Mislabeling Misinformation: Annotation Consistency Shapes Machine Learning for DIY Health Risks

Manon Pilaud[1][0009-0002-7586-6944], Alexandra J. Berges [2][0000-0002-0890-6291],
Ian McCulloh[3][0000-0003-2916-3914]

[1] Johns Hopkins University, Baltimore, MD 21218, USA
mpilaud1@jhu.edu
[2] Johns Hopkins University, Baltimore, MD 21218, USA
aberges1@jhmi.edu
[3] Johns Hopkins University, Baltimore, MD 21218, USA
imccull4@jhu.edu

Abstract. The rise of do-it-yourself (DIY) cosmetic procedures promoted on social media platforms such as TikTok has introduced new risks to public health, particularly as untrained individuals attempt at-home dermal filler injections. This study investigates the feasibility of detecting medical misinformation related to DIY fillers using machine learning techniques and examines the impact of annotation consistency on model performance. We collected and manually labeled 195 TikTok posts using a three-class schema: non-relevant, relevant-benign, and relevant-misinformation. Labels were assigned by both a medical expert and a technical contributor, with a subset re-labeled to assess intra-annotator agreement. Results showed moderate-to-substantial agreement within the same expert (κ = 0.624) but low agreement across annotators, revealing variability in label interpretation. A Random Forest classifier trained on different label subsets showed that annotations from the more internally consistent rater led to stronger model performance, particularly in precision. These findings underscore the importance of early investment in annotation quality and inter-rater validation when building AI systems for misinformation detection. We discuss the implications for public health surveillance and propose future work to scale content filtering and support qualitative review by experts.

Keywords: Medical misinformation · TikTok · Cosmetic procedures · Annotation consistency · Health informatics · Machine learning.

1 Introduction

Social media platforms have become prominent sources of health-related information, especially among younger audiences seeking advice on beauty, wellness, and cosmetic enhancements. While these platforms provide opportunities for public education and peer-to-peer support, they also serve as vectors for the rapid spread of unverified, misleading, or dangerous medical content [1-4]. In recent years, there has been a disturbing rise in user-generated content promoting do-it-yourself (DIY) cosmetic procedures, including at-home dermal filler injections [3-4]. These posts often trivialize complex medical interventions and lack warnings about the potential risks, such as infection, vascular occlusion, tissue necrosis, and permanent disfigurement [2-4]. As emergency room physicians report a surge in cases involving complications from such procedures, the need to understand and monitor this digital ecosystem has become urgent.

This study focuses on detecting and characterizing medical misinformation related to DIY filler injections on visual and video-centric platforms, specifically Instagram and TikTok. We hypothesize that social media content glamorizing or providing unqualified instruction on cosmetic injections contributes to real-world harm by encouraging medically unsupervised behavior. However, automated detection of this misinformation presents several challenges: the content is highly multimodal (involving images, videos, captions, and hashtags), often context-dependent, and interspersed with a significant volume of irrelevant or benign material [1]. Furthermore, misinformation comprises a small proportion of overall content, making class imbalance a major barrier to effective classification [5]. Our research aims to address these challenges by training a classifier to distinguish between irrelevant content, relevant benign information, and high-risk misinformation.

To support this work, we developed a custom data collection script to retrieve thousands of posts using keywords associated with DIY filler procedures. After filtering for language and engagement thresholds, we manually annotated a sample of 195 posts. Each post was labeled as non-relevant, relevant-benign, or relevant-misinformation by a board-certified emergency room physician with direct experience treating patients harmed by DIY cosmetic procedures. To assess labeling consistency, the physician re-labeled a subset of 43 posts at a later time. A computer scientist who developed the scraping script also independently labeled the full dataset, allowing us to measure inter-annotator agreement and examine how differences in domain expertise influence labeling outcomes. These labels served as the foundation for training and evaluating our classification models.

Our contributions are threefold. First, we provide an annotated dataset of social media posts related to DIY cosmetic injections with labels informed by clinical expertise. Second, we assess the reliability of human annotations and its implications for machine learning performance in low-prevalence misinformation settings. Third, we validate a machine learning classifier trained on expert-labeled social media posts to explore how labeling consistency affects performance. We pose the following research questions:

RQ1: What level of agreement exists between domain experts and non-experts when labeling medically relevant social media content?

RQ2: How does label consistency impact model performance?

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on medical misinformation and content moderation on social platforms. Section 3 outlines our data collection, labeling, and annotation validation processes. Section 4 presents our results, including inter-annotator agreement scores and classifier performance across label subsets. Finally, Section 5 discusses the implications of our findings for real-time misinformation monitoring, labeling protocol design, and future improvements in training data acquisition for rare-event classification problems in health informatics.

2 Background

The detection of medical misinformation on social media increasingly relies on multimodal machine learning approaches that combine textual, visual, and contextual features. Prior work has demonstrated that while these systems can identify misinformation in domains such as COVID-19 or anti-vaccine content, they often falter in niche applications—particularly DIY cosmetic advice—where misinformation is rare, richly contextual, and heavily image or video-based [6]. Multimodal deep learning methods have shown promise in detecting dangerous medical claims on social media platforms, yet they struggle with false positives and lack robust clinical validation [1,6]. Studies specifically targeting cosmetic misinformation remain sparse, with most relying on generic public-labeled datasets or keyword-based filtering that fails to capture nuanced medical advice. Consequently, these models often underperform in identifying visual evidence of harmful practices, such as self-administered filler injections, and are limited by a lack of expert-annotated training data [6,7].

Label noise and disagreement are inherent challenges in medical misinformation detection and significantly influence model generalizability and downstream utility. McCulloh et al. stress that effective leadership and quality control in data annotation—through detailed guidelines, annotator training, oversight, and adjudication protocols—lead to higher consistency and improved dataset quality [8]. In practical terms, their framework provides methods for annotator coordination, structured feedback loops, and error monitoring in large labeling efforts. Importantly, Nassar et al. demonstrated that removing an inconsistent annotator from an object recognition dataset increased model precision but caused a drop in recall due to the smaller dataset [9]. This finding underscores that while data quantity may be replenished over time, unchecked label inconsistency can embed long-term technical debt in AI systems if not corrected early.

Inter-annotator agreement further illustrates this dynamic relationship, acting both as a diagnostic of label quality and a predictor of model efficacy. Several studies show that Cohen's Kappa or Krippendorff's Alpha scores below moderate levels correspond to lower algorithmic precision and reduced trust in model predictions. For example, in the Nassar et al. study, low Krippendorff's Alpha among annotators directly led to reduced classifier precision, prompting recommendations for continuous annotation monitoring [9]. Other research in medical image segmentation advocates for hierarchical correction methods such as expert adjudication, consensus voting, or label refinement algorithms to effectively resolve disagreements and improve classification outcomes [10].

When evaluating classifier performance, particularly in medical misinformation settings, AUC, precision, recall, and F1-score provide richer insights than simple accuracy—especially in imbalanced or expert-labeling contexts. AUC (area under the ROC curve) measures a model's discrimination ability across all threshold levels; precision quantifies the reliability of positive predictions (i.e. how many predicted misinformation items are truly misinformation); recall indicates the proportion of actual misinformation correctly identified; and F1-score balances precision and recall. In low-prevalence settings, a model may achieve high overall accuracy by defaulting to the majority class, yet still fail to detect harmful misinformation. Moreover, these metrics depend on a defined "ground truth"; when labels derive from subjective human judgment, the concept of ground truth becomes fuzzy, magnifying the uncertainty in metric interpretation.

Depending on the application, researchers may choose to favor recall—ensuring that as much relevant content as possible is captured—or prioritize precision, so that the flagged content is highly relevant and actionable. For example, in public health surveillance, maximizing recall might be critical to ensure no harmful misinformation goes undetected, even at the cost of false positives. Conversely, in resource-constrained settings, a higher precision may be preferable to reduce the burden of manually reviewing false alarms. However, it is important to recognize that these thresholding decisions are made after the classifier has been trained. The underlying shape of the precision-recall trade-off curve—and the model's overall AUC—is determined during training and is heavily influenced by the quality and consistency of the labeled data. If annotation inconsistencies are present early in the process, they can constrain the model's learning potential and distort its decision boundary, ultimately limiting the range of viable thresholds for operational use. Therefore, investing in consistent labeling practices upfront is essential to achieving a model with a favorable precision-recall profile that offers flexibility for downstream decision-making.

Empirical guidelines for dataset sizing, such as those proposed by Koshute et al., help estimate the necessary number of labeled examples to develop reliable classifiers [11].

Their work suggests that modern classifiers—especially in complex, low-prevalence domains—often require thousands to tens of thousands of annotated instances per class to reach performance plateaus. In settings like DIY cosmetic misinformation, where harmful content arises in fewer than 5% of posts, constructing balanced datasets is particularly challenging; without careful annotation design, models risk being under-trained on the target concept or overfitting to noise.

In summary, effective systems for detecting medical misinformation—such as those addressing DIY cosmetic injections—require an integrated approach: leveraging expert-guided annotations, implementing rigorous annotation leadership and disagreement management, and evaluating models with appropriate, context-sensitive performance metrics. These principles guided our development of a relevance classification model capable of early filtering of irrelevant content and identification of harmful misinformation.

3 Methodology

To identify social media content related to do-it-yourself (DIY) cosmetic injections, we developed a custom web scraping pipeline targeting TikTok. The scraper was seeded with a manually curated list of hashtags and keyword-based search terms associated with cosmetic procedures. Queries included high-risk and tutorial-style tags such as #athomebotox, #athomefiller, #fillerdiy, #botoxdiy, and #hyaluronpen, as well as broader procedural terms like #botox, #filler, and their plain-text equivalents ("Botox", "Filler"). Each query was executed independently with the goal of retrieving up to 100 posts per term, although some terms—such as #filler—returned fewer posts due to platform-specific constraints. Posts were deduplicated based on video URLs, resulting in a final dataset of 1,180 unique TikTok posts. Metadata collected for each post included caption text, hashtags, username, timestamp, and user engagement metrics (likes and shares).

To ensure temporal and social relevance, we filtered the dataset to include only posts published in the six months prior to collection and ranked them by user engagement. From this filtered subset, the top 200 posts were selected for manual annotation. Posts were labeled using a three-tier relevance scale: 0 = Not Relevant (e.g., unrelated content), 1 = Relevant–Benign (e.g., medically related but not misleading), and 2 = Relevant–Misinformation (e.g., promotional or instructional content encouraging at-home filler use without medical supervision).

Three rounds of annotations were conducted. The first set was labeled by a data engineer with technical knowledge of the scraping pipeline (referred to as DE). The second set was labeled by an emergency room physician (MD1) with direct clinical experience treating complications from cosmetic procedures. The physician re-labeled a random sample of 43 posts at a later time (MD2) without referencing prior labels, allowing us to evaluate intra-annotator consistency. All annotations were stored using structured identifiers and merged using the unique video URL of each post.

To quantify agreement across annotators, we computed pairwise Cohen’s Kappa scores between DE, MD1, and MD2 for the subset of posts with overlapping annotations. Before analysis, we excluded any records with missing labels or posts marked as inaccessible or invalid. Annotated datasets were aligned on the video_url field, and confusion matrices were generated to visualize labeling discrepancies across each annotator pair. These visualizations were saved for documentation and later inspection.

To evaluate the potential for automated classification of social media content, we trained a Random Forest model to predict the relevance of posts using textual features derived from the caption text, hashtags, and username fields. Each feature type was independently vectorized using a term frequency–inverse document frequency (TF-IDF) transformation, and a Column Transformer was used to apply these transformations in parallel across input fields. This approach enabled the integration of heterogeneous text inputs into a unified model pipeline.

Model training and evaluation were conducted using 5-fold stratified cross-validation to maintain class balance across folds. The classifier's performance was assessed using macro-averaged metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). These macro-averaged metrics were selected to account for class imbalance and provide a more comprehensive measure of performance across all label categories.

In cases where individual test folds lacked representation from one or more classes—an expected occurrence given the low prevalence of certain labels—predicted probability vectors were padded to preserve a consistent output structure and ensure valid computation of macro-AUC scores. ROC curves were also generated for each subset of annotations (e.g., DE-only, MD1-only, MD2-only, and combinations) to enable visual comparison of classifier performance across different labeling sources.

4 Findings

Our analysis evaluated the consistency of human annotators and its impact on the performance of machine learning models trained to detect medical misinformation in DIY cosmetic content. A total of 195 social media posts were annotated by two individuals: an emergency room physician (MD1) and a data engineer (DE). The physician re-annotated 43 of these posts at a later time (MD2), allowing us to assess intra-annotator reliability. Inter-annotator agreement was measured using Cohen's Kappa. The results indicated moderate-to-substantial intra-annotator agreement between MD1 and MD2 (κ = 0.624), moderate inter-annotator agreement between MD1 and DE (κ = 0.453), and only slight agreement between MD2 and DE (κ = 0.058), suggesting notable variability in how different individuals interpret the labeling rubric and how the same individuals may differ in their repeat assessment of the same data.

Classifier performance was assessed using five-fold cross-validation across different training label sources and combinations. Models trained on DE-only labels achieved the highest AUC (0.8044) and F1-score (0.6097), outperforming all other configurations. In contrast, models trained solely on MD1 labels underperformed, achieving an AUC of 0.7484 and an F1-score of 0.4369, despite the full 195-sample training set. Interestingly, the model trained on the 43 double-annotated samples from MD1 and MD2 achieved an AUC of 0.8075—slightly higher than the DE-only model—along with strong F1 and precision scores, indicating that label consistency can partially compensate for smaller training sizes (see Figure 1 and Table 1).

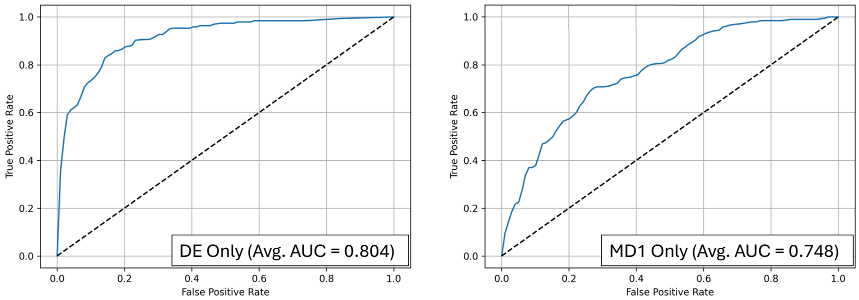

Figure 1. AUC curves for DE only and MD1 only, demonstrating that consistency is perhaps more important than medical expertise in labeling content.

Table 1. Average 5-fold random forest performance across training data sets.

	DE	MD 1	MD 2	MD 1+2	DE + MD1	DE + MD2
AUC	0.8044	0.7484	-	0.8075	0.7923	0.7425
F1-score	0.6097	0.4369	0.4584	0.6029	0.5229	0.5751
Precision	0.6557	0.5801	0.4458	0.6461	0.5198	0.6187
Recall	0.5957	0.4489	0.5333	0.5984	0.5297	0.5797
Accuracy	0.7761	0.5795	0.6972	0.6310	0.6743	0.6851
Training *n*	195	195	43	43	195	43

The model trained on MD2 labels alone (n = 43) performed better than MD1 alone across most metrics (F1 = 0.4584 vs. 0.4369; Accuracy = 0.6972 vs. 0.5795), even with a substantially smaller training set. This supports the hypothesis that MD2 labels—

presumably more internally consistent—led to better generalization. Moreover, combining DE with MD2 annotations resulted in improved precision (0.6187) and F1-score (0.5751) over DE + MD1 (F1 = 0.5229), again reinforcing the benefit of aligning with more consistent labels.

These findings suggest that annotator consistency has a direct influence on model performance—particularly on precision. While quantity of data matters for recall, as illustrated by the superior recall in DE-only and DE + MD1 models (both with n = 195), inconsistent labels can constrain the model's ability to learn accurate decision boundaries, reducing the precision and overall discriminative power. This aligns with previous research [9] showing that removal of inconsistent annotators can increase precision but often lowers recall due to decreased data volume. Our study replicates this trade-off, highlighting that while data quantity can be improved incrementally over time, label inconsistency—if left unaddressed—introduces persistent noise that can degrade long-term model quality.

In summary, our results demonstrate that careful attention to annotation quality, particularly intra-annotator consistency, is critical in the early stages of machine learning pipeline development for misinformation detection. Classifier thresholds and application-specific tuning decisions—such as favoring recall or precision—are made after training; thus, maximizing the potential of the model during training requires high-quality, reliable labeled data. As our findings show, consistent expert annotations—even at smaller scales—can yield superior model performance compared to larger, noisier datasets. These insights will inform future improvements to our labeling protocol and underscore the importance of consensus-based annotation strategies in domains with low-prevalence and high-context complexity.

5 Conclusion and Future Work

This study demonstrates the feasibility and importance of applying machine learning techniques to detect niche forms of medical misinformation, such as do-it-yourself (DIY) cosmetic filler injections, within noisy and highly contextual social media environments. Our findings show that, even in a low-prevalence and imbalanced setting, a classifier trained on well-labeled social media data can differentiate between non-relevant, benign, and misinformation content with meaningful performance. These results are particularly relevant for public health surveillance, where automated systems may serve as early-warning tools to flag high-risk content for expert review.

A central insight from this work is the critical role of label consistency in building high-performing models. We observed that inconsistent annotations, including those made by trained medical professionals, can significantly degrade model precision and reliability. Conversely, consistent labeling—regardless of annotator background—

yielded more generalizable and effective models. This challenges the assumption that subject-matter expertise alone is sufficient for high-quality annotation. Instead, it underscores the need for multiple annotators, clear labeling rubrics, and systematic inter-rater agreement assessment to ensure reliable training data. These processes help mitigate human error early in the pipeline, leading to stronger and more reproducible machine learning systems.

Despite the promising results, our study has limitations. The dataset was relatively small, and the low prevalence of misinformation restricted our ability to train class-balanced models. Additionally, the use of a three-class rubric may oversimplify the complex and often subtle nature of misinformation. Future work should expand the annotation schema to capture more nuanced misinformation signals, incorporate a more diverse pool of annotators to assess robustness across perspectives, and scale up data collection efforts. In particular, the trained classifier can be used to pre-filter large datasets, allowing experts to focus their attention on more likely candidates for misinformation—supporting qualitative analysis of emerging themes and informing timely, evidence-based public health interventions.

In summary, this work offers several key contributions: (1) it presents a novel application of machine learning to the detection of DIY cosmetic injection misinformation; (2) it provides empirical evidence that label consistency has a greater impact on model quality than annotator domain expertise alone; and (3) it offers practical guidance for integrating annotation quality control into the development of AI systems for health misinformation monitoring. These insights will inform future efforts to build scalable, trustworthy, and context-aware misinformation detection tools.

References

1. Wang Z., Yin Z., Argyris YA.; *Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning*, arXiv (2020). Supports your point about multimodal misinformation detection in health contexts
2. "Dermal Filler Videos With Low Quality Health Information Garner High Viewership on Social Media," *Dermatology Times* (2023). https://www.dermatologytimes.com/view/dermal-filler-videos-with-low-quality-health-information-garner-high-viewership-on-social-media. accessed 11 June 2025. Demonstrates that filler-related content on platforms like TikTok and Instagram is often misleading
3. "TikTok, phony doctors to blame for dangerous DIY lip filler trend …", ABC News (2021). https://abcnews.go.com/GMA/Style/tiktok-phony-doctors-blame-dangerous-diy-lip-filler/story?id=80722294. Accessed 11 June 2025. Highlights the proliferation of hyaluron pen usage on TikTok and associated health risks
4. "People Are Doing Their Own Filler at Home Now," Allure (2025). https://www.allure.com/story/at-home-lip-filler-and-botox. Accessed 11 June 2025. Confirms increased self-injection behaviors with real-world complications and unregulated substances
5. Faisal DR, Mahendra R.; *Two-Stage Classifier for COVID-19 Misinformation Detection Using BERT: a Study on Indonesian Tweets*, arXiv (2022). Confirms benefits of staged filtering approaches for noisy social data

6. **Zuhui Wang, Zhaozheng Yin, Young Anna Argyris**, *Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning*, arXiv (2020). https://arxiv.org/abs/2012.13968?utm_source=chatgpt.com
7. Emily P. Smith, ..., Social media dermatologic advice prone to misinformation, PMC (2022). https://pmc.ncbi.nlm.nih.gov/articles/PMC10315776/
8. McCulloh, Ian, et al. "Leadership of data annotation teams." 2018 International Workshop on Social Sensing (SocialSens). IEEE, 2018.
9. Nassar, Joseph, et al. "Assessing data quality of annotations with Krippendorff alpha for applications in computer vision." *arXiv preprint arXiv:1912.10107* (2019).
10. Warfield, Simon K., Kelly H. Zou, and William M. Wells. *Assessing Inter-Observer Agreement for Medical Image Segmentation*. National Library of Medicine, 2004. https://lhncbc.nlm.nih.gov/LHC-publications/PDF/Assessing_Inter-Annotator_Agreement_for_Medical_Image_Segmentation.pdf.
11. Koshute, Phillip, Jared Zook, and Ian McCulloh. "Recommending training set sizes for classification." *arXiv preprint arXiv:2102.09382* (2021).