Proxy Tasks Ensemble for Explainable Inference in Sensitive Data

Clara Ernesto
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
claraernestodecarvalho@usp.br

Sandra Avila
Instituto de Computação
Universidade Estadual de Campinas (UNICAMP)
sandra@ic.unicamp.br

Carlos Caetano
Instituto de Computação
Universidade Estadual de Campinas (UNICAMP)
caetanoc@unicamp.br

Leo S. F. Ribeiro
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
leo.ribeiro@icmc.usp.br

***Abstract*—Child sexual abuse imagery (CSAI) classification inherits several challenges by its nature. Due to the limited access to training data and the highly sensitive nature of the images in question, most solutions are not reproducible, not distributable, not explainable, and vulnerable to attacks. In this context, we propose the Ensemble for Inference in Sensitive data using Proxy Tasks (EISP) framework to realize explainable inference in sensitive data, without having to train directly in target images, by using an ensemble of Proxy Tasks related to CSAI detection, such as nudity detection, age estimation, and perceived gender classification.**

If the EISP system is provided with Proxy Tasks that correlate to the target task, it can be optimized with feature combination, realize predictions about the given task and provide explainability both by creating data visualizations, and extracting SHAP feature importance through each prediction. In this Work In Progress paper, we test the framework and explain its inner works using a public non-CSAI dataset.

I. Introduction

Child sexual abuse imagery (CSAI) classification is a task in deep need of automation, due to the severe trauma inflicted upon the victims, and the heavy burden on law-enforcement agents, who are psychologically harmed as they manually check the material [1].

Child sexual abuse affects 9% to 19.7% of girls and 3% to 7.9% of boys, and it is manifested in different manners such as sex trafficking or rape [2, 3, 4]. In this context, the victims suffer with intense lasting consequences, as depression, post-traumatic stress, and lifelong injuries such as chronic pain, which are lengthened by the online exposition of their suffered violence.

CSAI classification tooling has evolved and been modernized. However, model development still faces technical, legal, and ethical challenges in many of its steps. New systems, created with these obstacles in mind, are needed for explainability and distribution.

The hardships of CSAI classification are inherited from the nature of the data. Training models directly on CSAI data poses a risk of model inversion [5], as sensitive image information is represented in the model parameters; consequently, the fitted models cannot be distributed to the general public. Moreover, CSAI datasets, which are highly restricted by the law, are only accessible to law enforcement agents (LEA). Due to this necessary restriction, all research in the field needs to collaborate with law enforcement, and training directly on CSAI is rarely viable. With these constraints, Proxy Tasks [6] were introduced to make for viable development, distribution, and reproduction while evading exposure of sensitive information. The *Proxy Tasks* are formalized as related sub-tasks of the target task which consider the constraints present in the target task context (in the case of CSAI, it is not possible to guarantee LEAs will have access to specialized hardware, etc.).

In this paper, we propose the Ensemble for Inference in Sensitive data using Proxy Tasks (EISP) framework for classification in sensitive data, utilizing selected proxy tasks and ensembling them with an XGBoost model [7]. Although we have not yet conducted tests on real CSAI, our current results focus on proxy tasks relevant to this scenario, such as nudity detection, perceived gender classification, and age estimation. Aside from the final lightweight and quick-to-train XGBoost model, the proxy tasks were trained on public data, and therefore each model is free from the harms of possible model inversions; this makes for proxy models that are distributable and reproducible. Moreover, the XGBoost is the one model that cannot be distributed, as it is indirectly trained in sensitive data, but it can be trained in a short time span and with low compute resources, which facilitates its use and development within the constraints of LEAs. As a bonus feature from this design, having a separate model for each proxy task allows us to observe within the EISP framework the importance of each task for the final classification, making this also an explainable solution, an essential characteristic for LEAs.

We show how our framework[1] has been refined, tested, and its results explored using the Image Sentiment Dataset [8], a task meant to represent the ambiguities and nuances of CSAI

[1]All developed code is available at https://github.com/clr-cera/EISPCSAI.

through a public dataset. Therefore, our contributions are:

1) Propose and formalize our framework, guided by the literature on CSAI classification.
2) Test and evaluate its inner workings in a public dataset, using a different task still related to the selected Proxy Tasks.
3) Show and explain visualization results of the dataset using the Proxy Tasks.
4) Show different possible optimizations, using feature combinations and dimension reduction techniques.

II. Related Work

Existing work on CSAI detection has largely followed two parallel paths. The first path relies on hash-based matching techniques, which, while effective against known material, are easily circumvented by minor alterations and offer no capabilities for generalization to novel content [9, 10]. The second path explores content-based methods, from classical computer vision pipelines [11] to deep representation learning [12]. However, given the high sensitivity of CSAI data, training such models directly remains heavily restricted.

The concept of Proxy Tasks [6] has emerged as a practical solution, allowing researchers to develop models on related problems such as age estimation [13], nudity detection [14, 15], and scene classification [16], before applying them cautiously to CSAI-related tasks.

In this context, the selection of Proxy Tasks was heavily inspired by the work of Laranjeira et al. [17], which analyzed a Brazilian CSAI dataset with several different models and tasks.

Whilst applying it to a distinct task, the study of Reis et al. [18] is the most similar to ours in its execution; the authors present a framework for combining hundreds of distinct shallow Natural Language Processing features within a single ensemble. Their solution experiments with different combinations of features and presents explainability metrics and visualizations that highlight which features are more or less relevant to their task of Fake News classification. We apply this concept to CSAI Proxy Tasks, with the main distinction being our target task and our use of deep-learning-based feature vectors.

III. Our Approach

The problem we aim to solve is to realize predictions on top of sensitive images without directly using them to train any model, and to explain and visualize how the system predicts its output. Moreover, we aim to utilize the EISP framework on CSAI, so the framework was proposed and will be tested with its constraints in mind. The problem solution the EISP framework propose can be formalized as the following:

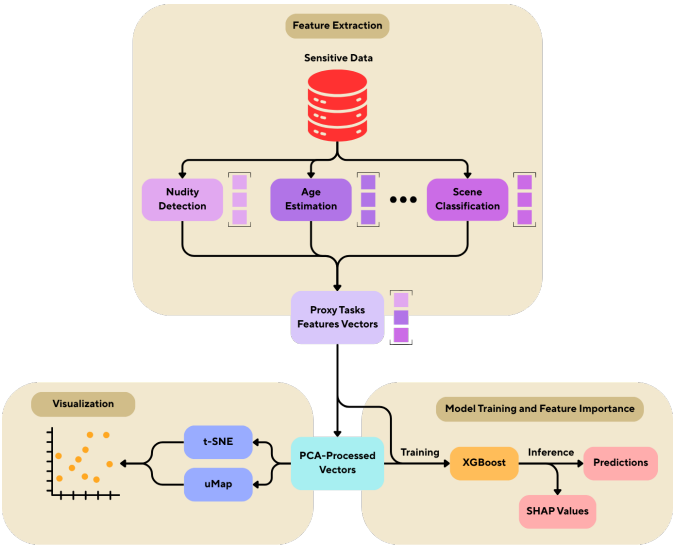

Fig. 1: Diagram of proposed EISP framework, consisting of three steps: Feature Extraction, Feature Space Visualization and Model Training and Feature Importance. In the Feature Extraction step, the sensitive data is processed by different Proxy Task models and their result is concatenated. After that the resultant feature vectors are processed by PCA. The PCA-processed vectors then go through the Visualization step, where they are input to the t-SNE and uMap techniques for projection onto a 2D plane. Parallel to this, the original feature vectors or the PCA-Processed vectors are used for training the XGBoost ensemble model, which is used for inference and produce SHAP Values and label predictions.

$$F : S \to L \tag{1}$$
$$PT : S \to FS \tag{2}$$
$$CFS = FS_0 \oplus FS_1 \oplus \ \ldots \ \oplus FS_n \tag{3}$$
$$CF : FS \to CFS \tag{4}$$
$$TI = \{(x, y) \mid x \in CFS, y \in L\} \tag{5}$$
$$ET : TI \to E \tag{6}$$
$$E = \{CFS \to L\} \tag{7}$$
$$s \subset S \tag{8}$$
$$F = ET((PT \circ CF)(s), L)(s) \tag{9}$$

where F is the prediction process, S is the sensitive data vector space, L is the labels vector space, PT is the process of extraction of each Proxy Task feature vector, FS is the vector space for each Proxy Task vector, CFS is the vector space generated by the concatenation CF of feature vectors, TI is the training input, ET is the ensemble training process, E is the ensemble function space, and s is a subset of S.

Given a set of pre-trained proxy task models, our proposed framework consists of three main parts, as shown in Figure 1: (i) Feature Extraction, (ii) Feature Space Visualization, and (iii) Model Training and Feature Importance.

For the first step, we have selected features that represent data related to CSAI. Our selection comprises models for age

estimation, perceived gender classification, nudity detection, scene classification, and object detection. This way, we can use these feature vectors to train the model, perform inference, and explain the model based on feature importance.

In the Feature Space Visualization step, we utilize visualization techniques to observe how the images are distributed and clustered according to their concatenated inferred features. This allows us to identify which features are most important for separating the images and analyze the dataset distribution.

Finally, we realize model training and inference using the feature vectors for the target task. For these preliminary results, we use XGBoost as the model, as it can be trained in a reasonable time span and has improved compatibility with SHAP values [19] for feature importance classification. Moreover, we train the model using combinations of the features to check if there is a combination that performs better or the same as using all vectors, and consequently, be able to select features to be dropped.

IV. Methodology

A. Dataset

We evaluated our approach on the Image Sentiment dataset [8]. It was built from disaster images on social media. A crowd-sourcing study annotated 4,003 images with 2,338 users, tagging them with emotions generated when viewing the images. The users answered five questions. The first two were numeric questions regarding emotion positivity and excitement, and the last three had multiple options to be selected, describing emotion keywords (Joy, Sadness, Fear, etc.) related to user evoked emotion, and image aspects that influenced feelings (Facial expressions, image color, image background, etc.). We expect it to be useful for testing the EISP framework and evaluating the selected features for CSAI detection, as the three last questions have options related to some of the selected Proxy Tasks. Even if the ensemble model is supposed to be a final solution for CSAI detection, testing this on a public dataset allows for a thorough evaluation of the framework itself before adding to the burden of partner LEAs. Testing on real CSAI is reserved as a final step and has not been concluded yet.

B. Feature Extraction

In Feature Extraction, we extract several feature vectors for each separate Proxy Task.

a) Face Detection: We first detect faces and extract face crops for use on the Age/Gender model and ITA value. For this task, we used the MTCNN model [20].

b) Age/Gender: The Age/Gender model [21] was trained for the task of estimating age and classifying the perceived gender of a face crop. We used the faces extracted from MTCNN, and averaged the resultant vectors, if there was no recognizable face, we initialized the vector as $\vec{0}$.

c) ITA value: The ITA value is computed from face landmark extraction [22, 23] using $ITA = \frac{\arctan(L-50)}{b} \times \frac{180}{\pi}$, where L and b are channels of the image in CIE-Lab space, respectively indicating luminance and amount of yellow. Therefore, the ITA value consists of a single $(1D)$ number, and if there are no recognizable faces it is initialized as 0.

d) Object Detection: For the Object Detection Proxy Task, we used YOLOv11x [24], which is a state-of-the-art model for the task. The vector is extracted from its backbone, as we want a general representation, but still related to the task.

e) Nudity Classification.: We used a pre-trained ViT Base model for Nudity Classification [25], where the vector is extracted as the first CLS token.

f) Scene Places: We used a pre-trained AlexNet model for Scene Classification [16] trained in the Places365 dataset, where we also extracted a feature vector from its architecture.

g) Scene Coelho et al.: We also utilized a Scene Classification model trained with CSAI constraints in mind, employing Few Shot techniques (from Coelho et al. [6]). This model was trained to output a feature vector for similarity in distance, facilitating feature extraction.

Age/Gender	ITA Skin Tone	Object	Nudity	Scene Places	Scene Coelho et al.
4096	1	768	768	256	384

TABLE I: Size of vector per feature.

We decided to use two models for Scene Classification to compare and test a model trained traditionally, and another trained using Few Shot Learning with the constraints of CSAI in mind.

After the extraction and concatenation, the resultant vector has size $6273D$, and each Proxy Task feature vector size can be seen in Table I. As the final vector has a high dimensionality, we also processed each feature separately with PCA.

C. Feature Space Visualization

We used t-SNE [26] and UMAP [27] to reduce the concatenated feature dimensionality to 2, so that we can plot and visualize the feature space. We also performed t-SNE analysis using each feature individually to visualize the image distribution per feature.

D. XGBoost Training and Performance

We trained XGBoost using both the original feature values and the PCA-processed feature values. In both trainings, we used regression for the first two questions and binary logistic for the last three. After training, we tested the models and stored the SHAP values for feature importance and their performance metrics (Mean Squared Error for regression and Balanced Accuracy for binary logistic). Finally, we trained the model using combinations of the PCA features, storing their performance, so that we can store which feature combinations achieve the best metrics, and consequently, which features can be dropped without a large impact on performance.

E. Implementation Details

We conducted the feature extraction and XGBoost training on one NVIDIA A5500. The training of proxy task models was explained in each of the referenced original papers. We trained the XGBoost model with `maximum depth=7` and `eta=0.1`.

F. Carbon Footprint

We conducted a carbon footprint analysis of our framework using the CodeCarbon tool [28], relating each task considered heavy on compute (our criterion was tasks that took more than 15 minutes to run) to its energy consumption and equivalent CO_2 emission. Each task processed 3678 images. (Table II).

Task	Energy (kWH)	CO_2-eq (kg)
Feature extraction	1.1376	0.1118
Training with feature combination	0.8059	0.0792

TABLE II: Average energy consumption and equivalent CO_2 emissions (CO_2-eq) for each compute-heavy step of our method.

V. Results and Analysis

A. Feature Extraction

For feature extraction, we processed all images with each model. Due to the high dimensionality of the final feature vector, one of our pipelines further processes the features with PCA, while maintaining a minimum of 80% of the variance. To keep dimensions consistent, we reduced each vector to a $256D$ projection (the smallest power of 2 size that still kept 80% of the variance for all features); the exception is the ITA skin tone feature that is already a single scalar value.

B. Feature Space Visualization

We applied both t-SNE and UMAP techniques to project the feature space onto two dimensions. Then we plotted it using colors to represent all annotated labels. We tested both plotting techniques with multiple hyperparameters; the final ones for t-SNE were: `perplexity=30`, and for UMAP were: `n_neighbors=50`, `min_dist=0.1`. The numeric questions are colored according to a heatmap of their values, and the boolean options from the multiple-choice questions are plotted separately. Both the t-SNE and UMAP plots similarly represented the clusters, and we present the t-SNE plots here.

In addition to visualizing the concatenated feature space, we have also plotted the space for each individual proxy task's vector to observe how images are distributed along each feature space. This visualization will help us understand the clusters in the original t-SNE.

From the visualization in Figure 2, we can observe three distinct parts of the dataset. On the right side, a cluster of images with recognizable faces, which may be attributed to the ITA value and age/gender feature vector, both of which depend on these recognizable faces. In the middle, we observe several small clusters; these clusters are explained by the behavior of the feature space from Coelho et al. [6]'s scene model, which we observed to also present several small clusters across all features. This suggests that the visualization techniques rely on this feature to group the data in clusters when there are no age/gender and ITA features.

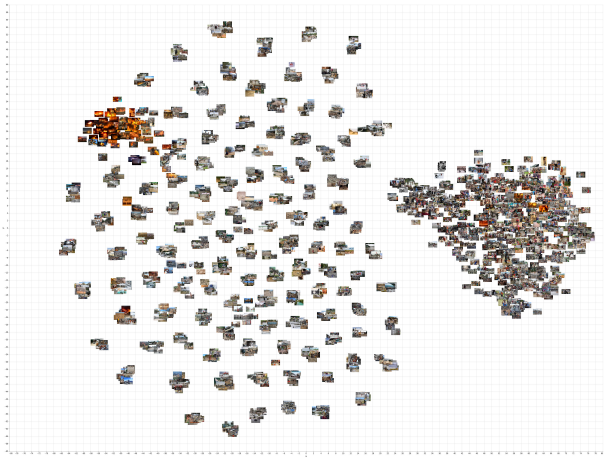

Fig. 2: t-SNE plot of Proxy Tasks feature vector space after PCA-processing, each point is rendered using its image. We recommend using zoom on a computer screen for this figure.

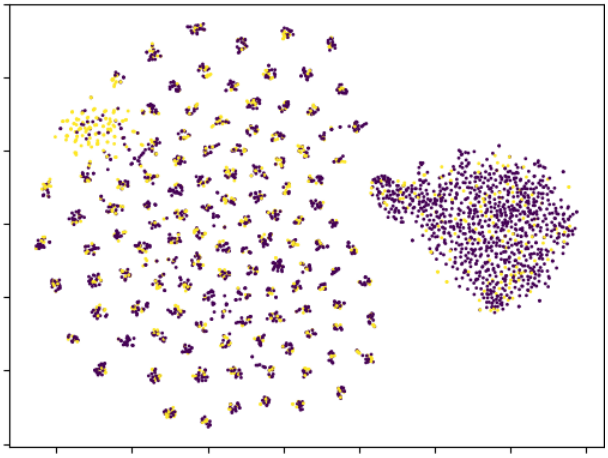

Fig. 3: t-SNE plot colored for question 3.3 label. Question 3 is a multi-option question asking which major keywords describe their evoked emotion after seeing the picture, and option 3 is the keyword *Fear*

Moreover, compared to the other individual feature plots, which do not present small clusters but an almost uniform distribution, we infer that Coelho et al. [6]'s scene model is extracting vectors that follow a more separate and discrete distribution, closer to the categorical output.

By examining the colored plots for each label, we can observe that in some cases, the feature space clearly separates images by their labels. This examination indicates that the data can be separated by identifying which cluster relates to each feature and label.

For example, as shown in Figure 3, which shows the t-SNE plot for question 3.3, which is a multi-option question asking which major keywords describe their evoked emotion after seeing the picture, and option 3 is the keyword *Fear*. We can observe in the plot that most images containing recognizable faces are not marked. In contrast, the leftmost cluster, which includes fire incidents, is mostly marked.

C. Model Training and Performance

When training the model, we did four different experiments. We used the original features, the PCA-processed features, different combinations of Proxy Tasks in original feature vectors, and different combinations of Proxy Tasks in PCA-processed feature vectors.

The Mean Square Errors (MSE) for numeric questions (Q1 and Q2) and the average balanced accuracy for multi-option questions (Q3, Q4, and Q5) are presented for each experiment. They are shown in Table III.

Experiment	Q1	Q2	Q3	Q4	Q5
Original Features	2.3207	1.2164	55.4%	52.9%	63.8%
PCA-Processed	2.5172	1.1661	54.2%	52.0%	62.1%
Combined Original	2.2535	1.1668	55.8%	53.5%	64.5%
Combined PCA	2.3130	1.1610	54.9%	53.0%	63.3%

TABLE III: Testing metrics for each of the five questions, for all experiments (all original features, all PCA-processed features, the best combination of original features and the best combination of PCA-processed features). The first two questions ask for numeric values related to positivity and excitement felt after seeing the image, are trained with regression and use Mean Squared Error (MSE) as a metric. The last three are multi-option questions with keywords to be selected related to the user-evoked emotion, and use balanced accuracy as a metric.

Overall, the performance scores are modest in each question average, due to most options not being related to the selected Proxy Tasks, this could possibly be improved with the selection of different tasks such as facial expression recognition or action classification.

The performance of the PCA-processed features was slightly inferior in most labels compared to the original ones. This can be explained to the dimension-reducing process, which results in smaller variances, leading to information loss.

However, when analyzing the most performing combination of PCA-processed features and the most performing combination of original features, both achieved similar metric improvements. Therefore, when optimized, both feature spaces can achieve better results with smaller inference time spans, while the PCA-processed features leads to even faster training and inference periods. However, the combination of features can lead to overfitting if the same optimizing behavior is not generalizable to other images. Additionally, as seen in Table IV, most of the best combinations only use three Proxy Tasks, and therefore the framework can be optimized to use far less feature vectors, which optimize training speed (if another training is needed), feature extraction time span on images to predict and inference periods.

Finally, we analyze its performance by each label. In the numeric ones, the RMSE was between 1 and 1.5. Since other inference works [29, 30] on the Image Sentiment Dataset employ different tasks, we lack a baseline for comparing RMSE metrics and analyzing the model's performance on numeric questions.

Feature Vectors	Q1	Q2	Q3	Q4	Q5
Original Features	N	INP	IONPC	AION	ION
PCA-Processed	ONC	AIONC	AON	ON	IN

TABLE IV: Best Proxy Task feature vectors combination, on both original vectors and PCA-processed vectors, for each of the five questions. The metrics used to determine the best combination were the same as those used in Table III. The combinations are described as the concatenation of letters relating to each Proxy Task, encoded as Age/Gender model (A), ITA (I), Object detection (O), Nudity detection (N), Scene model (P) and Coelho et al. Scene model (C).

By observing the multi-label options, we can see that several options have 50% of balanced accuracy. Therefore, the model cannot infer them using the selected features because the Proxy Tasks lack correlation with the specific options; alternatively, other sets of Proxy Tasks can be tested. In other ones, we can see higher values, reaching 83%. By reading the label's description and observing the visualization plots of the feature space, we can understand how the model is inferring the specific option.

As an example, at the second question, where the user is asked for a value in the Likert scale that represents its stimulation after seeing the image, the best models using the original features are ITA value, Nudity and Scene Classification. As there are several images of wildfires, or natural disasters, the Scene Classification model can relate this information to higher stimulation.

D. Feature Importance

In this last step, we calculate the aggregated Shapley values for each Proxy Task on each prediction, and observe how important each feature is for their inference. When analyzing labels that had 50% balanced accuracy, their SHAP values are not relevant because the model is not fitted to the desired objective. However, by examining the proportion of aggregated SHAP values for each feature, we can identify which features are more important, how they vary between labels, and how they differ between PCA-processed features and original features.

For an example, as shown in Figure 4, in question 5.1, which asks the user to mark the kind of information that influenced the most their evoked emotion, and option 1 describes: *Human facial expression, pose or gesture*; the XGBoost model has its output most influenced, in descending order, by the Nudity

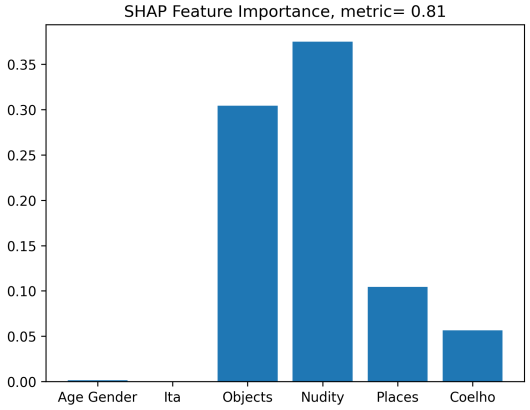

Fig. 4: Aggregated Shapley value per Proxy Task in question 5.1. Question 5 asks the user to mark the kind of information that influenced the most their evoked emotion, and option 1 describes: *Human facial expression, pose or gesture*.

detection model, the Object detection model, the Places 360 Scene model, and the Coelho et al. [6] Scene model.

Through this analysis, we found that most labels yielded high SHAP values for Nudity and Object models. These models also appeared in the most combinations with the best performance values, indicating that these two features are contributing the most to most label predictions.

VI. Conclusions and Next Steps

In this work, we proposed the EISP framework, a novel system for explainable inference in sensitive data. By using an ensemble of proxy tasks, our method can realize inference and provide explainability surrounding the sensitive data and target task. As observed on test results, if the framework is provided with Proxy Tasks that correlate to the target task, it can predict and explain both the data with its visualization and its inner workings with feature importance. Besides that, it can be trained with feature combination to optimize metrics and inference speed.

Future work will focus on testing EISP on CSAI, evaluating its performance, and analyzing its explanations. Moreover, we intend to make the framework available as a Python library, generalized to different kinds of target tasks.

References

[1] E. Bursztein, E. Clarke *et al.*, "Rethinking the detection of child sexual abuse imagery on the internet," in *The World Wide Web Conference*, 2019.

[2] J. Barth, L. Bermetz *et al.*, "The current prevalence of child sexual abuse worldwide: A systematic review and meta-analysis," *International Journal of Public Health*, 2013.

[3] N. Pereda, G. Guilera *et al.*, "The prevalence of child sexual abuse in community and student samples: A meta-analysis," *Clinical Psychology Review*, 2009.

[4] M. Stoltenborgh, M. H. van Ijzendoorn *et al.*, "A global perspective on child sexual abuse: Meta-analysis of prevalence around the world," *Child Maltreatment*, 2011.

[5] Y. Zhang, R. Jia *et al.*, "The secret revealer: Generative model-inversion attacks against deep neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1911.07135

[6] T. Coelho, L. S. F. Ribeiro *et al.*, "Minimizing risk through minimizing model-data interaction: A protocol for relying on proxy tasks when designing child sexual abuse imagery detection models," in *ACM Conference on Fairness, Accountability, and Transparency*, 2025, p. 1543–1553. [Online]. Available: https://doi.org/10.1145/3715275.3732102

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 785–794. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[8] S. Z. Hassan, K. Ahmad *et al.*, "Visual sentiment analysis from disaster images in social media," 2020.

[9] B. Westlake, M. Bouchard, and R. Frank, "Comparing methods for detecting child exploitation content online," in *European Intelligence and Security Informatics Conference*, 2012.

[10] Microsoft, "Tackling child sexual abuse strategy." 2009.

[11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, 2003, pp. 1470–1477.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[13] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep EXpectation of Apparent Age from a Single Image," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: IEEE, Dec. 2015, pp. 252–257.

[14] M. Polastro and P. Eleuterio, "Nudetective: A forensic tool to help combat child pornography through automatic nudity detection," in *Workshops on Database and Expert Systems Applications*, 2010.

[15] ——, "A statistical approach for identifying videos of child pornography at crime scenes," in *International Conference on Availability, Reliability and Security*, 2012.

[16] B. Zhou, A. Lapedriza *et al.*, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[17] C. Laranjeira, J. Macedo *et al.*, "Seeing without looking: Analysis pipeline for child sexual abuse datasets," 2022. [Online]. Available: https://arxiv.org/abs/2204.14110

[18] J. C. S. Reis, A. Correia *et al.*, "Explainable machine learning for fake news detection," in *ACM Conference on Web Science*, 2019, p. 17–26. [Online]. Available: https://doi.org/10.1145/3292522.3326027

[19] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[20] I. de Paz Centeno, "ipazc/mtcnn: v1.0.0," Oct. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13901379

[21] J. Macedo, F. Costa, and J. A. dos Santos, "A Benchmark Methodology for Child Pornography Detection," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2018, pp. 455–462.

[22] M. Merler, N. Ratha *et al.*, "Diversity in faces," 2019. [Online]. Available: https://arxiv.org/abs/1901.10436

[23] N. M. Kinyanjui, T. Odonga *et al.*, "Fairness of classifiers across skin tones in dermatology," in *Medical Image Computing and Computer Assisted Intervention*, 2020, p. 320–329. [Online]. Available: https://doi.org/10.1007/978-3-030-59725-2_31

[24] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," jan 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[25] AdamCodd, "Vit base nsfw detector," https://huggingface.co/AdamCodd/vit-base-nsfw-detector, accessed: 2025-08-10.

[26] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[27] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[28] K. Lottick, S. Susai *et al.*, "Energy usage reports: Environmental awareness as part of algorithmic accountability," 2019. [Online]. Available: https://arxiv.org/abs/1911.08354

[29] A. Pournaras, N. Gkalelis *et al.*, "Combining multiple deep-learning-based image features for visual sentiment analysis," Dec. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.6655366

[30] K. Ahmad, M. A. Ayub *et al.*, "Deep models for visual sentiment analysis of disaster-related multimedia content," 2021. [Online]. Available: https://arxiv.org/abs/2112.12060