

HUMAN PERFORMANCE IN DETECTING DEEPFAKE MEDIA: AN EXPERIMENTAL EVALUATION IN INDIA

ANINDITA MALIK¹, DEEPAK RAJ RAO G², KIRAN KUMARI³

¹Research Scholar, Department of Forensic Science, Lovely Professional University, Phagwara 144411, Punjab, India

²Professor of Practice, National Forensic Science University, New Delhi 110085, India

³Assistant Professor, Department of Forensic Science, Lovely Professional University, Phagwara 144411, Punjab, India

E-mail: ¹anindita96malik@gmail.com, ²deepakraj.rao@nfsu.ac.in, ³malikraw91@gmail.com

ABSTRACT

The rapid advancement of deepfake technologies has made it increasingly difficult to distinguish between real digital content from Artificial Intelligence (AI) generated media. This study examines human efficiency to detect deepfakes through a controlled media-classification experiment. It was hypothesized that the human ability to detect deepfakes would be low in general and vary significantly across the demographic and behavioral characteristics such as age, occupation, social media usage, and prior familiarity with deepfake technology. Participants classified a series of images and videos as deepfake or real and identified the perceptual cues influenced their decision. The responses were analyzed using Signal Detection Theory (SDT), Dual-Process Theory (DPT), and the Heuristic-Systematic Model (HSM), to interpret the decision accuracy and cognitive processes underlying it. Results confirmed the hypothesis: participants showed substantial difficulty distinguishing genuine content from manipulated content; with misclassification rates exceeding 53% across categories. Detection accuracy differed meaningfully by age, profession, use of social media and familiarity with deepfake technology, with senior, more experienced, and more digitally active participants performing better. Participants relied predominantly on surface level cues like media quality and facial expressions instead of systematic analysis. The findings reveal the critical gap between the advancement of generative AI and society's ability to detect it, emphasizing the need for stronger digital media literacy initiatives and hybrid human-AI detection frameworks to ensure information safety.

Keywords: *Deepfake, Artificial Intelligence, Synthetic Media, Generative Adversarial Networks, Deepfake Detection Framework*

1. INTRODUCTION

One of the significant issues of the present-day information systems is that digital media can be easily changed and spread quickly online. The fact that such information can get widely distributed (particularly through social media platforms like - facebook, youtube, etc.) means that it is threatening to democracy and the quality of public dialogue [1-4]. It is more and more necessary to discern between genuine information and information that is not real, given that digital accessibility has been widening all over the world and especially in the Indian region where using digital platforms has become one of the main sources of news and information. Therefore,

figuring out whether such content is authentic and credible is becoming more complicated and complicated [2,5].

One of the most alarming aspects of digital manipulation is deepfake technology which involves the generation of photorealistic images of real people using artificial intelligence. The name came into popular usage in 2017 when an involuntary AI-generated content featuring face swapping appeared on the Internet [6]. Since then, the technology of deepfakes has grown from a simple technological novelty to a full-blown tool for deception. Modern deepfakes allow for the replication of a person's look, voice, behavioral characteristics so accurately that the untrained eye cannot spot the difference between artificial and real media [7].

The introduction of Generative Adversarial Networks (GANs) for the first time happened in 2014 by Goodfellow et al. [8]. This technology innovatively uses the deep learning architecture of adversarial framework whereby the generative model aims to create convincing images of objects in contrast with the discriminative model that tries to identify the fake images created by the generative model (see Figure 1). By employing this iterative competition, the generative model acquires the necessary skills and knowledge such as the surface texture, lighting consistency, and spatio-temporal correlation [9, 10]. Besides that, as the implementation of encoder-decoder structures and convolutional feature extraction methods appeared, even more advanced techniques of face-swapping and facial reenactment whereby the features can be changed with no damage to the spatial structure and with minimum probability that certain evidence of the manipulation is traceable have appeared [11–13].

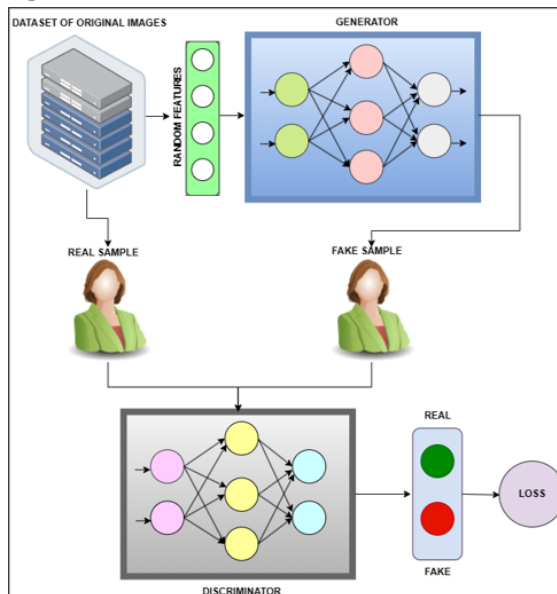


Figure 1: Function of Generative adversarial network

The proliferation of open-source platforms and modular deepfake systems has significantly eased the process of making synthetic media [14, 15]. As a result, individuals are no longer just passive observers of edited content; rather, they have learned to expect such content as a regular part of their online lives. This process of normalization is apparent in the Indian context. An analysis conducted by Pi-Labs called Digital Deception Epidemic: 2024 Report on Deepfake Fraud's Toll on India [16] shows that the number of

incidents related to deepfakes has risen 550% since 2019, and losses caused by them may reach around ₹70,000 crore in 2024. Deepfake-related fraud makes up nearly 40% of all cases of cybercrime associated with AI in India. In response to this problem, India's Ministry of Electronics and Information Technology (MeitY) has released official advisories obligating digital media platforms to eliminate content that has been flagged for deepfake within 36 hours after a formal appeal has been made [17]. The issue of deepfake technology and content has been clearly illustrated by the high-profile incident involving a deepfake video depicting Rashmika Mandanna that became viral in November 2023 and various politically-modified videos appearing before the 2024 general elections held in the country (Figure 2).

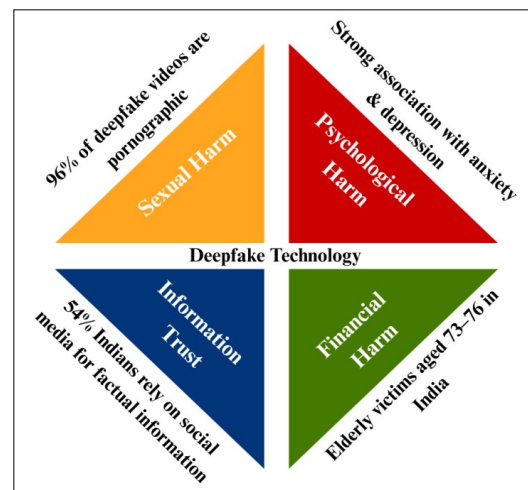


Figure 2: Societal impact dimensions of deepfake technology in India

Although the systems that utilize Convolutional Neural Networks (CNNs) and Transformer architectures for detection purposes have become more sophisticated, they still have limited availability in the real world and are virtually absent from mainstream platforms such as Facebook, Instagram, and Twitter/X, where fake media can be found [12, 18]. As such, average users have to rely on visual heuristics to judge the authenticity of media and this isn't an effective approach. In a study conducted by Nightingale and Farid [19], it was demonstrated that AI-generated faces were often perceived to be more trustworthy than the real ones and that the participants were correct only 48.2% of the time when judging which images were real and which were synthetic. A similar study by Groh et al. [18] involving 15,016 people has shown that human observers as well as the AI detection systems have shown similar levels

of error when classifying deepfake materials. These studies confirm that, as long as there are no verification tools, human perception is not sufficient for the reliable identification of high-quality synthetic media.

Despite the increasing amount of available evidence, the human detection capacity itself has not been studied in detail as compared to the detection technologies. The majority of the existing studies concentrate on either the performance of a particular algorithm or on samples derived from Western countries so the results cannot be generalized to such countries as India, where digital literacy, social media habits, and AI-assisted content are different. Furthermore, there are not many studies that investigate how particular demographic features and previous experience with deepfake technology influence decisions made by people and how they react to AI-assisted misinformation in underdeveloped digital economies.

Cognitive biases influence how people judge and trust digital media. Vaccari and Chadwick [20] discovered that exposure to political deepfakes may not necessarily lead to deception but rather creates epistemic uncertainty that reduces people's trust in legitimate information sources. Other studies show that detecting such content is related to the previous exposure to AI-generated material, indicating that cognitive processing, as well as lived experiences, determine their susceptibility to synthetic media [18, 19].

Overall, these results indicate that little study has been done on how well untrained people can detect deepfakes on their own in the South and Southeast Asian contexts. Because past studies have focused on English-speaking populations in the West, the findings cannot be generalized to India due to differences in social media usage and behavior, digital literacy levels, and cultural practices regarding AI technology. The current research aims at providing the first empirical evidence on deepfake detection capabilities of people living in North India by looking at not only the general level of accuracy, but also at the demographic and cognitive determinants of this capability. It was hypothesized that overall accuracy would be low and that it would differ depending on demographic and behavioral variables, like age, social media engagement, occupation, and knowledge of deepfake technology.

To this end, the study will address the following research questions:

RQ1: What percentage of individuals living in Delhi/NCR and/or Punjab will be able to determine if an image or video is real or a deepfake without the use of any technical means?

RQ2: What specific perceptual cues or cognitive heuristics do people use to evaluate whether or not digital media is real?

RQ3: How does people's prior exposure to deepfake technology and use of social media impact the accuracy with which they can determine whether an image/video is a deepfake?

The research outlined in this paper has numerous contributions to the novel field of deepfake detection. The first contribution is the development of a systematic method to assess the ability of individuals to distinguish between real and fake media under controlled circumstances. The second contribution is that the relationships between the demographic characteristics of subjects, their behavior on social media, their knowledge of deepfakes, and their detection performance were identified. And thirdly, with the help of SDT, DPT, and HSM, the different perceptual cues and cognitive strategies that people utilize while judging authenticity are described in detail. The last contribution is the emphasis on the limitations of human perception in this area and justification for combining human technology and AI-based detection methods in future research.

As a result of contributing to the developing empirical literature on human exposure to AI-based misinformation, the present study provides the practical recommendations for the digital media policy makers, public awareness programs, and digital forensics training in the Indian context.

2. MATERIALS AND METHODS

2.1 Study Objective

The focus of this research is to study the extent to which people can recognize AI-generated deepfakes by studying participants' responses (Figure 3).

The research has three main goals: the first one is to evaluate how well the experiment participants can identify real media from AI-generated ones; the second goal is to analyze whether the detection rates differ based on demographic factors, previous experience with deepfakes, occupation, and frequency of using social media. The third objective is to identify visual and cognitive cues that people use to assess whether or not a video has been manipulated, including facial expression, eye movements, mouth, and teeth shape, detail in the background, and quality of media content. The research is conducted among people from India and aims to analyze the difficulties that individuals face when identifying artificial media content.

The proposed model underscores the importance of organized experimental testing over the traditional methods used for perception research, whereby the same set of media stimuli is equally presented to all participants in order to lessen the possible impact of outside variables on responses.

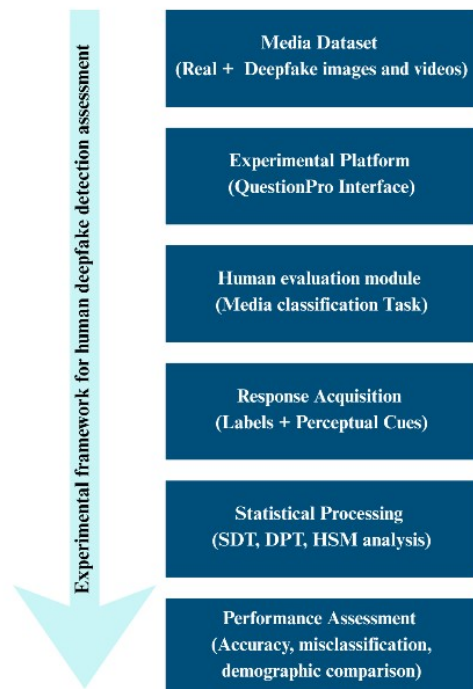


Figure 3. Experimental framework for human deepfake detection assessment

2.2 Experimental design and media classification framework

The experimental framework was created for the purpose of evaluating participants' aptitude

for differentiating authentic media from media generated by AI and determining the key visual and cognitive indicators that were used in their classification. Previous studies showed that people tend to make the distinction between original and modified content based on visual signals and cognitive heuristics [6, 20]. Hence, the framework was developed using established theoretical frameworks and incorporated into the study in three stages (Figure 4): (1) Conceptual design and media sourcing; (2) Assessment Platform Development and deployment; and (3) Participant recruitment and Data Collection.

The framework for conducting the study is standardized so that the methodology remains consistent. In particular, the presentation of media items, collecting responses, grouping participants, and processing data were done according to specific protocols that eliminate bias and ensure reliability.

Figure 4 depicts a comprehensive process used in this study. The framework merges theory, media development, participant involvement, the gathering of responses, and data analysis.

Stage 1: Conceptual Design and Media Selection

This theoretical framework uses Signal Detection Theory (SDT), a famous psychological theory used to know how humans can differentiate between noise and signal under conditions of uncertainty [21], which is appropriate for understanding how to tell manipulated media from real media. The research work is based upon the notion of real media as a signal and manipulated media as noise, as the participants have to classify the media in uncertain conditions. Prior research shows that human accuracy in this domain is usually quite near to random guessing [18].

Participants in the study needed to label the media production being shown as real or deepfake. The media was divided into groups according to type, containing four real photos, four deepfake images, one real video, and one deepfake footage, all of which is from publicly accessible databases available for research and academic purposes. The use of both video and image types of media enables evaluating the effectiveness of the detection, as the processes of the creation of videos and photos differ [12].

The design of the task was additionally influenced by Dual-Process Theory (DPT) [22], which makes a distinction between a quick and intuitive way of making judgments (System 1) and a slow and analytical way of processing (System 2). This model is particularly pertinent for deepfake classification since people usually make a ground swift intuitive judgment before proceeding to a thorough evaluation. Previous studies indicate that first impressions govern our judgment of digital media frequently, especially in the context of social networks [20]. Consequently, the current research investigated what cues people rely on while making their decision, such as unusual facial expression, inconsistent movement of the eyes or mouth, appearance of unusual visual artifacts, background inconsistencies, low picture quality, gut feeling, and so on. Demographic data was also collected, including age, gender, occupation, education background, frequency of social media usage, as well as prior experience with deepfakes [23].

Stage 2: Assessment Platform Development and deployment

The assessment form was created by using QuestionPro, an online tool for taking surveys that provides respondents with different types of answers and allows different types of media to be used in the survey. These platforms are usually employed for research on perceptions and decision-making since they allow controlling the presentation of stimuli and recording answers [18].

The development of the survey was guided by the Heuristic-Systematic Model (HSM) [24], according to which people rely on heuristics, such as their appearance, realism, and context affinity when they do not have enough cognitive resources or a high level of motivation to conduct systematic evaluation. In the case of deepfake identification, the heuristics used include facial expression differences, artificial features, contextual differences, and lighting issues. The respondents could select from the list of heuristics the ones used in their evaluation, such as facial expression differences, abnormalities in mouth or eyes, problems with teeth, context differences, image quality problems, unnatural features, intuition, and others. This approach reflects the Source Credibility Theory, according to which the evaluation of content authenticity depends on the perception of the realism of its source [25].

Stage 3: Participant recruitment and Data Collection

The method of recruitment comprised online media, social media portals, chat applications, and academic mediums. Recruitment was restricted to the areas of Delhi-NCR and Punjab due to high internet utilization, smartphone availability, and social media exposure [26, 27]. For each form of media, respondents were asked to express their opinions using a three-point rating scale (Agree / Disagree / Not Sure), consistent with standard direct-judgment response formats used in SDT-based research to capture response confidence [21].

The assessment platform was first used in a pilot study phase to validate that (1) all multimedia elements worked properly, (2) the also easy to use, and (3) the instructions clearly stated what to do. Respondents took part in the study voluntarily and provided informed consent before taking part.

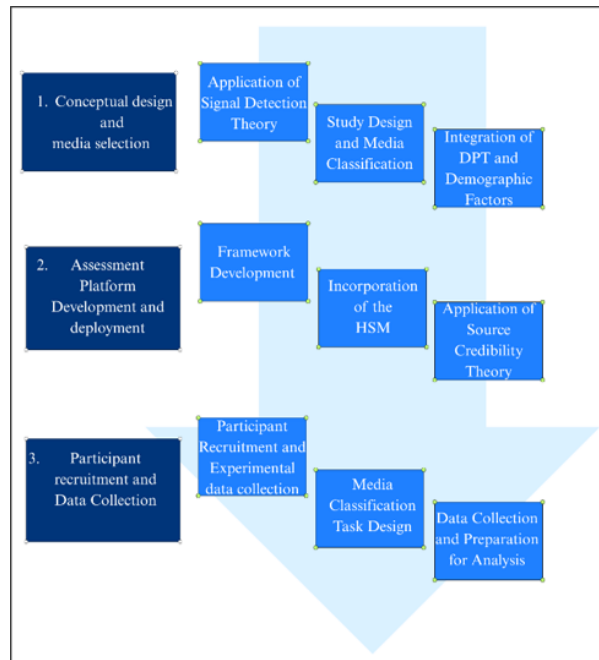


Figure 4. Phases of Experimental design and media classification framework development

QuestionPro was utilized to collect the data automatically and to extract the data for the purpose of conducting statistical analysis. Information was then categorized based on set criteria, including demographic details, previous experience of deepfakes, social media habits, and type of media in order to allow for assessment and

analysis of detection results according to behavioral and demographic measures. Evaluation process included descriptive statistics and cross-tabulations in order to make comparisons regarding the accuracy of classification taking into account different demographics and behaviors as per the guidance of the SDT-based framework implemented in the design of the study.

3. DATA ANALYSIS AND RESULTS

3.1 Descriptive Statistics and Bivariate Relationships

A group of 357 people assessed various media samples, which comprised four real images, four deepfake images, one real video, and one deepfake video. The evaluation of each sample was made public through the use of three options: "Agree", "Disagree", and "Neither Agree nor Disagree".

Subjects showed significant issue in differentiating between authentic and deepfake content. Detection rates of deepfake media types were roughly 34.8% to 72.4% based on demographic differences, while that of authentic media was somewhat similar, around 30.7% to 77.5%. This indicates that an individual factor is what contributes to differences in detection performance. Participants' detection ability appears to be affected by the Signal Detection Theory, where subjects see a high level of perceptual overlap. In addition, this results in unusually high false-positive and false-negative rates.

The overall descriptive statistics reveal that the performance of participants varies considerably, indicating that the success in detecting deepfakes depends on a range of factors rather than on a specific demographic factor. Therefore, the following analysis looks into the effects of each factor individually.

3.2 Comparison of Perceived Authenticity between Manipulated and Authentic Media

Misclassification was prevalent in both groups, regardless of media type, with 53.8% for authentic images and 54.1% for deepfake images respectively. Among those in the "Other" occupations, misclassification of authentic media reached 56.9%. This shows that there is significant ambiguity in the participants' judgments concerning visual authenticity irrespective of media type. The high levels of misclassification indicate that the

judges are depending on heuristic judgment instead of detailed analysis which is in line with SDT since it predicts that as the differentiability of the signal and noise decreases, the accuracy of judgment declines.

3.3 Identification Performance across Demographic Groups

3.3.1. Age

The detection accuracy improved with participants' age. People 35 years and older were the most accurate in detecting both authentic (76.2%) and deepfake (68.9%) media and those in the youngest group of 18-24 years had the least accuracy in detecting deepfake media (54.2%) which suggests that the older participants analyze and process digital content more effectively, which is in accordance with DPT which proposes that older adults do much more analytical processing (System 2) than intuitive (System 1) processing.

Table 1. Accuracy as per age group

Age Group	Correct Deepfake (%)	Correct Real (%)
18-24	54.20	69.00
25-34	63.50	72.80
35+	68.90	76.20

3.3.2. Occupation

Similarly, there was variation in detection accuracy based on type of occupation (Table 2). Those who worked in academia or research scored highest in terms of correct identification of deepfakes (47.4%), followed by professionals (41.3%), and students (34.8%) and those in the "Other" category had a lower rate (36.2%). The same pattern could be observed when authentic media was identified, where the "Other" category had the highest rate of misidentification (56.9%). This is probably due to differences in levels of analytical engagement and exposure to digital content among participants, which fits into the overall high rates of errors presented in Section 3.2.

Table 2. Accuracy as per occupation

Occupation	Correct DF (%)	Incorrect DF (%)	Correct (%)	Incorrect (%)
Student	34.80	54.10	30.70	53.80
Working Professional	41.30	46.20	40.50	46.70
Academic /Research	47.40	44.80	42.10	47.40
Others	36.20	55.20	31.00	56.90

3.3.3. Familiarity with Deepfake Technology

The study found a solid association between people's knowledge about DeepFake Technology (DFT) and their capability to recognize DFTs. Those participants who knew the most about DFTs had the highest percentage rate (72.4%) in being correct when recognizing DFTs, representing an improvement of 28.6 % greater than participants who did not have much understanding about DFTs (43.8%) as shown in table 3. The results are also strongly supported by SDT, which suggests that previous knowledge enhances an individual's ability to recognize differences between 'signals' vs. 'noise' for a specific area of expertise.

Table 3. Accuracy as per familiarity with deepfake

Familiarity Level	Correct DF (%)	Correct Real (%)
Low	43.80	61.50
Medium	58.60	69.30
High	72.40	75.60

3.3.4. Accuracy as per gender

The accuracy of the detections, according to gender, was somewhat different but it differed insignificantly. The male respondents demonstrated a slightly more accurate identification of

manipulated media (61.5%) and authentic content (71.0%) than the female respondents (54.3% and 65.7%, respectively). The respondents who categorized themselves as "other/prefer not to say" had an average performance (58.0 and 66.2%). The gender effect is still observable through these results.

Table 4: Accuracy as per gender

Gender	Correct DF (%)	Correct Real (%)
Male	61.50	71.00
Female	54.30	65.70
Other/Prefer not	58.00	66.20

3.3.5. Social Media Usage

A positive influence of social media activity on the detection success was discovered as well. Participants who used social media activities with high frequency managed to identify 77.5% of not manipulated contents and 69.3% of manipulated media content, meanwhile the low-activity users managed to recognize 60.4% and 49.1% (as shown in table 5), respectively. It indicates that a greater experience with the digital media provides a better understanding of visual imagery and the content identification success.

Table 5: Accuracy by social media activity level

Activity Level	Correct DF (%)	Correct Real (%)
Low	49.10	60.40
Moderate	59.70	70.10
High	69.30	77.50

3.3.6. Perceptual Indicators Used by Participants

The participants pointed out the particular visual cues that were used for making authenticity assessments (as shown in Table 6). Media quality problems (64.0%) were cited most often, followed by facial expression abnormalities (58.8%), suggesting that those taking part in the research relied on the surface visual abnormalities in evaluating the information they received. Interestingly, 21.1% of answers were based solely on intuition ("gut feeling") and did not involve a specific recognizable clue, which illustrates the usefulness of the DPT concept of quick intuitive (System 1) thinking. The apparent inconsistency in

error rates and the reliance on intuitive thinking point to the low useful value of this type of quick reasoning for detecting deepfake content.

Table 6: *Perceptual cues used in authenticity judgments*

Category	Description	Percentage (%)
Image / Video Quality Issues	Poor quality, blur, pixelation, distortion, low resolution	64.00
Facial Expressions & Emotions	Unnatural, weak, awkward, mismatched expressions	58.80
Background & Lighting Issues	Fake background, lighting/shadow mismatch, blending issues	40.40
Eyes Anomalies	Weird eyes, gaze, eye shape, blinking, edited eyes	33.30
Gut Feeling / Intuition	Guess, instinct, "seems fake/real", no clear reason	21.10
Face-Body / Neck Mismatch	Misalignment of face, neck, head, body proportions	22.80
Movement / Lip-Sync Issues	Unnatural movement, lip-sync mismatch (videos)	16.70
General Unrealistic Details	"Looked unreal", artifacts, strange features (non-specific)	19.30
Clear / Natural Appearance (Real)	Expressions/background looked natural or real	13.20
No Explanation / Null Response	“.”, “None”, “Nil”, “Idk”, blank	14.00

4. DISCUSSION

The results provide evidence on the inadequacies of human perception in judging AI media. Participants adopted different perception methods while making judgments but the levels of performance in finding genuine content were very low overall. This confirms worries about the pace of generative AI's development ahead of people's ability to detect synthetic media. The research offers valuable information on how credible AI-generated content is perceived by regular people and how hard it is to make the distinction between real and fake content.

4.1 Human Limitations in Detecting Deepfakes

The research gives significant information regarding the ways through which people recognize the originality of media created by AI. The findings suggest that most participants find it complicated to tell if the media is authentic or altered, especially since the artificial information closely looks like that of a real person. One of the main observations is that the detection accuracy among different groups is low, indicating that being unaware of whether something is original or altered is a general problem. Such a conclusion is in line with the earlier studies showing that people have difficulties distinguishing artificial from original information [18,19]. As the differences in the artificial media are decreasing due to the advances in the generation technologies, the trust in the senses will likely also diminish and thus people will have to rely on hints that are not always correct. This situation has clear consequences for the reality where one makes quick judgments about authenticity without previous checking (for example, social media users posting content) as no one can be sure that the attentive viewer has overcome the perceptual gap.

The human evaluation in this case includes a certain amount of uncertainty, mainly because altered media tends to look similar to the true representations [27]: the perceptual signal and noise almost completely overlap with each other. As a result, the evaluators must categorize the analyzed content while being unsure, which creates problems with the application of objective assessment criteria and leads to high numbers of false-positive and false-negative decisions [21]. The data obtained indicates that continuing progress made in creating deepfakes reduces efficiency of traditional visual evaluation, making one wonder whether human perceptual analyzation will be enough for ensuring reliable verification of digital evidence in important

areas, including media, law enforcement, and forensic science.

4.2 The Role of Familiarity and Digital Exposure

In general, the participants that knew about deep fake technology proved to be more accurate in identifying manipulated content, which highlights the importance of technological literacy and media education for critical digital engagement with different digital content [20]. People versed in deep fake methods seem to be more inclined with treating different digital media in a skeptical way and being less torn away in evaluating the media to distinguish between fake and original content as a result.

Constant engagement with different types of digital media seems to enhance visual experience that aids in recognition of images. Those who have a habit of using digital media on a regular basis are able to learn how to identify a visual object better [23]. Nevertheless, being affected by the exposure leads to dependency on automatic behavior and indicates the necessity of critical media consumption rather than passive one. The fact that there is a connection between familiarity level and efficiency of detection proves that learning through exposure leads to growth in perceptual abilities, but not completely. The initial effect tells about the possibility of examining familiarity from the point of view of it being just a part of the wider concept of media literacy.

4.3 The 'Liar's Dividend' and Declining Trust in Visual Media

Respondents tend to confuse real media with fake media quite often, indicating a pattern of uncertainty in how people evaluate digital media per se, and not just synthetic media. This is likely because of the general uncertainty that users have in determining whether visual information is real or not when consumed online. As deepfake technology starts producing cheaper and better photorealistic deepfakes, any visual media, including true content, will become subject to greater scrutiny, which Chesney and Citron [7] refer to as the "Liar's Dividend," such that the existence of synthetic content already causes doubts about true evidence. This loss of default confidence in visual media brings more trouble than just individual misjudgments, as it puts those conditions when true evidence may be regarded as false, which

has ramifications not only for the use of visual evidence in court but also for journalism and political communication.

4.4 Cognitive Decision-Making and Perceptual Strategies

The perceptual cues mentioned by the participants act as one of the sources of understanding regarding the cognitive process involved in deepfake detection. People followed three main signals, especially; flaws in image quality, inconsistencies in emotions depicted by the actors, and discrepancies in the background. From this aspect they are similar to the DPT. According to DPT, people tend to use heuristic approaches in cases when the subject matter is visually complicated and verifiable only through extensive effort [22].

Many participants appeared to form an initial judgment through intuitive, System 1 processing, before some engaged in more analytical, System 2 evaluation, for instance, identifying specific irregularities in lighting or facial structure. This combination of intuitive and analytical thinking works in accordance with the previous perceptual studies of deepfake recognition [6]. The fact that both analytical and intuitive processing systems are simultaneously at work indicates that mistakes in deepfake recognition do not necessarily imply that the participant has not applied analytical thinking.

4.5 Implications for Media Literacy and Detection Technologies

The findings indicate that there is a pressing need for more public education on AI-generated media and for educational programs that address this issue. Interestingly, participants who were well informed about deepfake technology were unable to differentiate between authentic and manipulated content by visual judgment alone. Thus, in order to deal with challenges posed by deepfake technology, it will be essential to pair human perceptual judgment with automated detection systems, especially considering the evidence that machine learning-based detection systems might be able to detect small visual discrepancies that human beings might not [13]. Based on the findings of accuracy as per gender and perceptual indicators used by participants for deepfake identification, one can conclude that

literacy interventions and technological detection tools are compatible rather than interchangeable: literacy enhances initial judgment capability, while automated tools are needed to detect residual errors that still exist in most cases even among the educated users.

5. IMPLICATIONS

5.1 Digital Forensic Implications

The results provide a reference point for human capability in detecting deepfakes, informing efforts to create hybrid human-AI systems in forensic technology (such as tools for decision-making, frameworks for explainable deepfake detection, and workflows for forensic triage). The research produces further evidence of the ability of humans to classify real and manipulated media in a precise laboratory setting. The specific patterns among different demographics, including the differences between weak and experienced participants, suggest that the forensic triage process should be focused on the need for independent verification of media suspected to be used by the inexperienced users. The findings suggest that relying purely on humans is not effective as generative technologies are developing rapidly, and manipulation will become more sophisticated [18, 19]. A clear contribution of this research lies in its evidence regarding the relationship between the experience of users in social media, the knowledge of deepfakes, and the ability to detect them.

5.2 Research Implications

This research provides a methodological approach that could provide insight into future studies that would look into the influence of deepfake on skepticism, trust, and decision-making. This study provides data whereby participants are shown various forms of media with different types of visual clues and impression signals such as image quality, facial expressiveness, and attributes of background. The data is very useful in future research on how people process and interpret visual information and also learn about the cognitive strategies they use to evaluate the media content that has undergone changes. This research can also be used in the process of identifying population groups that are more at risk of being misled by deepfake.

5.3 Practical and Educational Implications

The results obtained from this study present certain ideas that are relevant not only from theoretical perspective but also for the field of media literacy and comprehension of artificial media and misinformation spread. The proliferation of synthetic media leads to an increase in the number of cases in which people get deceived. The data concerning perceptual cues presented in Table 6 shows that participants relied on media quality (64.0%) and facial expression (58.8%) cues more than on those cues that would be more informative, such as lip-sync or alignment of movements. Thus, it would be beneficial for the training programs to teach individuals to rely more on such cues. This research provides information for the development of training programs, educational materials, and campaigns aimed at raising awareness. Educators and policymakers can benefit from understanding that people tend to make decisions based on appearance. In this light, it seems that combining approaches that rely on both technological advancement and the human-centered approach is the most effective way of addressing problems brought about by deepfake media.

6. CONCLUSION

The purpose of this investigation was to determine how good participants are at distinguishing between genuine and AI-produced media using an experiment that was controlled. The result is that mediating identification is much harder than expected, and the ability to identify accurately depends on the participants' involvement with digital media and also on how well they understand deepfake technology. This work increases knowledge of the struggles faced by people not only in identifying deepfakes but also in dealing with the transforming technology of digital media, thus paving the way for prospective findings that would create solutions for the current issue.

This is an important benchmark study focusing on human perceptual performance rather than computer detection algorithms. The experimental approach developed in this study can be applied to other groups and other generations of Artificial Intelligence, as well as different forms of digital forgery, supporting continued progress in digital forensics and human-machine collaborative detection.

7. LIMITATIONS

The selection of participants was restricted to only The participant recruitment process was restricted to two locations, namely Delhi-NCR and Punjab. Thus, this may limit the generalizability of the results in other parts of India or possibly beyond. In addition to that, the study was based entirely on human observation in a controlled laboratory setting and did not use any forensic detection tools for the identification of deepfakes. Furthermore, the study concentrated on a small number of selected images and videos instead of a larger one which would make it applicable in more contexts.

8. FUTURE RESEARCH DIRECTIONS

The experimental design can be extended to study larger and more diverse groups of people by involving more types of AI generated images and videos, such as multimodal deepfakes. Moreover, if this approach is combined with AI forensics technologies, this will lead to the emergence of better human-AI cooperation for deepfake detection.

REFERENCES:

- [1]. Figueira Á, Oliveira L. The current state of fake news: challenges and opportunities. *Procedia computer science*. 2017 Jan 1;121:817-25.
- [2]. Borges L, Martins B, Calado P. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*. 2019 Jun 26;11(3):1-26.
- [3]. Qayyum A, Qadir J, Janjua MU, Sher F. Using blockchain to rein in the new post-truth world and check the spread of fake news. *IT Professional*. 2019 Jul 15;21(4):16-24.
- [4]. Anderson J, Rainie L. The future of truth and misinformation online.
- [5]. Britt MA, Rouet JF, Blaum D, Millis K. A reasoned approach to dealing with fake news. *Policy Insights from the Behavioral and Brain Sciences*. 2019 Mar;6(1):94-101.
- [6]. Westerlund M. The emergence of deepfake technology: A review. *Technology innovation management review*. 2019;9(11).
- [7]. Chesney B, Citron D. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*. 2019;107:1753.
- [8]. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*. 2014;27.
- [9]. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019* (pp. 4401-4410).
- [10]. Mirsky Y, Lee W. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*. 2021 Jan 2;54(1):1-41.
- [11]. Zhang T. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*. 2022 Feb;81(5):6259-76.
- [12]. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion*. 2020 Dec 1;64:131-48.
- [13]. Verdoliva L. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*. 2020 Jun 12;14(5):910-32.
- [14]. Swathi P, Sk S. Deepfake creation and detection: A survey. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) 2021 Sep 2* (pp. 584-588). IEEE.
- [15]. Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC. Deepfakes: Trick or treat?. *Business horizons*. 2020 Mar 1;63(2):135-46.
- [16]. Pi-Labs, "Digital Deception Epidemic: 2024 Report on Deepfake Fraud's Toll on India", Pi-Labs, Dec. 2024.
- [17]. Ministry of Electronics and Information Technology (MeitY), Government of India, "Union Government issues advisory to social media intermediaries to identify misinformation and deepfakes", Press Information Bureau, New Delhi, India, Nov. 7, 2023.
- [18]. Groh M, Epstein Z, Firestone C, Picard R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*. 2022 Jan 4;119(1):e2110013119.
- [19]. Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the*

- National Academy of Sciences. 2022 Feb 22;119(8):e2120481119.
- [20]. Vaccari C, Chadwick A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*. 2020 Feb;6(1):2056305120903408.
- [21]. Green DM, Swets JA. *Signal detection theory and psychophysics*. New York: Wiley; 1966 Jan 1.
- [22]. Watson K. D. Kahneman.(2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux. 499 pages. *Canadian Journal of Program Evaluation*. 2011 Sep;26(2):111-3.
- [23]. Guess A, Nagler J, Tucker J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances*. 2019 Jan 9;5(1):eaau4586.
- [24]. Chaiken S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*. 1980 Nov;39(5):752.
- [25]. Hovland CI, Weiss W. The influence of source credibility on communication effectiveness. *Public opinion quarterly*. 1951 Jan 1;15(4):635-50.
- [26]. Madhusudhan M. Use of social networking sites by research scholars of the University of Delhi: A study. *The International Information & Library Review*. 2012 Jun 1;44(2):100-13.
- [27]. Jain N, Jindal V, "Users' perception towards smartphone applications in Delhi NCR: Demographic perspective," *ShodhKosh: Journal of Visual and Performing Arts*, vol. 4, no. 2, pp. 4326-4332, 2023.