MedGore: An Approach and a Dataset for Identification of Sensitive Medical Images

Soumya Gayen, Rory Mulcahey, Russell Loane, Dina Demner-Fushman, Deepak Gupta
National Library of Medicine, NIH, HHS
{gayens,rory.mulcahey,russell.loane,ddemner,guptadk}@nih.gov

Abstract

Medical images are invaluable in illustrating health issues for the patients. While biomedical publications are a good source of such images, some of the images are not appropriate for the patient viewing without a warning. To enable development of automated tools for selection of patient-safe images and generation of warnings, we created a dataset MedGore of over 78,000 sensitive medical images and 183,000 non-sensitive images published in the biomedical literature. The sensitive content includes gore, severe disease, nudity, surgical openings, internal organs, and other medical images of this nature. The set of the manually identified seed 300 images was expanded using a combination of human curation and a nearest neighbor clustering algorithm. The quality of the automatically labeled images was evaluated manually, yielding a total of more than 4,000 doubly-manually annotated images. The automatically labeled images proved to approach the utility of the manually labeled images for training the models in our experiments that validated the dataset in the task of labeling unseen images using the image features, the figure captions or both.

Keywords: multi-modal image labeling, sensitive medical images, patient-oriented image labeling

1. Introduction

The need for a dataset that effectively identifies and manages sensitive medical images is becoming critical in the realm of digital health technology. Current resources are inadequate in providing tools for content moderation of explicit content, leaving gaps in medical image repositories. Our dataset developed using images from the Open-i image search engine (Demner-Fushman et al., 2012) addresses these shortcomings by offering a large collection of over 78,000 images deemed sensitive for a non-medical professional general user.

General public usually experiences discomfort with specific types of content, especially images displaying severe trauma, surgical details, bodily fluids, internal organs, nudity, and other images of this nature, refer to Figure 1. This discomfort underscores the urgency of developing this dataset. To develop the dataset, we used the k-Nearest Neighbors clustering algorithm (Gupta et al., 2023) that allowed us to quickly manually review images grouped together as sensitive or safe.

Informal comparative analysis with existing datasets and content moderation tools, such as Amazon Rekognition [1], Microsoft Azure AI [2], and other AI-based solutions for detecting explicit content[3], revealed that while these tools provide robust general-purpose content moderation, they lack the specificity needed for effective moderation of medical images. For example, Amazon Rekognition and Azure AI are highly effective in identifying adult content and other explicit material, but these tools do not capture the nuances of medical image content. Additionally, resources like the GitHub repository on content-moderation-deep-learning[4] list various datasets and tools for detecting nudity, violence and other sensitive media, yet they lack the specificity required for medical images.

Our approach, inspired by solutions in similar domains such as gore classification and censoring (Larocque, 2021) and adult content image recognition (Karamizadeh et al., 2023), marks a novel contribution by specifically targeting the filtering of peer-reviewed medical images for public consumption. These efforts have resulted in a dataset that not only enhances content moderation capabilities but also provides a valuable resource for developers and researchers aiming to implement similar tools and novel algorithms for detecting sensitive medical images.

While datasets narrowly focused on specific images related to surgical (Carstens et al., 2023), laparoscopic (Ríos et al., 2023), or endoscopic (Borgli et al., 2020) procedures exist, there is no comprehensive dataset designed for the purpose of

[1] https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html

[2] https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview

[3] https://hivemoderation.com/

[4] https://github.com/fcakyon/content-moderation-deep-learning (Akyon and Temizel, 2023)

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 296–306
12 May 2026.

moderating explicit medical images for public consumption. The contributions of this work, therefore, are: 1) a large dataset that fills the critical gap of having various medical images marked as sensitive or not; 2) an approach to automatic labeling of images leveraging a small manually annotated seed set, and 3) establishing a strong benchmark for sensitive image identification using text (if available) and image features.

Figure 1: The images shown are blurred representations of graphic or sensitive content, including but not limited to severe trauma, surgical procedures, exposed internal organs, bodily fluids, and nudity. The complete, unaltered dataset contains more explicit material that may cause psychological distress.

2. `MedGore` Dataset

The medical images were sourced from the Open Access Subset of PubMed Central (PMC) using the Open-i search engine. The PMC Open Access Subset includes only materials that may be freely downloaded, reused, redistributed, and publicly shared under the applicable open-access license. The search restricted to photographs yielded 788,472 images.

An image was labeled as Sensitive or Gore if it depicted visible blood, tissue, organs, surgical procedures, trauma, burns, amputations, infections, or other content that could be disturbing outside a clinical or medical context. Images lacking such explicit or invasive visual content were labeled as Non-Sensitive or Normal. The curated seed images served as the foundation for automated dataset expansion using similarity-based retrieval methods, ensuring high-quality and semantically consistent samples across both categories.

2.1. Dataset Generation

Starting with manually curated set of 300 images per class, additional images were retrieved using 1 to 10 nearest neighbors from a K-Nearest Neighbors (KNN) search method (Gupta et al., 2023) . The KNN retrieval was performed separately for the Non-Sensitive and Sensitive categories. To ensure diversity and balance, the resulting images were further organized by their corresponding clinical journals. From these, 1,000 images per class were randomly sampled, proportionally weighted by the number of images available per journal. The selected 2,000 images were then manually annotated using an in-house image annotation tool with three labels — Blur, Maybe, and Ok — where annotators determined whether an image required blurring, was uncertain, or was safe for display. Annotation was performed by eight annotators, grouped into four random pairs Each annotator reviewed 250 images per round. Following annotation, disagreements and Maybe cases were resolved through consensus discussions among the authors. Multiple rounds of evaluation and sampling were conducted to refine the dataset, resulting in a final balanced collection of 2,000 Sensitive and 2,000 Non-Sensitive images.

The 1,000 seed images from each of the Non-Sensitive and Sensitive classes were utilized to expand the collection using the KNN search. The subsequently manually annotated additional 1,000 images from each category are reserved as test set. To expand each category, the first 99 nearest neighbors were considered for each seed image. Through this KNN-based retrieval, a total of 183,706 Non-Sensitive and 78,025 Sensitive images were generated, as shown in Figure 2. We call this set `AutoMedGore`. Subsequently, we examined the relationship between the retrieval quality and the distance from the seed image within the KNN neighborhood to assess accuracy of the images.

2.2. Nearest Neighbor Homogeneity Estimation

Let's assume that if an image is classified as Sensitive or Non-Sensitive, there exists a probability π that its nearest neighbors (NNs) in the KNN space would also share the same classification. To empirically evaluate this relationship, pairs of annotators manually reviewed 10 images per NN rank from ranks 1 to 99, randomly sampled across clinical journals to ensure balanced representation. To enhance statistical robustness, NN ranks were grouped into bands of 20 ranks (i.e., 1–20, 21–40, etc.), resulting in 200 measurements per band. The resulting data were analyzed using a Beta-Binomial Bayesian Model, which estimates

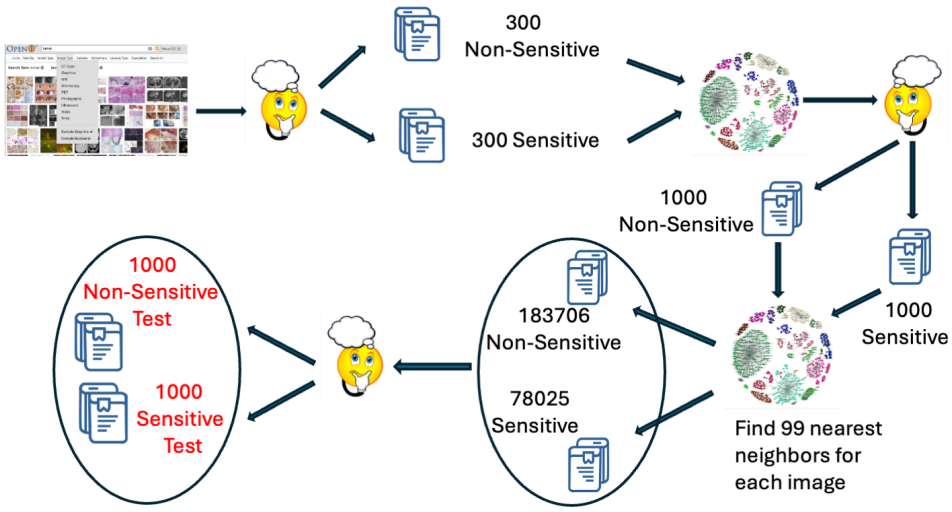


Figure 2: Illustration of MedGore dataset generation starting from 300 Open-i images to KNN based search and expansion a total of 183,706 Non-Sensitive and 78,025 Sensitive images

the probability π of a correct classification based on the number of correct and incorrect observations. A non-informative prior was used, with parameters $\alpha = a+1$, $\beta = b+1$ (where, a = number of NNs confirmed sensitive by human review and b = number NNs shown not-sensitive by human review) corresponding to a uniform prior distribution for π in the absence of prior measurements. The model has a probability density function, PDF(p , α, β), over possible values, p , for the actual value, π,

$$\text{PDF}(p, \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

where B is the Beta function.

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The PDF has mean $\mu = \frac{\alpha}{\alpha+\beta}$ and variance $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, which is all we need.

Our estimate for π is,

$$\pi = \frac{\alpha}{\alpha+\beta} \pm \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

We evaluated three cases: combined non-sensitive and sensitive images, non-sensitive images only, and sensitive images only. For each case, we tested several sets of equal-width nearest neighbor (NN) bands. Across all configurations—irrespective of NN rank or band width—the probability of correct classification remained approximately constant. Specifically, the combined case achieved $\sim 84\%$ accuracy, the sensitive-only case $\sim 76\%$, and the non-sensitive only case $\sim 92\%$. These findings indicate that the likelihood of an image being classified as sensitive or non-sensitive is roughly constant out to the distance of the 100th nearest neighbor of the seed images.

To further validate this observation, we adjusted the band widths and conducted additional reviews until the uncertainty was sufficiently reduced, ensuring that the probabilities associated with each band were distinguishable. Surprisingly, the classification probability remained stable across the first 99 nearest neighbors, suggesting that the boundary between blurred and sensitive images is not reached within this range. Expanding the index to include more nearest neighbors is necessary to investigate the transition region more comprehensively.

To compare and analyze the effectiveness of the automatically generated dataset, we consider two collections of the same size as the `MedGore` training set from `AutoMedGore` for training the models. In the first set `AutoMedGore-First3`, we sampled images of size 2000 from the closest three nearest neighbors; in another set called `AutoMedGore-Last3`, we sampled images of size 2000 from the farthest three nearest neighbors. We split the set as a training and a validation set, similar to the `MedGore`.

2.3. Inter-Annotation Agreement

Before reconciling the disagreements, we analyzed inter-annotation agreement using Gwet's AC1 coefficient (Gwet, 2008) and Cohen's Kappa (Wongpakaran et al., 2013). Gwet's AC1 was shown to provide a more stable inter-rater reliability coefficient and be less affected by prevalence and marginal probability than Cohen's Kappa (Wongpakaran et al., 2013). We obtained Gwet's AC1

pairwise values ranging from 0.92 to 0.94, Cohen's Kappa pairwise values ranging from 0.83 to 0.88 and F1 scores between 0.9 and 0.93, indicating substantial to near-perfect inter-annotator agreement and demonstrating that our annotation process is both reliable and consistent across different image categories.

2.4. Data Analysis

We analyzed the annotated images, which are labeled as Non-Sensitive or Sensitive based on the visual information they contained. In particular, we wanted to estimate the proportion of blood-like and skin-like regions in those images that are annotated to be blurred for public viewing. Following the literature on color detection (Chaves-González et al., 2007, 2010; Kakumanu et al., 2007), we aim to find the red color regions in the HSV color space and skin regions in the YCbCr color space. In the HSV color space, we seek the hue value corresponding to the red color, which shows the blood-like regions. We use the thresholding approach used in (Shaik et al., 2015) to detect the skin-like regions in the image. We hypothesized that if an image contains a high hue value and has been detected as a skin-like region, it is likely to be annotated as Blur. Towards this, we computed the proportions of blood-like and skin-like regions in a given image by aggregating the detected red hue scores and skin-like regions using thresholds and averaging over all values in the respective color spaces. For a given dataset, we normalized these scores using z-scores. Therefore, we assign a single estimated blur (EB) score for each image as the maximum of the blood-like and skin-like region scores. We have provided the histogram (Fig. 3) for the training and test set of `MedGore`, `AutoMedGore-First3`, and `AutoMedGore-Last3`. The histograms show that the average EB score (0) is mostly assigned to the Non-Sensitive images, and above average (>0) is assigned to the Blur images. This trend is consistent across the different datasets, as shown in Fig. 3a, 3b, 3c and 3d. It also signifies that the automatically created dataset `AutoMedGore-First3`, and `AutoMedGore-Last3` follow a similar data distribution to that of the manually assessed `MedGore`.

3. Benchmarking `MedGore`

3.1. Problem Statement

The task of sensitive medical image (SMI) classification intends to categorize a given medical image $\mathcal{I}$ (and associated text $\mathcal{T}$) into one of the classes[5] {'*sensitive*', '*non-sensitive*' }. Formally, we want to learn a parametrized function f which take $\mathcal{I}$ and $\mathcal{T}$ (if available) and predict the class $y = f(\mathcal{I}, \mathcal{T} : \theta)$.

3.2. Approaches

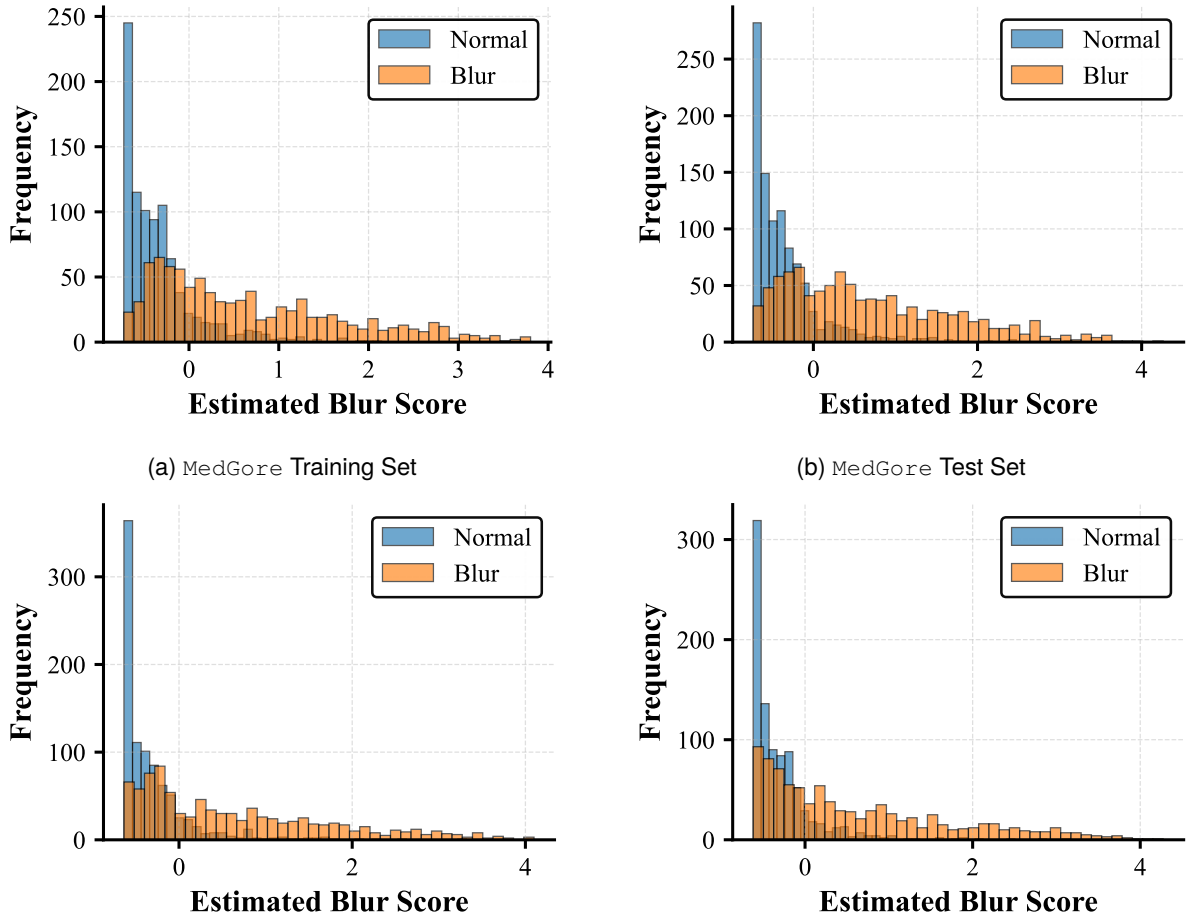
We benchmark the `MedGore` across three settings, each reflecting a different modality configuration.

Monomodal (Text): We performed the experiments with the pre-trained language models such as BERT (Devlin et al., 2019) variants ($\text{BERT}_{\text{Base}}$, $\text{BERT}_{\text{Large}}$) and RoBERTa (Liu et al., 2019) variants ($\text{RoBERTa}_{\text{Base}}$, $\text{RoBERTa}_{\text{Large}}$), which are pre-trained using self-supervised objective functions on a large-scale dataset. We also utilized the recent LLMs, such as LLama3 (Grattafiori et al., 2024) models (Llama-3-8B-Instruct, Llama-3.3-70B-Instruct) and Mistral (Jiang et al., 2023) (Mistral-7B-Instruct-v0.3), which are fine-tuned on an instruction following dataset. Additionally, we included the latest Qwen (Yang et al., 2025) family models (Qwen3-8B, Qwen3-14B), which integrate the thinking mode for complex tasks requiring multi-step reasoning with the non-thinking mode for rapid, context-driven responses into a single unified framework. We fine-tune the PLMs on the training set, and for LLMs, we experiment in the zero-shot setting with the prompt shown in Fig. 4.

Monomodal (Vision): We utilized a variety of vision models to train them for the task of SMI classification. Specifically, we used the ResNet-50 (He et al., 2016) model (microsoft/resnet-50), which is built upon a convolutional neural network (CNN) with residual connections, and the ViT (Dosovitskiy et al., 2020) model (google/vit-base-patch16-224), a pure transformer architecture that processes images as sequences of patches, equivalent to tokens in textual data. We also employed the Swin Transformer (Liu et al., 2021) model (microsoft/swin-base-patch4-window7-224), a hierarchical Transformer that computes self-attention within shifted windows. The shifted window design makes self-attention more efficient by focusing on local windows, while still allowing connections between them. Additionally, we included recent DINOv2 (Oquab et al., 2024) models (dinov2-small, dinov2-base, dinov2-large) that are designed to stabilize and accelerate discriminative self-supervised learning as models and datasets scale.

Multimodal (Text+Vision): The monomodal experiments are extended to multimodal, where we experiment with a variety of models that are equipped to process textual and visual modalities to learn the complexity in information. In particular, we utilized the CLIP (Radford et al., 2021) model (clip-vit-base-patch32), which is capable of learning a shared embedding space between image and text,

[5] For the benchmarking purpose and reporting all the results and analysis, we will refer to the *non-sensitive* images to be labeled as **Normal** and *sensitive* images as **Blur**.



(a) MedGore Training Set

(b) MedGore Test Set

(c) AutoMedGore-First3

(d) AutoMedGore-Last3

Figure 3: The estimate blur score vs. frequency of the images across the created different variants of the MedGore dataset.

Prompt : LLM Prompt

A medical image should be blurred if it depicts blood, tissue, organs, severe trauma, or decomposition in a manner that may be disturbing or inappropriate for general public viewing outside medical or clinical contexts. You are a classifier. Determine whether the image needs to be blurred based on the textual description of the image provided below. Respond strictly with one word: "BLUR" if the image should be blurred based on its textual description; otherwise, respond with "NORMAL".

{Textual Description}

Figure 4: Zero-shot prompt for monomodal (text) setup using large language models.

and the BLIP2 (Li et al., 2023) (blip2-flan-t5-xl) model, which leverages both frozen pre-trained image models and language models for the visual reasoning task. We also utilized the instruction-tuned large language-vision (LLVM) models such as Qwen2-VL (Wang et al., 2024) (Qwen2-VL-7B-Instruct) model and Llava (Liu et al., 2023) model (llava-1.5-7b-hf), which are fine-tuned on large-scale instruction-following datasets to offer competitive zero-shot performance. A recent multimodal large language model (MLLM), InternVL (Zhu et al., 2025) model (InternVL3-1B-hf), is used in our experiments, which is built upon multimodal Pre-Training and mixed preference optimization and has shown superior performance across a variety of languages and tasks. All the experiments of multimodal setting are done in a zero-shot setup with the prompt shown in Fig. 5, except CLIP, where we consider the text as the candidate class labels (Normal and Blur).

3.3. Experimental Setups

3.3.1. Datasets

We perform the experiments on human annotated image collections, MedGore and automatically

Prompt : LLVM and MLLM Prompt

A medical image should be blurred if it depicts blood, tissue, organs, severe trauma, or decomposition in a manner that may be disturbing or inappropriate for general public viewing outside medical or clinical contexts. You are a classifier. Determine whether the image needs to be blurred based on the given textual description and the image. Respond strictly with one word: "BLUR" if the image should be blurred based on its textual description; otherwise, respond with "NORMAL".
{Textual Description}
{Image}

Figure 5: Zero-shot prompt for multimodal setting using large language-vision models.

generated collection, `AutoMedGore-First3` and `AutoMedGore-Last3`. We split the set as a training and a validation set of size 1800 and 200, respectively. We train the models on the training set of `MedGore`. In all experiments, the validation dataset was used to select the best-performing model checkpoint, and the model was evaluated on the test set `MedGore`.

3.3.2. Modalities

Monomodal (Text): We consider two different sources of textual data from `MedGore` to perform the experiments. The first source is a **caption** of the image provided in the PubMed articles, and the second source is the **mention** sentence where the corresponding figure was referred to in the article. We trained two variants of each PLM on the training set and evaluated the model on the respective test set. Since we performed all the experiments for LLMs in a zero-shot setup, we evaluated the model on the test set, considering both **caption** and mention **textual** descriptions.
Monomodal (Vision): We fine-tune each vision model on the training set, considering only the image modality, and evaluate on the test dataset.
Multimodal (Text+Vision): Similar to the Monomodal (Text), we utilized two sets of textual descriptions to perform the experiments: caption and mention. All the experiments are done in a zero-shot setup on the test set.

3.4. Implementation Details

We set the maximum length of the textual description to 512 tokens for all the experiments. All the pre-trained models were used from the Hugging Face (Wolf et al., 2020) library. For the PLM's fine-tuning, we set the learning rate=$2e-5$, epoch=10, and batch size =16. We performed the zero-shot experiments on LLMs and VLLMs with the following configuration: maximum new tokens=4, and 4-bit quantization. For the vision models, we fine-tune them with the learning rate=$2e-5$, epoch=10, and batch size =16.

Evaluation Scheme We evaluated the performance of the models in terms of macro precision, recall, and F1-score. We also computed the area under the curve score for those models that produce the probability scores for the predicted class.

4. Results, Analysis and Discussion

4.1. Results

We have reported the results with Monomodal (Text), Monomodal (Vision), and Multimodal (Text+Vision) in Tables 1, 3. For the Monomodal (Text) setting, when training with the captions of the images, the $\text{RoBERTa}_{\text{Large}}$ achieved the maximum F1-score of 92 on the ground-truth test set of the `MedGore` dataset. With the same model, we reported the F1-score of 87.35 when training with the mentions of the images. The absolute performance drop of 4.65 when switching from captions to mentions shows that captions are the most accurate representation of the images. The other models (variants of BERT and $\text{RoBERTa}_{\text{Base}}$) also obtained the competitive F1-score. With the LLMs under the setting of Monomodal (Text), we obtained significantly lower performance except for the Mistral-7B-Instruct-v0.3, which reported an F1-score of 75.87 with Captions of the images and 65.65 with mentions of the images.

When considering the visual modality for the SMI classification task, we obtained better results compared to the textual modality. Under this setting, the model swin-base-patch4-window7-224 achieved the highest F1-score of 98.19, and other visual models also reported competitive performance, ranging from 94.01 (resnet-50) to 97.28 (dinov2-large). The results show that the fine-tuned visual models offer better discriminative information for the SMI classification task compared to the textual modality.

For the Multimodal (Text+Vision) setting, we conducted experiments using a zero-shot setup with five different models and reported the results in Table 1. For the clip-vit-base-patch32 model, we use the candidate labels (Blur and Normal) as text and compare the similarity with the image. The label with the highest similarity was assigned to the image. The clip-vit-base-patch32 model achieved an F1-score of 33.46 and the lowest precision score 25.15. We utilized the corresponding textual description (caption or mention) with the prompt shown in Fig. 5 for the remaining multimodal models. The model llava-1.5-7b-hf achieved

Model Name	Caption				Mention			
	Precision	Recall	F1-Score	AUC	Precision	Recall	F1-Score	AUC
Monomodal (Text)								
$BERT_{Base}$	91.96	91.77	91.74	96.11	87.66	85.98	85.77	94.02
$BERT_{Large}$	92.01	91.92	91.9	96.72	88.05	86.23	86.01	92.77
$RoBERTa_{Base}$	91.58	91.28	91.23	96.27	87.45	85.78	85.56	93.86
$RoBERTa_{Large}$	92	92.01	92	96.82	88.42	87.47	87.35	94.56
Mistral-7B-Instruct-v0.3	79.22	76.5	75.87	NA	67.17	66.12	65.65	NA
Llama-3.3-70B-Instruct	50.15	50	33.65	NA	73.8	53.19	40.4	NA
Llama-3-8B-Instruct	75.16	50.05	33.58	NA	75.16	50.05	33.58	NA
Qwen3-14B	37.34	47.22	34.54	NA	47.3	49.71	35.36	NA
Qwen3-8B	35.86	49.35	33.61	NA	46.51	49.71	34.88	NA
Multimodal (Text+Vision)								
clip-vit-base-patch32	25.15	50.00	33.46	NA	25.15	50.00	33.46	NA
blip2-flan-t5-xl	25.52	31.98	22.42	NA	17.65	30.66	22.41	NA
InternVL3-1B-hf	56.70	50.77	36.92	NA	58.48	51.43	38.83	NA
llava-1.5-7b-hf	40.24	40.15	40.14	NA	41.60	41.38	41.21	NA
Qwen2-VL-7B-Instruct	48.15	34.21	24.37	NA	50.35	34.21	24.24	NA

Table 1: Performance comparison of the monomodal (text) and multimodal (text+vision) approaches on the test set of the MedGore dataset. The 'NA' refers to models that generate an answer rather than predict with a confidence score.

Model Name	Caption				Mention			
	Precision	Recall	F1-Score	AUC	Precision	Recall	F1-Score	AUC
Monomodal (Text) with AutoMedGore-First3								
$BERT_{Base}$	84.83	82.87	82.56	93.8	86.34	83.29	82.85	93.35
$BERT_{Large}$	88.78	86.34	86.06	96.23	88.03	86.83	86.67	94.1
$RoBERTa_{Base}$	86.73	82.95	82.41	95.07	86.4	82.74	82.22	92.36
$RoBERTa_{Large}$	88.12	87.22	87.1	95.78	87.11	84.69	84.36	93.09
Monomodal (Text) AutoMedGore-Last3								
$BERT_{Base}$	86.34	83.29	82.85	93.35	86.51	83.84	83.46	93.2
$BERT_{Large}$	88.03	86.83	86.67	94.1	86.57	84.53	84.25	93.22
$RoBERTa_{Base}$	86.4	82.74	82.22	92.36	85.87	81.84	81.23	92.42
$RoBERTa_{Large}$	87.11	84.69	84.36	93.09	87.03	85.28	85.04	93.73

Table 2: Performance comparison of the monomodal (text) approaches by training the model on AutoMedGore-First3 and AutoMedGore-Last3 on the test set of the MedGore dataset.

Models	Precision	Recall	F1-Score	AUC
MedGore				
vit-base-patch16-224	96.87	96.75	96.73	96.75
resnet-50	94.21	94.03	94.01	94.03
swin-base-patch4-window7-224	98.2	98.19	98.19	98.19
dinov2-base	96.44	96.43	96.43	96.43
dinov2-large	97.34	97.29	97.28	97.29
dinov2-small	96.15	96	95.97	96
AutoMedGore-First3				
vit-base-patch16-224	94.52	94.05	94	94.05
resnet-50	89.8	88.09	87.91	88.09
swin-base-patch4-window7-224	94.96	94.55	94.51	94.55
dinov2-base	91.21	89.85	89.71	89.85
dinov2-large	92.46	91.35	91.24	91.35
dinov2-small	91.02	90.58	90.52	90.58
AutoMedGore-Last3				
vit-base-patch16-224	94.32	93.89	93.85	93.89
resnet-50	92.12	91.05	90.94	91.05
swin-base-patch4-window7-224	93.81	93.2	93.14	93.2
dinov2-base	91.96	90.94	90.84	90.94
dinov2-large	92.01	90.65	90.52	90.65
dinov2-small	92.48	91.74	91.67	91.74

Table 3: Performance comparison of the monomodal (vision) approaches by training the model on MedGore, AutoMedGore-First3 and AutoMedGore-Last3 on the test set of the MedGore dataset.

the highest F1-score of 40.14 amongst all multimodal models with the balanced precision and recall scores. The model InternVL3-1B-hf obtained the highest precision score of 56.70 and recall score of 50.77. The model predicted most of the 1929 images into the Blur category out of a total of 1988 images in the test set, which yielded a high F1-score (66.9) of class Blur but a very low F1-score (7.1) for Normal images, leading to the lower macro-F1 score of 36.92.

4.2. Analysis

To analyze the effectiveness of the automatically created dataset, we conducted additional experiments, selecting samples using two different strategies (AutoMedGore-First3 and AutoMedGore-Last3) to create a training set for comparison with the human-annotated training set MedGore. The experimental results depicted in Tables 1, 3 show that when the model was trained on the training set of MedGore, it performed bet-

ter in monomodal (text) and monomodal (vision) settings. Following this, we aim to analyze the effect of model performances when they are trained on AutoMedGore-First3 and AutoMedGore-Last3 datasets. For the monomodal (text) setting, we reported the results in Table 2 by training the models with AutoMedGore-First3 and AutoMedGore-Last3 datasets and evaluating on the same ground-truth dataset, which is part of MedGore collection. We obtained the best results with the RoBERTa$_{\text{Large}}$ model, achieving an F1-score of 87.1 and 84.36 with AutoMedGore-First3 considering the caption and mention, respectively, as the textual description of the images. Comparing these results with the MedGore training set, we can say that though there is a performance decrement, it has also to be noted that the AutoMedGore-First3 is our automatically created dataset, which is achieving stronger results compared to the resource and time-consuming construction of the human-annotated MedGore dataset. We observe similar trends when the AutoMedGore-Last3 was considered for training the models. We extend the same experiments with a monomodal (vision) setting and reported and compared the results in Table 3. We observe that the best-performing swin-base-patch4-window7-224 model achieved an F1-score of 94.51 and 93.14 for AutoMedGore-First3 and AutoMedGore-Last3, respectively. These results allow us to make two claims: **(1)** our automatically created dataset is producing comparable results for the task of SMI classification, **(2)** the quality of the automatically generated datasets remains similar as we descend into nearest neighbor ranking while choosing the samples.

4.3. Discussion

The analysis of NN homogeneity estimation using the Beta-Binomial Bayesian Model demonstrates that clusters formed around the seed images are consistent and homogeneous. The probability of correct classification remains fairly constant at approximately 84% for distances out to the 100th NN, indicating that the retrieved images are visually very similar to their corresponding seeds. A higher level of homogeneity is observed for non-sensitive images ($\sim 92\%$) compared to sensitive images ($\sim 76\%$), likely due to the greater number and visual uniformity of the non-sensitive seeds locally. Overall, these results suggest that clusters within the first 99 NNs are well separated and maintain strong class consistency.

The KNN expansion resulted in 183,000 non-sensitive images and 78,000 sensitive images. This imbalance arises because 77.45% of non-sensitive images are unique, whereas only 31.85% of sensitive images are unique. This indicates that sensitive image clusters exhibit greater overlap, while non-sensitive image clusters are more widely distributed across the feature space. The higher redundancy among sensitive images can be attributed to their visual similarity in image features, such as the predominance of red tones in surgical photographs. Additionally, a small proportion of non-photographic images, such as MRI or X-ray scans, are included in the non-sensitive category. Also some photographs were also filtered during processing, further contributing to this distributional difference.

The experiments on a variety of monomodal and multimodal approaches reveal that the task of SMI classification is well handled by a hierarchical vision transformer (Dosovitskiy et al., 2020), which processes images in patches. However, ResNet50 achieved the highest performance, even though it was not one of the computationally inexpensive models evaluated in our study. We also find that the monomodal (text) model with textual descriptions (captions or mentions) as input for images is not better than the monomodal (vision) model. It can also be attributed to our visual analysis (see Section 2.4), which shows that images with blood-like and skin-like regions tend to produce higher estimated blur scores. We observed around 7% performance decrement between the best-performing monomodal (text) and monomodal (vision) models. In our experiments, we found that neither LLMs nor VLLMs under a zero-shot setting could perform better on the SMI classification task. Given the better performance of the lightweight vision model, we did not extend our study to fine-tune those LLMs and VLLMs, which require substantial computational resources, as this would not be a practical solution for model deployment. We also observe that the high performance of the SMI classification could be due to a binary classification setup, which makes the model learn discriminative features between normal and gory images. A more challenging problem could be fine-grained gory image classification, where the model needs to identify a specific type of gory image, if any is present. We will consider this extension in future work.

5. Conclusion

In this work, we introduce a large dataset of medical photographs, their captions and corresponding discussions in the full text of the medical articles from which the images were extracted. The dataset was generated to enable development of approaches to automated detection of sensitive medical images. We introduce a KNN-based approach to rapid labeling of the images and an approach to NN homogeneity estimation using the Beta-Binomial Bayesian Model. Finally, we establish mono- and multi-modal benchmarks for image

labeling using image and text features.

6. Ethical Considerations and Limitations

The images in the MedGore collection are sourced from the Open Access subset of PubMedCentral. The dataset may be freely downloaded, reused, redistributed, and publicly shared under the applicable open-access license. To the best of our knowledge, the journals that publish the images require patients' concent and IRB review. These images are publicly available at https://openi.nlm.nih.gov. While the dataset we presented is large, covers multiple aspects of sensitive image content, and enables highly accurate image classification, our analysis shows that some unknown sensitive image types may be present in Open-i and other medical image collections. We believe that the presented approach to the identification of such images will enable filling these gaps in future work.

7. Data Availability

MedGore Dataset is available for download at https://bionlp.nlm.nih.gov/MedGore_Dataset.zip.

8. Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

9. Bibliographical References

Fatih Cagatay Akyon and Alptekin Temizel. 2023. State-of-the-art in nudity classification: A comparative analysis. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. 2020. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283.

Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. 2023. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):1–8.

Jose M Chaves-González, Miguel A Vega-Rodríguez, Juan A Gómez-Pulido, and Juan M Sánchez-Pérez. 2007. Colour spaces study for skin colour detection in face recognition systems. In *International Conference on Security and Cryptography*, volume 2, pages 171–174. SCITEPRESS.

Jose M Chaves-González, Miguel A Vega-Rodríguez, Juan A Gómez-Pulido, and Juan M Sánchez-Pérez. 2010. Detecting skin in face recognition systems: A colour spaces study. *Digital signal processing*, 20(3):806–823.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Deepak Gupta, Russell Loane, Soumya Gayen, and Dina Demner-Fushman. 2023. Medical image retrieval via nearest neighbor search on pre-trained image features. *Knowledge-based systems*, 278:110907.

Kilem L. Gwet. 2008. *Computing inter-rater reliability and its variance in the presence of high agreement*. Advanced Analytics LLC.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. 2007. A survey of skin-color modeling and detection methods. *Pattern recognition*, 40(3):1106–1122.

Sasan Karamizadeh, Saman Shojae Chaeikar, and Alireza Jolfaei. 2023. Adult content image recognition by boltzmann machine limited and deep learning. *Evolutionary Intelligence*, 16(4):1185–1194.

William Larocque. 2021. *Gore classification and censoring in images*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Featured Certification.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Manuel Sebastián Ríos, María Alejandra Molina-Rodriguez, Daniella Londoño, Camilo Andrés Guillén, Sebastián Sierra, Felipe Zapata, and Luis Felipe Giraldo. 2023. Cholec80-cvs: An open dataset with an evaluation of strasberg's critical view of safety for ai. *Scientific Data*, 10(1):194.

Khamar Basha Shaik, P Ganesan, V Kalist, BS Sathish, and J Merlin Mary Jenitha. 2015. Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, 57:41–48.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.