applied to media forensics, including deepfake detection. However not many studies have looked at fairness, in deepfake detection especially when it comes to demographic groups [1, 4].

3. SOURCES OF BIAS IN DEEPPFAKE DETECTION

3.1 Dataset Bias

The biggest problem with deepfake detection is the way the training datasets are made. Some popular tests, like FaceForensics++ and DFDC use video footage that anyone can see. It does not have enough different kinds of people to show what the whole world is like. So when models are trained with these datasets they get better at making decisions for the kinds of people that are seen a lot in the training data. Deepfake detection has this issue because the training datasets are not good enough. Deepfake detection models learn from these datasets and deepfake detection is not fair, to all people because of this. When deployed on data from underrepresented groups, error rates increase substantially.

The deepfake generation methods that are used to make samples in these datasets are based on certain techniques that people used when they were made. Now we have methods to make deepfakes especially those that use diffusion models. These new methods can make things that are different, from what the model is used to finding. This means the model will have a time finding these new deepfakes and it will make mistakes. The deepfake generation methods and the model will not work well together because of this.

3.2 Model Architecture Bias

Convolutional networks and transformer models can pick up demographic information in their learned representations. This happens even when they are not specifically trained on labels.

The reason is that facial appearance is often linked to attributes.

Convolutional networks and transformer models use facial appearance, as a shortcut to make predictions.

As a result they can inadvertently encode information.

This issue arises because facial appearance can be an indicator of demographic attributes. Models may thus develop different decision boundaries for faces that share demographic characteristics, leading to systematically higher false positive or false negative rates for certain groups. Adversarial probing experiments have confirmed that such demographic leakage is widespread in current detection models.

3.3 Evaluation Bias

Model evaluation can be biased in a way. This happens when we do not measure fairness. Usually people report how accurate a model is overall. This can hide big

differences in how well the model works for different groups of people. For example a model can be correct 92 percent of the time. It can be correct 98 percent of the time for one group of people and only 76 percent of the time for another group. If we do not look at the results, for each group separately we will not see this problem. We will not fix it. Model evaluation and fairness are important. Model evaluation is when we check how well a model is working. Fairness is when a model works well for all groups of people. If we do not check for fairness during model evaluation we might not see that a model is not working well for some groups. This is a problem because model evaluation is used to decide if a model is good or not. Model evaluation is used to check the accuracy of a model. Accuracy is when a model is correct. Model is a computer program that makes predictions.

4. EMPIRICAL EVIDENCE & ANALYSIS

4.1 Cross-Dataset Performance

The thing about deepfake detection is that it does not work well when you use a different set of pictures to test it. If you use the pictures that you used to teach the model it can tell if a picture is fake about 90 to 95 percent of the time. If you use a different set of pictures like the ones from Celeb-DF or DFDC the model is only correct about 70 percent of the time or less. This is a problem with deepfake detection models. They are not good at finding pictures in general. Instead they are just good at finding pictures in the specific set of pictures they were taught with. Deepfake detection models are not learning what makes a picture a deepfake in general. They are just learning what makes a picture a deepfake in the set of pictures they were taught with. This means that deepfake detection models are not very good, at finding deepfakes in pictures they have never seen before.

More recent studies have extended these evaluations to include newer datasets generated with diffusion-based methods. The performance degradation is even more severe in these cases, with some well-performing models reducing to near-chance accuracy. This indicates that the representational gap between GAN-based training data and diffusion-based test data is sufficient to completely invalidate learned detection heuristics.

4.2 Demographic Disparities

Several studies have found differences in how well deepfake detection works for different people. Research shows that real content is more likely to be labeled as fake for people with darker skin and for women. These differences are not isolated incidents.

A closer look at the numbers shows that accuracy differences between groups are often than 15-20 percentage points.

This is a deal especially when compared to areas like financial credit scoring or medical diagnosis where accuracy is crucial.

These results point to a problem with the data used to train deepfake detection models.

The people and footage used to train these models seem to be from certain groups.

This leads to biases, in the models.

To fix this we need to change how we collect and prepare data. We should make sure that the data used to train these models includes a range of people.

This way deepfake detection can work fairly for everyone.

The way we gather data needs to be overhauled to ensure representativeness.

Deepfakes and real footage should reflect the diversity of the population. That way trained models can make accurate judgments.

4.3 Explainability Analysis

Explainability tools like Grad-CAM and SHAP help us understand how trained deepfake detection models make predictions.

These tools show that models often focus on things like compression patterns, watermarks or background characteristics in the data than the actual signs of a deepfake, which are changes to the face.

This is a problem because these things are not the same across all data and they can be different for groups of people.

As a result deepfake detection models may not work well when they are tested on data and they may also be unfair to certain groups of people.

The models rely on these artifacts instead of the real signs of a deepfake.

This reliance causes the models to be brittle when tested across datasets.

It also leads to demographic bias because these artifacts are not evenly spread across different demographic groups, in the dataset.

5. FAIRNESS METRICS AND MEASUREMENT APPROACHES

We need to turn fairness ideas into things we can measure. This requires a set of fairness metrics. The people who work on detecting deepfakes are starting to use fairness metrics from areas of fair algorithm work but they still have not agreed on a standard set of fairness metrics. Fairness metrics are important for fairness measurement approaches. The deepfake detection community is using fairness metrics, for fairness measurement approaches. Fairness metrics standardization is still not complete.

5.1 Group Fairness Metrics

Demographic Parity requires that the positive prediction rate, here interpreted as the deepfake classification rate, be equal across demographic groups. This measure is easy to understand. It does not consider the real differences in the amount of deepfake content in different groups. Equalized Odds is a way to look at this because it makes sure that the true positive rates and false positive rates are the same for all groups. This means that the detection performance and false alarm rates are fair for everyone. The False Positive Rate Disparity looks at the difference in alarm rates, between the groups that are most affected and the groups that are least affected by deepfake content. Deepfake content is a problem and the False Positive Rate Disparity is an important measure because it shows how often real content is wrongly flagged and this can cause harm to people who create this real content, specifically the individuals whose real content is mistakenly flagged as deepfake content.

5.2 Individual Fairness Metrics

Individual fairness requires that similar individuals receive similar predictions, where similarity is defined with respect to the detection task rather than demographic attributes. Measuring individual fairness in deepfake detection is methodologically challenging because defining appropriate similarity measures for faces is non-trivial. Recent work has proposed contrastive learning-based approaches to enforce individual fairness constraints during training, showing promising results in reducing both demographic disparities and cross-dataset brittleness [2].

6. BIAS MITIGATION STRATEGIES

6.1 Data-Level Interventions

Data-level interventions focus on improving the demographic diversity and representativeness of training datasets used for deepfake detection.

6.2 Algorithm-Level Interventions

Fairness-aware training objectives change the classification loss function to include penalties for differences in performance across demographics.

This way the model learns to be fair and accurate. Adversarial debiasing techniques add another classifier that tries to guess attributes from the models internal workings.

The main model is trained at the time to minimize deepfake detection loss and maximize the uncertainty of the auxiliary classifier.

The adversarial objective helps the main model learn representations that're useful for deepfake detection and fair across demographics.

Domain adaptation and learning approaches are also used to improve performance on new datasets.

These approaches train models to quickly adapt to domains with limited supervision.

By doing they reduce the reliance on specific details, in the dataset that can make models fragile and biased. This helps models to generalize better and be more robust.

Models trained with these approaches are less likely to be biased.

6.3 Post-Processing Interventions

Threshold calibration methods change the decision limit for each group so that fairness measures are the same.

This approach is easy to do. It makes us wonder if equal limits really mean fairness or just a practical fix that doesn't deal with the models underlying problems. Human-in-the-loop systems are another way to handle things after the fact.

They have a person review cases where the model's not very confident or where certain groups might be more likely to be misclassified.

Human review can help in cases where model confidence's low.

It also helps for groups that might have a chance of being misclassified.

Models can have biases.

Threshold calibration and human review are two ways to deal with these biases. The goal is to make the model fair, for all groups.

7. PROPOSED FRAMEWORK FOR EQUITABLE DEEPFAKE DETECTION

We looked at what people're saying and the facts we have. We think there should be a four-phase framework to help make deepfake detection systems that're fair and really work. This framework is supposed to guide people when they are making and using these deepfake detection systems. The goal is to have deepfake detection systems that're fair and robust.

Phase 1: Foundational Readiness

In the beginning people who have a stake in this project figure out what they mean by fairness. Set some basic rules for how many different types of people should be included in the groups they are studying. They do this by looking at all the groups they have to see if any types of people are missing deciding how they will measure if the system is being fair and setting up a system to make sure someone is responsible for making sure the system is fair. This part of the project also includes making plans for how they will keep track of the types of people using the system and make reports, about it once it is being used by people.

Phase 2: Targeted Piloting

Model development and evaluation are conducted in controlled settings with explicit demographic stratification. We make sure that our training methods are fair from the start. We do not just add fairness on as a fix. We test our models on datasets to see how well they work for different groups of people and with different methods. We also look at how the models make decisions to check if they are relying on things that're not relevant. What we learn from this helps us make our models better before we use them widely.

Phase 3: Iterative Scaling

We take models that're fair and work well in tests and use them in real-life situations that are more diverse. We keep an eye on how they work for different demographic groups. If a model makes mistakes we look at what went wrong especially if the mistakes happen often with certain groups of people or with certain types of fake data. We regularly update our models when new methods for making data appear and when we get more data, from different demographics. This helps our deepfake detection models stay accurate and fair.

Phase 4: Governance & Sustainability

Fairness is something that we need to think about for a time. We have to make sure that fairness is part of the rules that institutions follow. This means we need to have people from check that everything is fair. We also need to

make sure that institutions are open about how they're doing. They need to share information about how different groups of people're affected.

We should also have ways for people to tell us if they think something is not fair.

It is an idea for countries to work together on this. We can look at things like the IEEE P7003 bias framework and UNESCO ethics guidelines. This will help us make sure that everyone is doing things in a way. Fairness is important. We need to make sure that fairness standards are the same everywhere. This way we can avoid having rules, in different places.

8. DISCUSSION

The evidence in this paper shows that fairness and robustness in deepfake detection are crucial for the legitimacy and trustworthiness of these systems.

Deepfake detection systems can be unfair to groups of people.

This happens because only overall accuracy is reported and not how well the system works for groups.

These unfair systems are used in areas, such as checking content in courts and verifying identities.

So if deepfake detection systems are not fair they can cause harm.

The good news is that the technical community has started to develop tools to address these challenges.

Some of these tools include fairness- training, debiasing, domain adaptation and involving humans in the design process.

Combining strategies seems to work better than using just one.

However the challenge is not just technical; it is also about how organizations and regulations work.

Without rules and accountability there is incentive to make these systems fair.

There is a trade-off between fairness and accuracy in some cases.

Making a system fair can slightly reduce its accuracy.

This trade-off depends on the situation and who is affected by errors.

It is essential to discuss and agree on these trade-offs openly than making them implicit in the system design.

This open discussion is a part of responsible AI development and deepfake detection systems must be fair and robust.

The use of deepfake detection systems requires fairness and robustness to ensure that they are legitimate and trustworthy.

Fairness and robustness, in deepfake detection are essential.

9. CONCLUSION AND FUTURE DIRECTIONS

This paper looks at the problems of bias and fairness in detecting deepfakes. It uses information from studies and suggests a practical way to make detection systems more fair. The main thing we found out is that current deepfake detection models are not good at detecting all types of people and are not very good at working with sets of data. We need to work on this problem from sides, including the data we use the algorithms, how we test them and how we govern them.

Looking to the future there are a things that we need to focus on. First we need to make datasets of deepfakes that include many different types of people and are made with the latest methods. Second we should use frameworks to make sure that our detection systems are fair to each person not just to groups of people. Third because the technology to make deepfakes is getting better fast our detection systems need to be able to adapt and change all the time. Lastly we should do long-term studies to see how deepfake detection systems affect communities so we can make good policies.

Deepfake detection is not about solving a technical problem it is also about doing what is right for society. When you are on the internet it is getting really tough to figure out what is real and what is a deepfake. The systems that are supposed to keep the internet safe for us need to be systems that we can trust. This means these systems have to be accurate and fair to everyone and they have to be transparent and accountable to the people who use them.

Deepfake detection systems have to be very good at finding deepfakes and they have to be fair to all people so we can trust deepfake detection systems. We need to make sure that deepfake detection systems are good for everyone who uses the internet.

Making deepfake detection systems is a job, for the people who create deepfake detection systems. The people who use deepfake detection systems also have a job. We all need to be able to trust deepfake detection systems. The people who create deepfake detection systems and the people who use these deepfake detection systems have to work so we can trust these deepfake detection systems. The people who make deepfake detection systems need to do a job so the people who use deepfake detection systems can get accurate results, from the deepfake detection systems. This way we can all trust the deepfake detection systems to do what they are supposed to do.

REFERENCES

- [1] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu, "Improving Fairness in Deepfake Detection," Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV), pp. 1234–1243, 2020.
- [2] A. Hou, L. Lin, J. Li, and S. Hu, "Rethinking Individual Fairness in Deepfake Detection," arXiv preprint arXiv:2507.14326, 2026.
- [3] S. Agarwal, H. Farid, and Y. Gu, "Cross-Dataset Generalisation in Deepfake Detection: Biases and Remedies," IEEE Trans. Information Forensics and Security, vol. 19, pp. 456–469, 2024.
- [4] K. Zhang, M. Chen, and R. Wang, "Fairness Auditing in Multimedia Forensics: A Case Study on Deepfake Detection," ACM Multimedia Conf., pp. 212–221, 2022.
- [5] IEEE Standards Association, "IEEE P7003: Algorithmic Bias Considerations in AI Systems," IEEE Standards Draft, 2021.
- [6] UNESCO, "Ethics of Artificial Intelligence Global Report," UNESCO Publishing, 2023.
- [7] OECD, "Recommendation on AI Fairness and Transparency," OECD Digital Policy Papers, 2025.