

Enhancing Human Safety through Modesty Guard: A Comprehensive Review of OpenCV-Based Surveillance Systems

Abstract—The increasing deployment of reconnaissance organizations in civic and isolated places has raised concerns about privacy, ethics, and cultural sensitivities, particularly when it comes to the unintended capture of nudity or other immodest imagery. In diverse environments where cultural norms dictate specific standards of modesty, there is a growing need for systems that can automatically detect and manage nudity in surveillance footage without infringing on individuals' rights or exposing sensitive material unnecessarily. This system utilizes artificial intelligence (AI) and computer vision to detect nudity in real-time video feeds, offering customizable responses such as blurring, alerting security personnel, or securely archiving sensitive footage. This paper outlines the technical architecture, deployment strategies, and deep learning algorithms of the Modesty Guard system, emphasizing its potential applications in various environments where surveillance and modesty concerns intersect. Through the use of advanced AI and real-time video processing, the system promotes ethical surveillance practices while respecting cultural diversity and privacy.

Keywords— Neural networks, deep learning, ML, AI technology, and surveillance systems.

I. INTRODUCTION

With the increasing prevalence of reconnaissance organizations in both civic and isolated places, concerns about privacy, ethics, and cultural sensitivities have become more pronounced. Surveillance cameras, while essential for maintaining security and monitoring activities, often capture individuals in situations that may unintentionally expose them to privacy violations. One particularly sensitive issue is the unintentional recording of nudity or immodest imagery. In diverse societies, where cultural and religious norms around modesty vary, the inadvertent capture of such footage can lead to ethical dilemmas, public outcry, and even legal challenges. From a social science perspective, nudity can mean different things. In a narrow sense, nudity simply means that people are without clothes. In a broader perspective, the feeling of being naked comprises a mostly negative affect according to which people feel embarrassed, exposed, or unprotected for some reason without having control. This feeling is accompanied by the fear that personal privacy and intimacy limits are threatened. In the context of technology-assisted living

environments, this has been reported in different usage situations, but evidently most often whenever camera technologies were involved. Traditionally, surveillance systems focus on security, with limited attention paid to the ethical implications of what is being recorded. However, as societies become more interconnected and diverse, there is a growing demand for systems that not only maintain safety and security but also respect individual privacy and cultural values. This includes the need for technology that can detect and appropriately handle nudity or immodest imagery in surveillance footage, preventing exposure that could be considered offensive or inappropriate in specific contexts. In the computer vision community, plenty of detecting strategies, skin color models and scientific foundations have been proposed and researched for years to detect and recognize the pornographic nature of multimedia files. The first attempt of screening pornographic images used skin region segmentation as a pre-processing step before finding elongated regions for possible limbs.

II. RELATED WORK

Jamshid Bacha, et.al.,...[1] provided a system that uses the gathered KAU-Memes dataset 2582 to identify objectionable substance in comments and stop such annoyance from being shared on social media platform. The latter blends the freshly created memes from a number of repugnant and distasteful tweet datasets with the "2016. U.S. Election" dataset collection. Actually, the suggested model is validated using the KAU-Comments dataset, which comprise symbolic pictures and the accompanying text. We evaluate how well three recommended deep learning algorithms perform in terms of training and identifying objectionable content in memes. According to the authors' understanding and examination of the literature, this is the first method for detecting inappropriate content in memes that is from on You Only Look Once (YOLO). According to discover, For SSD-MobileNet V2, the developed model achieved 81.74% and 84.1% mAP and F1-score, while for YOLOv4, it achieved 85.20%, 84.0% mAP and F1-score. With the highest mAP, F1-score, precision, and recall—88.50%, 88.8%, 90.2%, and 87.5%, respectively—YOLOv5 showed the best description, respectively. This method suggests a new paradigm for improved offensive material identification in

unstructured data across diverse social media platforms. The recently evaluate framework was methodically practical to two iterations of YOLO and SSD Mobile Net with modify settings in order to assess accuracy and speed.

Sonali Samal, et.al,...[2] A database of categorised terms is produced here, and the words are then check for any offensive words. If the user says any vulgar terms, it will be sent to the Watchlists, that will filter those phrases out. Recovering model performance, reducing the computational load of the wanted model, and subtracting HP tuning are all made possible by the proposed TL-based attribute fusion progression. FFP transfers the learned data to control the classification process after combining low-level and mid-level images of the top-stage trained models. The creation of a tagged obscene picture dataset is one of the main achievements of our suggested approach. The Pix-2-Pix GAN architecture is used by GGOI to train deep learning models. To achieve training stability, ii) model architectures are modified by incorporating batch normalisation and a mixed pooling strategy; iii) A database of categorised terms is fashioned here, and the words are then chequered for any distasteful words. If the user says any vulgar terms, it will be sent to the Watchlists, that will filter those phrases out. By performing end-to-end image detection, outperforming models are selected to be integrated with the FFP; and IV) a TL-based picture finding technique is created by retraining the fused model's last layer. After thorough experimental analyses of benchmark datasets like the NPDI and Pornography 2k, the GGOI dataset is created. With average classification accuracy, comprehension, and F1 score of 98.50%, 98.46%, and 98.49%, respectively, the process TL model with fused MobileNet + DenseNet network outperforms active techniques.

Ishaani Priyadarshiniet.al,...[3] created necessitates the use of techniques that will demonstrate enhanced performance and shorter model creation times, as suggested in previous proposals to solve the issue. A database of part terms is formed here, and the words are then checked for any unpleasant words. If the user says any vulgar terms, it will be sent to the Watchlists that will filter that phrase out. . In order to detect hate speech on social media, we suggest a transport learning approach that uses a pre-trained model for data testing, hence increasing the reusability of the model. The routines of our suggested model are compared to those of Naive Bayes, Decision Trees, and Support Vector Machines (SVM), which are also the study's baseline methods, using the unigram and bigram language designs. We recommend two transfer learning models: the GloVe Model, which uses LSTM for the same purpose, and Google's Word2vec model,

which uses LSTM. Metrics like predict, recall, F-1 score, and support are used to validate the effectiveness of the proposed models for categorizing hate speech, offensive speech, and neutral speech. The overall accuracy of the models' performance on several datasets has been assessed. Comprehensive experimental study and findings show that the suggested models perform the baseline algorithms under consideration and are notably resilient for identifying hostile and offensive speech.

Chandradeep Bhatt et.al,...[4] created only for people who are not known with social media. The current research employs machine learning techniques to automatically identify the bullies' harsh word usage to stop this harassment. Following detection, if a child was bullied, it could use the child's dataset to determine the bullying and, in the unlikely event that the child was the victim of bullying, it would be determined from the dataset. If any abusive language is found, the developers will be alerted, the appropriate action will be taken, and an email alert will be sent if any abusive text is found in the conversation. Thus, the suggested approach is effective in identifying cyberbullying on social media and can be turned into a web application that requires users to link their social media profiles. Both parents who want to know what is happening with their children and the children themselves will find this project helpful. Since a lot of children develop social media profiles to stay up to date with their schooling, this idea will undoubtedly be helpful. Additionally, as everything is now done online, we can use this effort to stop bullying people online. My goal is to identify cyberbully words and their types and to send out an SMS alert if any are found. Consequently, we selected six machine learning techniques for classification and used count Vectorizer and time frequency - inverse manuscript frequency to extract features as a bag of words. I employed Naïve Bayes, K-NN, RF, SVM, Decision Trees, and Logistic Regression. When temporal frequency-inverse document frequency is used to extract features, support vector machines generate 91.98%.

Felipe Gonzalez-Pizarroet.al,...[5] The proposed study uses OpenAI's CLIP to do a multimodal examination of anti-Semitism and Islam phobia on 4chan's /pol/, which advances research efforts to identify and comprehend toxic speech on the Web. The Contrastive Learning methodology is used by this sizable pre-trained model. Using Google's Perspective API and human explanation, we develop a mechanism to identify a set of nasty textual expressions that are antisemitic and Islamophobic. Next, we find photos that closely resemble our antisemitic/Islamophobic textual words using OpenAI's CLIP. Our draw near detects 173K posts with 21K antisemitic/Islam phobic

pictures and 246K posts with 420 nasty statements by analysing a dataset of 66M post and 5.8M descriptions uploaded on 4chan's /pol/ over an 18-month period. We discover, among other things, that OpenAI's CLIP model has a precision tally of 0.81 (F1 score = 0.54) for identifying hostile material. We find that CLIP performs better than two baselines put out by the literature in provisions of precision, accuracy, and F1 score when it comes to identifying antisemitic and Islam phobic pictures. The necessity to develop additional techniques for identifying hateful imagery is

further highlighted by the fact that antisemitic/Islam phobic metaphors is posted now a comparable figure of pillars on 4chan's /pol/ as opposed to anti-Semitic/Islam phobic stylistic terms. Lastly, in order to help academics improved understand anti-Semitism and Islam phobia, we offer easily reached (upon demand) a dataset of 246K postings that surround 420 anti-Semitic/Islam phobic statements and 21K maybe anti-Semitic/Islam phobic pictures (spontaneously recognised by CLASP).

Table 1: Literature survey table

| S.NO | TITLE | TECHNIQUE | MERITS | DEMERITS |
|------|---|-----------------------------|--|---|
| 1 | A Deep Learning-Based Framework for Offensive Text Detection Unstructured Data for Heterogeneous social media | YOLO | Heterogeneous social media refers to different platforms with varying data types, formats, and languages. | This can result in high operational costs, particularly when scaling for large social media platforms. |
| 2 | Obscene image detection using transfer learning & feature fusion | Deep learning | This is beneficial for rapid deployment. | This could result in biases or underperformance when dealing with niche or rare obscene content types. |
| 3 | A transfer learning approach for detecting offensive & hate speech on social media platforms | LSTM | This reduces the time and computational resources required for training compared to building a model from scratch | The quality and biases of pre-trained models significantly affect the final performance |
| 4 | Detection of cyber-bullying in social-media using classification algorithms of machine learning | Machine Learning | This is particularly valuable for platforms with a high user base, where manual moderation would be impractical. | Classification algorithms may struggle with false positives (flagging non-bullying content as bullying) and false negatives (failing to detect actual cyberbullying). |
| 5 | Understanding and Detecting Hateful Content Using Contrastive Learning | CLIP | These models can capture the nuances of offensive language in free-form text, regardless of its structure or format. | This could lead to ethical dilemmas regarding user data privacy and surveillance. |
| 6 | Offensive Language Detection in Spanish Social Media: Testing | Natural language processing | This project explores multiple approaches, from simpler | Transformer models, while powerful, require significant |

| | | | | |
|----|---|-------------------------|---|---|
| | From Bag-of-Words Transformers Models | | technique like Bag-of-Words (BoW) to more advanced models like transformers | computational resources. |
| 7 | SOLD: Sinhala offensive language dataset | Deep learning | This makes the dataset valuable for researchers and developers who want to work on hate speech and unpleasant language detection in lesser-resourced languages. | Unlike English and other widely spoken languages, Sinhala lacks extensive NLP resources , such as pre-trained models and existing datasets . |
| 8 | Innovative deep learning techniques for monitoring aggressive behaviour in social media posts | Deep learning | These models can handle nuances in language, including sarcasm, slang, and contextual aggression , which are often missed by traditional methods. | This can be expensive, especially for smaller social media platforms or organizations with limited resources. |
| 9 | A multilingual offensive language detection technique based on transfer learning from transformer fine-tuning model | Transfer learning model | It helps determine which models are best suited for Spanish offensive language detection, making it easier for future projects to adopt the right | Smaller social media platforms or research teams with limited infrastructure may find it challenging to implement these models efficiently. |
| 10 | Classification of Virtual Harassment on Social Networks in Ensemble Learning Techniques | K Nearest Neighbour KNN | This can lead to better detection of virtual harassment, capturing more complex patterns that single models might miss. | Ensemble methods are generally more complex to implement and tune than single-model approaches. |

III. BACKGROUND OF THE WORK

Several existing algorithms and techniques have been developed for nudity detection and image processing, leveraging machine learning, computer vision, and image analysis. These algorithms can be categorized based on their move towards and underlying methodologies.

A. Skin Color Detection

Histogram-Based Approach: This method utilizes color histograms to identify skin-tone pixels in an image. It analyzes the distribution of colors in the RGB or HSV

color space and applies predefined thresholds to classify pixels as skin or non-skin.

- **Gaussian Mixture Models (GMM):** GMMs model skin color distributions by creating a statistical representation of skin tones. The algorithm identifies pixels that fit within the skin color distribution, effectively distinguishing skin from non-skin regions.

B. EDGE DETECTION ALGORITHMS.

- **Clever authority Indicator:** This unearthing algorithm is a more step process that detects edges in images by identifying areas of rapid intensity change. It involves grade calculation, non-maximum containment, double thresholding, and edge tracking. This algorithm helps delineate body contours, making it easier to identify shapes associated with nudity.
- **Sobel Filter:** The Sobel filter applies convolution with Sobel kernels to detect edges in images. It highlights the horizontal and vertical gradients, enabling the identification of prominent features and boundaries in the image, which can be useful for nudity detection.

C. ML ALGORITHMS

- **Support Vector Machines:** SVMs are supervised learning classical expenditure for classification tasks. In nudity detection, SVMs can be trained on character extracted from images (e.g., color histograms, texture description) to classify images as containing nudity or not. The SVM constructs a hyperplane in a multi-dimensional space to disconnect different classes.
- **Principal Component Analysis (PCA):** Reduces dimensionality while retaining important features, helping to improve the efficiency and effectiveness of models.
- **Texture Analysis:** Techniques like Local Binary Patterns (LBP) can capture textural patterns that distinguish nudity from non-nudity images.

IV. PROPOSED WORK

Convolutional Neural Networks algorithm is a powerful deep learning architecture widely used for image cataloguing everyday jobs, counting nudity detection. A typical CNN consists of several layers that automatically learn hierarchical description from images. The process begins with an input layer that receives the raw image data, which is then process through multiple convolutional layers. These layers are valid convolutional filters to the input, detecting specific features like edges and textures, and generating feature maps that highlight the presence of these learned features. Foundation functions, such as ReLU (Rectified Linear Unit), initiate non-linearity, enable the network to model hard patterns. In order to decrease computational burden and mitigate overfitting, pooling layers—which frequently use max pooling—lower the altitudinal portions of the quality atlases though preserve the utmost crucial information. The final

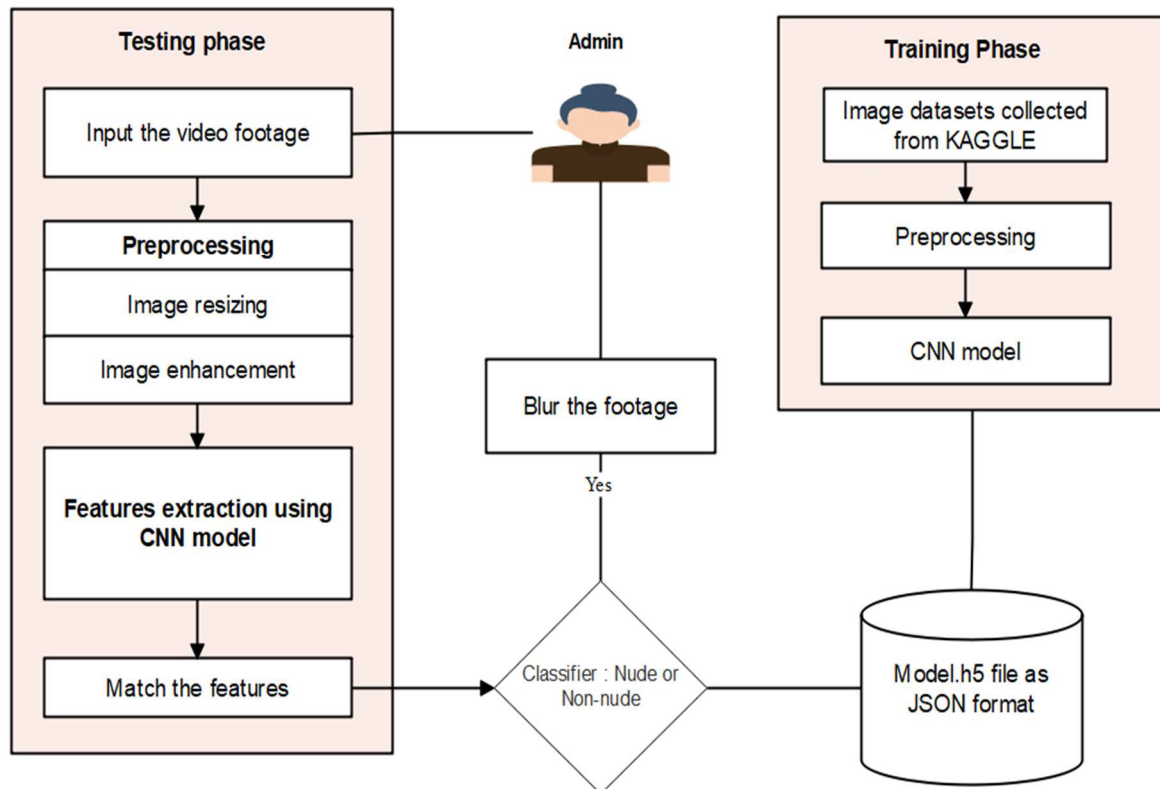
classification is then carried out by absolutely attached layers that flatten the high-level characteristics. Usually, a sigmoid occupation is used for binary classification and a softmax function for multi-class categorization. Training a CNN for nudity detection involves several steps: first, a diverse and well-labeled dataset of images, containing both nude and non-nude content, must be prepared and augmented to enhance variability. The dataset is divorced into test, validation, and training sets. Training set is utilised to back propagate across several epochs in order to minimise a loss function. While the final technique is assessed on the test set using method like correctness, precision, recall, and F1-score, model performance is tracked on the validation set to adjust hyper parameters and avoid overfitting. CNNs can efficiently categorise pictures and provide reliable nudity identification in a variety of scenarios by utilising these strategies. **Data Collection and Preprocessing:** The system begins by collecting a dataset of nude and non-nude images from various sources. These images are preprocessed to standardize their size, format, and orientation. Preprocessing techniques may include resizing, cropping, normalization, and augmentation to enhance the dataset's quality and variability. **Feature Extraction:** Once preprocessed, the images undergo feature extraction to capture relevant information for image classification. This involves analyzing the images to identify discriminative patterns, textures, and structures associated with different skin types. Feature extraction techniques may include traditional image processing algorithms or deep learning-based methods to extract high-dimensional feature representations. **Model Development:** The extract features are used to train a machine learning form for image classification. The system may employ deep learning architectures such as Convolutional neural network algorithm known for their effectiveness in image classification tasks. The representation is trained using labeled data, with an emphasis on optimizing classification performance and generalization.

Model Evaluation and Validation: The accomplished classical is calculated and validated expending disconnects test datasets to measure its enactment and robustness. The model's classification presentation is measured using evaluation measures such area under curve (AUC), sensitivity, specificity, and accuracy. Techniques for cross-validation may also be used to guarantee the accuracy of the findings. **Integration and Deployment:** Once validated, the trained model is integrated into the system for practical use. This may involve developing a user-

friendly interface or application that allows professionals to upload video footages and obtain classification results in real-time. If the footage

contains nudity content means, automatically the blur footage and send alert to admin. Figure 1 shows the proposed work

Fig. 1. Proposed Framework



V.INVESTIGATIONAL RESULTS

A database of categorised terms is shaped here, and the words are then chequered for any offensive words. If the user says any improper terms, it will be sent to



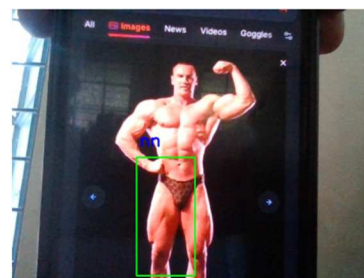
| Image class | Sample image | Image count |
|-------------|---|-------------|
| Non-nude |  | 2000 |
| Nude class |  | 2000 |

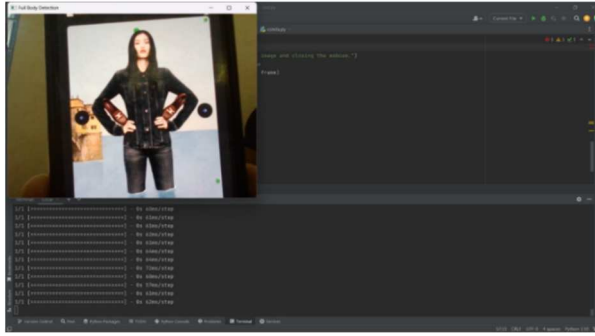
Table. 2. Sample Imagery table the Watchlists, that will filter those phrases out . This study allows us to examine two classes, such as nude and non-nude, by entering picture data in the form of JPG designs that are unruffle. Sample imagery in datasets is shown in the above table Extract the body features and build the model using CNN algorithm. The result shows in fig 2.

Fig 2:
layer



construction

CNN



Presentation metrics like precision, problem-solving, correctness, empathy, and specificity can endure without being creative in analyzing the system's evolution.

- True positive (TP): Total number of absolute positive predictions
- The number of false positives, or faulty positive detection, is known as the false positive (FP).
- Number of true negatives, or perfect negative predictions, is known as the true negative (TN).
- The number of genuine negatives resulting from faulty negative detection is known as the false negative (FN).
- Accuracy: Accuracy (ACC) is the percentage of flawless forecasts to the entire test data. The range is 0.0 at the lowest point and 1.0 at the highest point.

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} \times 100$$

| ALGORITHM | ACCURACY |
|------------------------------|----------|
| PRINCIPAL COMPONENT ANALYSIS | 50% |
| SUPPORT VECTOR MACHINE | 65% |
| CONVOLUTIONAL NEURAL NETWORK | 80% |

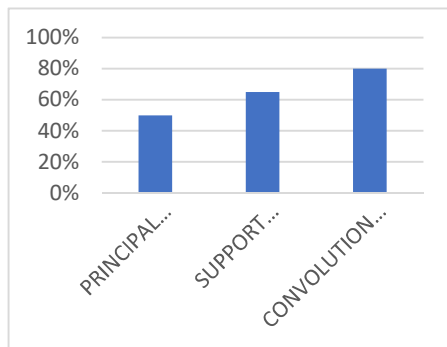


Fig 3: Accuracy rate

The preceding graph demonstrates that the suggested CNN approach outperforms the lack of a trend towards

The training accuracy can be shown in fig 4

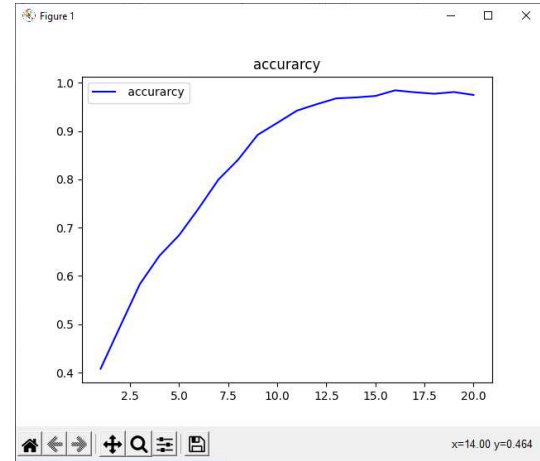


Fig 4: Training accuracy

VI. CONCLUSION

The development of a Modesty Guard System utilizing CNN-based classification for nudity detection represent an important development in the fields of surveillance and image processing. Using Convolutional Neural Networks Effectively, the system can learn and recognize complex patterns associated with nudity by analysing visual content from various sources. The combination of robust preprocessing techniques, diverse and well-structured datasets, and state-of-the-machine learning algorithms allows for accurate categorization and real-time detection, addressing the critical need for privacy and cultural sensitivity in public and digital spaces. As concerns regarding nudity and immodesty continue to grow in various contexts, implementing a reliable nudity detection system not only enhances security measures but also promotes a respectful and considerate environment. Future advancements in CNN architectures, the effectiveness and dependability of nudity detection systems will be further enhanced by transfer learning and hybrid techniques. By prioritizing ethical considerations and ongoing improvements in technology, the Modesty Guard System can serve as an essential tool in maintaining appropriate standards in visual content across multiple platforms

REFERENCES

- [1] Bacha, Jamshid, et al. "A deep learning-based framework for offensive text detection in unstructured data for heterogeneous social media." IEEE Access 11 (2023): 124484-124498.

- [2] Samal, Sonali, et al. "Obscene image detection using transfer learning and feature fusion." *Multimedia Tools and Applications* 82.19 (2023): 28739-28767.
- [3] Priyadarshini, Ishaani, Sandipan Sahu, and Raghvendra Kumar. "A transfer learning approach for detecting offensive and hate speech on social media platforms." *Multimedia Tools and Applications* 82.18 (2023): 27473-27499.
- [4] Kumar, Vishal & Bhatt, Chandradeep & Goyal, Parul & Dubey, Ghanshyam & Singh, Shalini. (2024). DETECTION OF CYBER-BULLYING IN SOCIAL-MEDIA USING CLASSIFICATION ALGORITHMS OF MACHINE LEARNING. *Community practitioner: the journal of the Community Practitioners' & Health Visitors' Association*. 21. 12. 10.5281/zenodo.11203380.
- [5] González-Pizarro, Felipe, and Savvas Zannettou. "Understanding and detecting hateful content using contrastive learning." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 17. 2023.
- [6] Molero, José María, et al. "Offensive Language Detection in Spanish Social Media: Testing From Bag-of-Words to Transformers Models." *IEEE Access* (2023).
- [7] Ranasinghe, Tharindu, et al. "Sold: Sinhala offensive language dataset." *Language Resources and Evaluation* (2024): 1-41.
- [8] Han, Huimin, et al. "Innovative deep learning techniques for monitoring aggressive behavior in social media posts." *Journal of Cloud Computing* 13.1 (2024): 19.
- [9] El-Alami, Fatima-zahra, Said Ouatic El Alaoui, and Nouredine En Nahnahi. "A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model." *Journal of King Saud University-Computer and Information Sciences* 34.8 (2022): 6048-6056.
- [10] Azeez, NureniAyofe, and Emad Fadhal. "Classification of Virtual harassment on social networks using ensemble learning techniques." *Applied Sciences* 13.7 (2023): 4570.
- [11] Akyon, Fatih Cagatay, and Alptekin Temizel. "State-of-the-Art in Nudity Classification: A Comparative Analysis." *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2023.
- [12] Kim, Seungkwon, Sangyeon Kim, and Seung-Hun Nam. "A Framework for Portrait Stylization with Skin-Tone Awareness and Nudity Identification." *ICASSP 2024-2024*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.
- [13] AlDahoul, Nouar, et al. "Evaluation of Content Moderation Software for Nudity and Pornography Detection in Various Scenarios."
- [14] Riccio, Piera, Thomas Hofmann, and Nuria Oliver. "Exposed or Erased: Algorithmic Censorship of Nudity in Art." *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024.
- [15] Dewan, Ritu, et al. "A Nudity Detection Algorithm for Web-based Online Networking Platform." *2023 10th International Conference on Based Deep Learning System on Rank Images Classification.* *Computer Systems Science & Engineering* 47.2 (2023).
- [17] Momo, Mhd Adel, et al. "Evaluation of Convolution and Attention Networks for Nudity and Pornography Detection in Sketch Images." *2023 IEEE Symposium on Computers & Informatics (ISCI)*. IEEE, 2023.
- [18] Laddach, Agnieszka. "Against the Nudity in Art: Eliasian Reading of National Conservative Catholic Habitus." *Open Theology* 10.1 (2024): 20240003.
- [19] Dubettier, Adrien, et al. "A Comparative Study of Tools for Explicit Content Detection in Images." *2023 International Conference on Cyberworlds (CW)*. IEEE, 2023.
- [20] Rautela, Kamakshi, et al. "Obscenity detection transformer for detecting inappropriate contents from videos." *Multimedia Tools and Applications* 83.4 (2024): 10799-10814.