INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

iCensr: A Web Image Detection and Censorship Plugin Utilizing the YOLO Deep Learning Method

Justin Gil B. Cruzada
justin_cruzada@yahoo.com
StartupLab Business Center, Philippines

Reynaldo A. Lomboy Jr.
reynaldo@archerscontactsolutions.com
Archers Contact Solutions, Philippines

Jesse Kabel N. Ruiz
jesseruiz197@gmail.com
P&A Grant Thornton, Philippines

Jerian R. Peren
jerian.peren@lpu.edu.ph
Lyceum of the Philippines University Cavite, Philippines

Abstract: The research introduced and developed a content moderation tool designed for Chromium-based browsers such as Google Chrome. It delved into assessing the effectiveness of YOLO v8 within iCensr, a browser plugin aimed at improving online browsing by ensuring a secure web environment. The primary objective of the plugin is to detect and censor objectionable images, including those depicting nudity, violence, and illicit drugs, across diverse websites, thus regulating content exposure online. The study evaluated the models of iCensr using Mean Average Precision (mAP). A total of 58 participants assessed the iCensr plugin through a Likert-scale survey based on ISO/IEC 25010:2023 acceptability standards. The outcomes of the evaluation suggest that iCensr is deemed "Highly Acceptable," indicating its potential to contribute to safer online interactions. The research underscores the significance of digital tools like iCensr in mitigating online risks and fostering a secure online environment for users of all ages. Additionally, the researchers recommend that future developers and researchers expand the censorship categories, implement other techniques, create a mobile version, and acquire better datasets for enhancing its functionality and effectiveness.

Keywords: objectionable images, browser plugin, content moderation, YOLOv8, iCensr Plugin, Online Security

INTRODUCTION

In the modern digital era, social computing serves as a bridge connecting diverse cultures through online platforms such as blogs, wikis, social networking sites, and Internet discussion forums (Kim, 2022). As of 2022, the average daily global social media usage was 151 minutes. Notably, young individuals are highly engaged with social media, with the Philippines standing out for having the longest average time spent on these platforms, as online users dedicate three hours and fifty-three minutes daily (Dixon, 2023). This context underscores the growing prevalence of visual content online and the increasing need for advanced tools to ensure the safety and integrity of online platforms.

The internet contains graphic and inappropriate content, including not safe for work (NSFW) material, explicit imagery, propaganda, crime, violence, and profanity. Despite the availability of safety features such as secure search, parental controls, and content filters, users, particularly children, frequently encounter harmful content. These protective measures may be ineffective when children demonstrate greater internet proficiency than their guardians. Alarmingly, children as young as 8 or 9 years old often stumble upon inappropriate content online (Severen, 2022).

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

Object detection is a vital aspect of computer vision. Its primary purpose is to identify visual elements belonging to predefined categories, such as humans, animals, vehicles, or architectural structures, within digital imagery like photographs or video frames (Boesch, 2023). This technology is instrumental in enabling researchers to identify particular content categories unsuitable for minors or individuals with specific preferences, thereby ensuring a more tailored and responsible experience in media consumption.

The moderation of NSFW or explicit web content has been extensively researched, with solutions leveraging image classification technology and various machine learning algorithms. Studies by Izzah et al. (2018), Bhatti et al. (2018), Bicho et al. (2020), and Kim (2022) primarily focus on explicit content but fail to encompass categories such as violence, propaganda, or crime, which also significantly impact web users (Stubbs et al., 2022; Internet Matters, 2018).

To enhance the effectiveness of the developed plugin, this study suggests using the state-of-the-art (SOTA) deep learning model YOLO version 8. This research aims to address the critical concern of identifying and mitigating unwanted or potentially harmful images in real-time, contributing to ensuring the security and appropriateness of online spaces (Bicho et al., 2020). In this era of information abundance, swiftly and accurately detecting and censoring objectionable images plays a pivotal role in safeguarding the online experience of young users, general audiences, and content providers.

This study explores YOLO-v8 (You Only Look Once version 8), a state-of-the-art deep learning algorithm known for its remarkable image classification capabilities and near-human-level proficiency (Hussain, 2023). The aim is to create an effective solution for enhancing digital safety and user experience.

REVIEW OF RELATED LITERATURE

The widespread availability of digital content on online platforms has led to a significant problem concerning the increased presence of objectionable images. These objectionable images are prevalent across the internet and can have various negative effects on users, particularly children and adolescents. These adverse effects include psychological and behavioral impacts, desensitization, normalization, and the potential for harm (Zulfiqar, 2021; Vinney, 2023; Facing History & Ourselves, 2016; Lin et al., 2020; Khan et al., 2020). With the easy accessibility of visual media, maintaining safe and secure digital spaces has become increasingly challenging (Internet Matters, 2018).

To contribute to addressing this issue, this study created a plugin designed to identify and censor objectionable images. The proposed plugin leverages the capabilities of deep learning algorithms, which process data by drawing inspiration from the human brain to help identify patterns in various forms of data, including images (Gillis et al., 2023). To identify and filter out images containing objectionable content, the process involves using object detection techniques to accurately recognize and classify elements within images (Wu et al., 2020).

The approach for object detection is You Only Look Once (YOLO), which streamlines the process, making object detection more efficient and unified (Kundu, 2023). Among the versions of YOLO, YOLO v8 ensures efficient and better real-time performance (Bhalerao, 2023). By incorporating this model, the plugin is expected to effectively moderate content, thereby fostering a safer online environment.

The utilization of the YOLO v8 variation Nano is anticipated to enhance the plugin's ability to swiftly and accurately detect objectionable content (Rath, 2023). Through this approach, the study addresses the challenge of increasing objectionable images online and provides a practical solution for content moderation.

Figure 1 illustrates the architecture of the algorithm that was used for the study, which was the YOLOv8 based on Wang et al. (2023). The YOLOv8 algorithm was built upon a convolutional neural network (CNN) that contains 3 modules, which are the Backbone, Neck, and Head. Mainly, the YOLOv8 algorithm can be divided into two main parts: the backbone and the head.

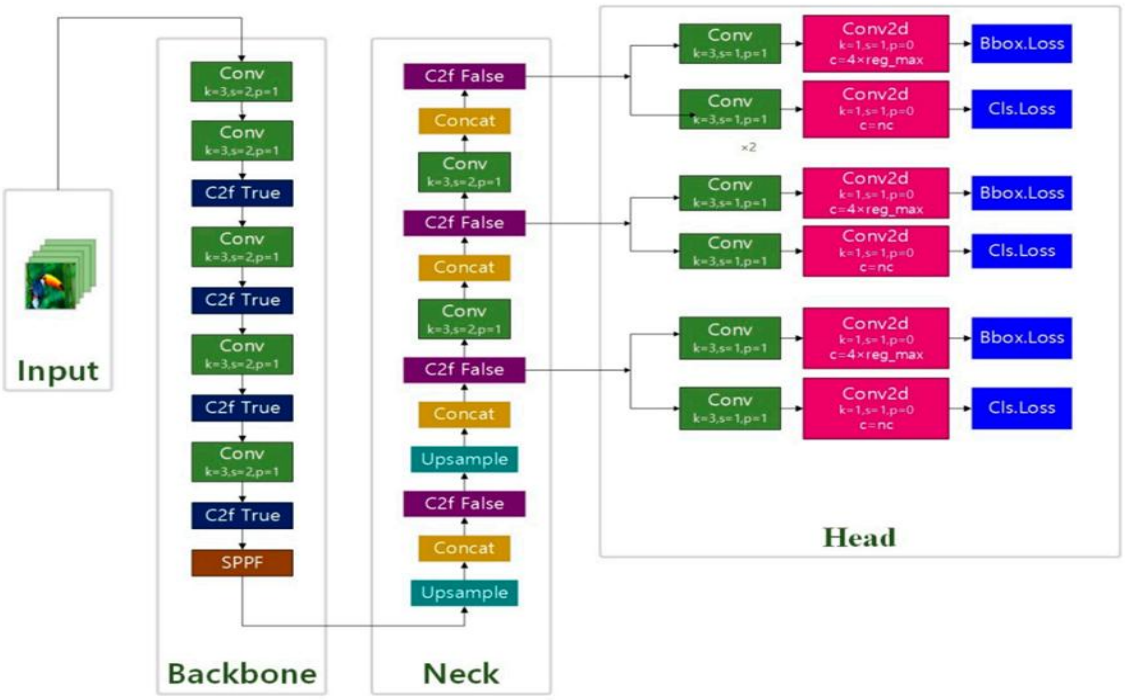
The backbone of YOLOv8 is a modified version of the CSPDarknet53 architecture. This architecture consists of 53 convolutional layers and employs cross-stage partial connections to improve information flow between the different layers. Specific changes in the backbone network and neck module from YOLOv5 to YOLOv8 include the kernel of the first convolutional layer has been changed from 6x6 to 3x3 and all C3 modules are replaced by C2f, resulting in more skip connections and additional split operations.

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Figure 1

YOLOv8 Architecture



YOLOv8 Algorithm

The neck module in YOLOv8 is based on the YOLOv7 ELAN design concept, replacing the C3 module of YOLOv5 with the C2f module. However, this module includes a lot of operations, such as split and concat that are not as deployment-friendly as before.

The head of YOLOv8 consists of multiple convolutional layers followed by a series of fully connected layers. These layers are responsible for predicting bounding boxes, abjectness scores, and class probabilities for the objects detected in an image. The key feature of the head module in YOLOv8 is the use of a self-attention mechanism, which allows the model to focus on different parts of the image and adjust the importance of different features based on their relevance to the task. And the ability to perform multi-scale object detection using a feature pyramid network. This network consists of multiple layers that detect objects at different scales, allowing the model to detect large and small objects within an image. The head module in YOLOv8 has been changed from the original coupling structure to the decoupling one, and its style has been changed from YOLOv5's Anchor-Based to Anchor-Free.

The architecture of YOLOv8 is highly customizable, allowing users to easily modify the model's structure and parameters to suit their needs. It supports various backbones, such as EfficientNet, ResNet, and CSPDarknet, giving users the flexibility to choose the best model for their specific use case.

Objective of the Study

The general objective of this study is to develop an efficient plugin capable of detecting and censoring objectionable images in real-time, thereby preventing such content from being displayed and enhancing the overall user experience.

Specifically, the study aims to achieve the following objectives:

1. Develop a robust deep learning model based on the YOLO v8 architecture, optimized to efficiently detect a wide range of inappropriate and harmful images across web platforms by training the model to recognize various categories of offensive content.
2. Design and develop a browser plugin that integrates the YOLO-based detection model into Google Chrome, with real-time censorship of objectionable images and user access controls for sensitive content.

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

3. Evaluate the performance of the YOLO v8-based plugin using the YOLO v8 Library from Ultralytics, specifically assessing its detection capabilities through metrics such as Mean Average Precision (mAP).
4. Evaluate the plugin's quality and acceptability using a researcher-designed survey aligned with ISO/IEC 25010:2023 standards, focusing on functional suitability, performance efficiency, compatibility, usability, reliability, maintainability, flexibility, security, and safety.

METHODOLOGY

The researchers utilized a quantitative descriptive approach to thoroughly determine the performance of each model, featuring YOLO v8, on web image detection and censorship. Using standardized datasets and random assignment of images, the study aims to ensure reliability, attributing any observed performance variations specifically to the model under scrutiny. Multiple iterations were conducted to ensure reliability and facilitate result reproducibility. The system development process model used in this study was the Agile Model. Agile is a dynamic and flexible approach to software development that emphasizes collaboration, adaptability, and incremental progress.

Fundamental Algorithm Used

The development of iCensr utilized the YOLOv8 Algorithm in the creation of the software project, enabling real-time object detection with high accuracy and efficiency. By leveraging the advancements in deep learning, YOLOv8 was particularly well-suited for filtering inappropriate images from the web thanks to its speed and precision. This algorithm was integrated into iCensr to identify and categorize explicit content, ensuring a reliable and robust filtering mechanism. The project also incorporated techniques such as transfer learning to fine-tune the model on specific datasets, enhancing its performance in detecting nuanced and context-specific content.

Figure 2.

Yolov8 Prediction Simulation

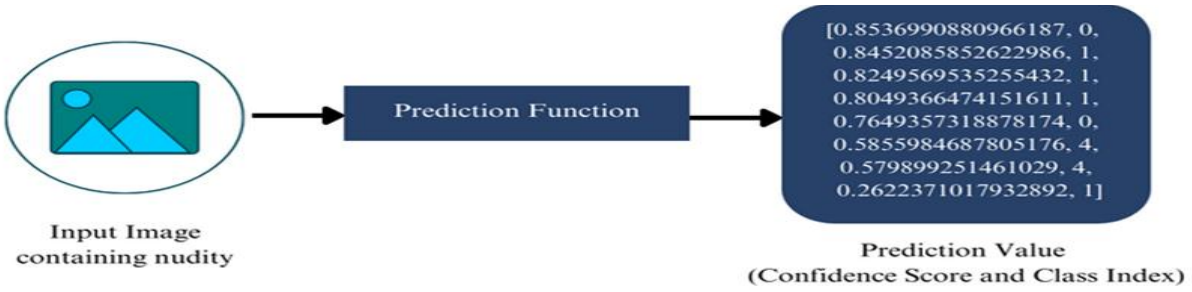
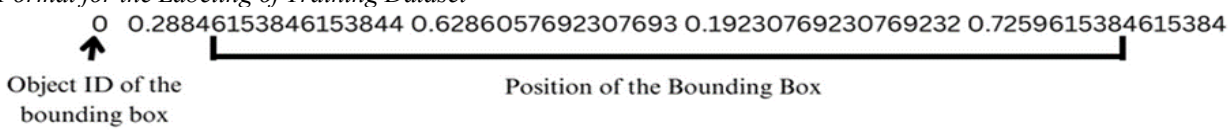


Figure 2 shows the researcher's simulation through Yolov8 Models that was utilized in the iCensr browser plugin. The following list would give context to the values of the Class Index mentioned in the figure, names: ['EXPOSED_BELLY', 'EXPOSED_BREAST_F', 'EXPOSED_BREAST_M', 'EXPOSED_BUTTOCKS', 'EXPOSED_GENITALIA_F', 'EXPOSED_GENITALIA_M']. It is demonstrated in the figure that YOLO v8 would produce the confidence score for each class detected in the input image. The confidence score would be read as a type of percentage, so if the confidence score was 0.8536… on class index 0, it is telling us that the model is 85.36% sure that the object it is detecting was named "EXPOSED_BELLY," since the class index 0 in the used Yolov8 model was named "EXPOSED_BELLY." ICensr would only fetch the confidence score and class index from the model, as it would be the values to be used in determining whether the content should be censored or not.

Figure 3

Format for the Labeling of Training Dataset



Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Figure 3 shows the values of the declared bounding box and what it would look like. Object ID would point at the classes you'll declare on the ".yaml" that would contain the directory for your training data and validation data and the name of the classes on your model. This step is necessary when you want to use your own annotated dataset.

The study used the annotated datasets that were provided publicly by Roboflow Users Tiem for the Nudity dataset, jaishreeram for the Violence dataset, and CBAIT for the Illicit Drugs dataset rather than creating these datasets from scratch.

$$m = [p(c_1), p(c_2), \ldots, p(c_n)] \qquad (1)$$

Equation 1 would act as a set that would hold the value of the declared classes mentioned during the training. $p(c_n)$ aims to represent the value of a class, in the context of this study, the value of it can be the object cannabis, violence, human interaction or human genitalia.

$$o = [x, y, \sqrt{w}, \sqrt{h}] \qquad (2)$$

Equation 2 was also a set that aims to hold the result bounding box produced by the YOLOv8 model predictions or the training bounding box fed to the YOLOv8 algorithm in model training.

Figure 4

Sample visualization of Bounding Box Position

Source : (https://docs.ultralytics.com/datasets/detect/#ultralytics-yolo-format)

Figure 4 shows the graphical representation of how bounding box values were used. Values x, y, √w, and √h show the positional values of the bounding box produced by the prediction or the bounding box used to train a YOLOv8 model. The x and y show the location of the center point of the bounding box, while √w shows the width of the bounding box, and √h shows the height of the bounding box.

$$n_{out} = [(n_{in}) + 2p - k)/s] + 1 \qquad (3)$$

Equation 3 was used to calculate the output feature to be used during the training process. Extracted features would then be used on the Conv Block to describe the characteristics that the image has for the algorithm to create predictions. This formula was utilized during the usage of the Conv Block based on the YOLOv8 Algorithm Architecture.

$$SiLU(x) = x\,(1/(1+e^{e-x})) \qquad (4)$$

Equation 4 denotes the Activation Function used in the Conv Block, which was the SiLU Function. SiLU Function is one process in YOLOv8's Conv Block wherein it would act as its activation function for the processed image. This activation function would help enhance the network's ability to learn and represent complex relationships in image data, leading to improved performance in tasks such as object detection.

$$mAP = 1/n \; \Sigma\_(i=1)^{n} AP^{i} \qquad (5)$$

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

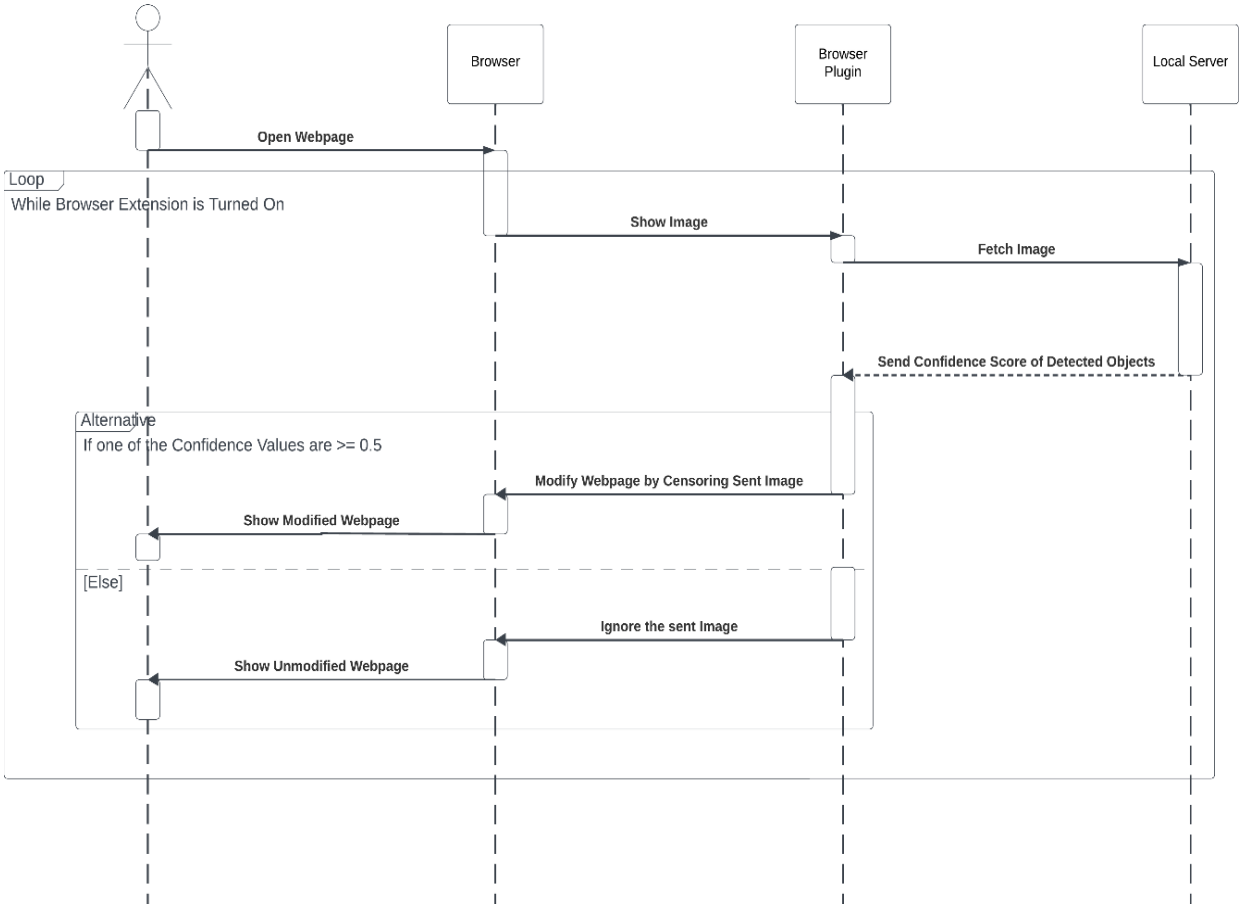
Equation 5 shows the equation for calculating the mAP. The researchers used mAP to evaluate the performance of the model through average precision across all categories, providing a single value to compare different models. mAP incorporates the precision-recall curve that plots precision against recall for different confidence thresholds that provide a balanced assessment of precision and recall by considering the area under the precision-recall curve. Also, mAP handles multiple object categories by calculating each category's average precision separately and taking their average across all types (hence the name mAP).

System Design

Figure 5 illustrates the Sequence Diagram of the iCensr Browser Plugin. This diagram outlines the sequence of events and message exchanges between key components of the system. The primary actor in this diagram is the end-user of the iCensr Browser Plugin. The objects involved are the browser application, the iCensr browser plugin, and the server responsible for running the YOLOv8 object detection script.

Figure 5

Sequence Diagram of ICensr Plugin



The process begins when the iCensr Plugin detects that a webpage with images is opened in the browser, provided the plugin is active. The plugin then sends the image to the server for analysis, where the YOLOv8 script evaluates the image and generates a confidence score for the detected content. If the confidence score is below 0.5, the plugin modifies the webpage to display a censored version for the end-user. This process continues until the end-user deactivates the plugin.

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Figure 6

Activity Diagram of ICensr Plugin

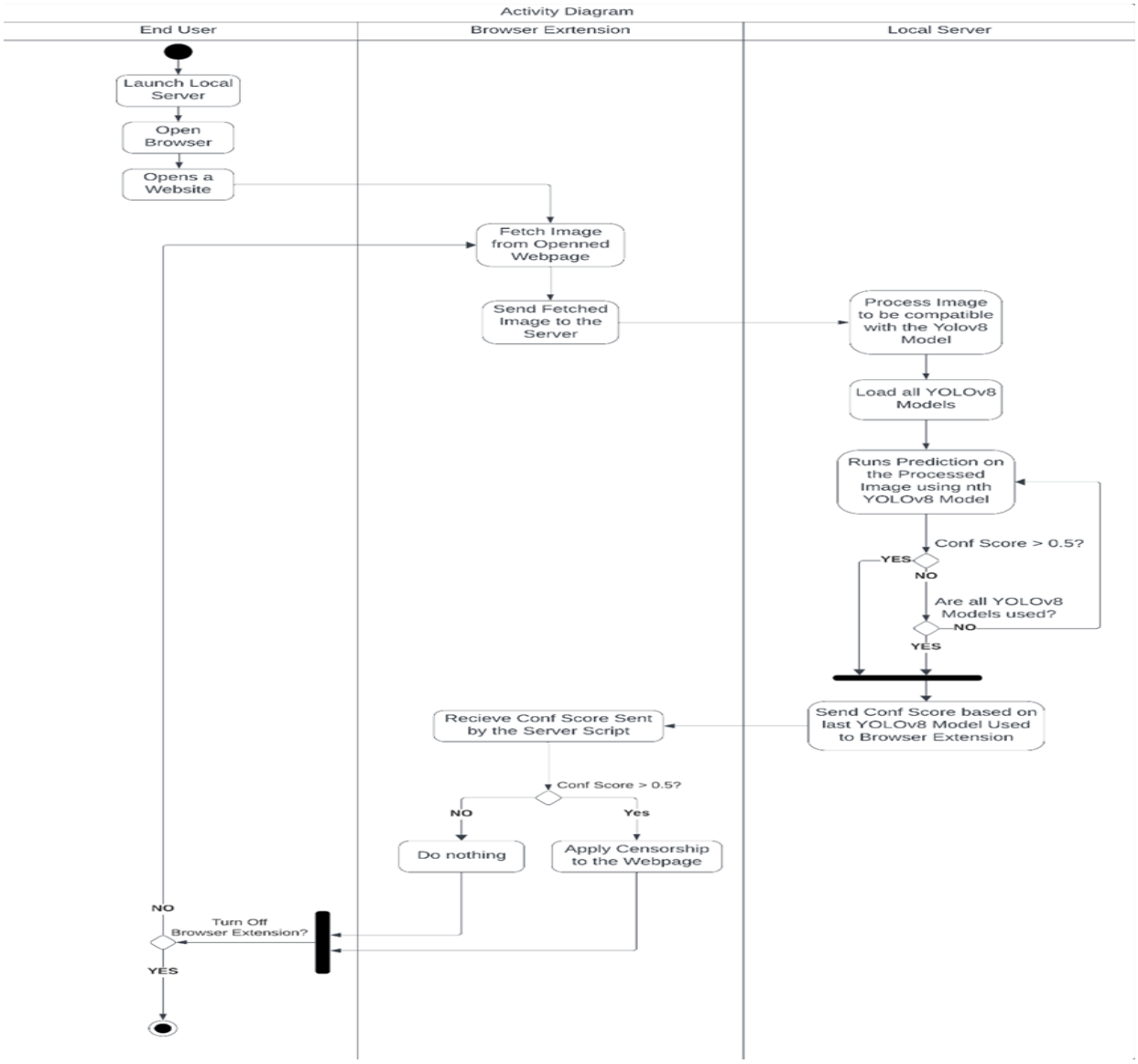


Figure 6 presents a rough sketch of the processes within the iCensr browser plugin, represented through an Activity Diagram. The swimlanes show the key actors and objects involved: the End User, the iCensr Browser Plugin, and the Server Script responsible for running the YOLOv8 object detection. According to the diagram, the iCensr Browser Plugin begins by fetching images from a webpage once the end-user opens a page containing images. These images are sent to the Server Script, where object predictions are made. The server then analyzes the confidence score of each prediction to determine if it qualifies as "Mature Content" (i.e., Confidence Score ≥ 0.5). If the score is below 0.5, the plugin checks additional YOLOv8 models to categorize the content (e.g., nudity, violence, illicit drugs). Once the server has either detected mature content with a sufficient confidence score or exhausted all models with low scores, it sends the confidence score back to the browser plugin. The plugin then evaluates whether the confidence

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

score is high enough to warrant censorship. If it is, the plugin modifies the webpage to censor the image. If not, no modification is made. The plugin operates continuously unless manually deactivated by the end user.

Figure 7

System Architecture of iCensr

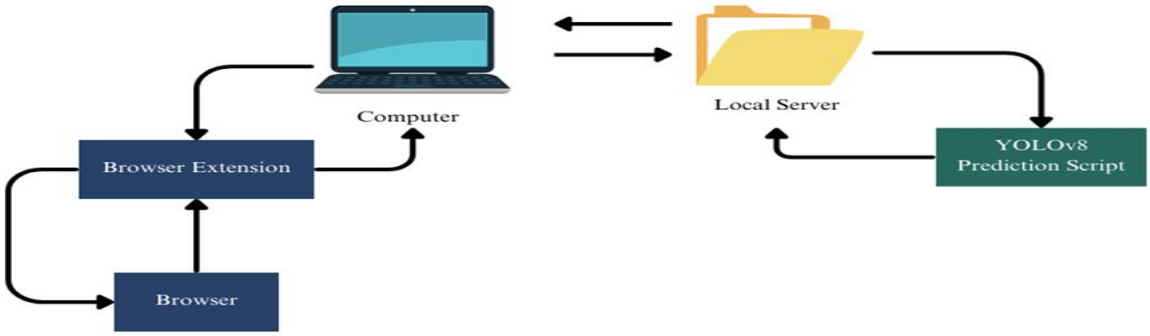


Figure 7 shows the flow of the processes when iCensr was used. The system architecture illustrates that the Browser Extension first acquires the data in the form of an image and then requests the computer to send data to the Local Server. The Local Server then ran the yolov8 prediction script to determine whether the received image was objectionable or not. Once the results were produced by the prediction script, the result was then sent by the Local Server back to the Browser Extension. The Browser Extension decides whether to censor the processed image on the browser based on the results received from the Local Server. The cycle of this process is to be repeated and can only be stopped when the user decides to close the browser or turn off the actual browser plugin.

Participants of the Study

A total of 58 participants took part in the study. This group included Technical-Vocational Livelihood (TVL) Information Communication and Technology (ICT) students from Tanza National Trade School, as well as 15 IT professionals, specifically software developers. All participants were regular users of online media and had likely encountered objectionable images, making them ideal candidates for the study. Their diverse experiences provided valuable insights into the performance and effectiveness of the iCensr plugin.

Research Instruments

Data were gathered through a survey questionnaire implemented using Forms to assess the efficacy of the ICensr Browser Plugin. Participants were required to use the plugin for at least 15 minutes on their web browser and then complete a survey questionnaire. The survey follows the structure outlined in ISO/IEC 25010 2023, including an introduction about the survey and plugin, data privacy information, participant personal details, and a Likert scale to evaluate the plugin's performance.

Table 1

Verbal Interpretation of the Mean Interval for Acceptability of iCensr

Point	Mean Interval	Verbal Interpretation
4	3.26-4.00	Highly Acceptable
3	2.51-3.25	Acceptable
2	1.76-2.50	Unacceptable
1	1.00-1.75	Highly Unacceptable

Table 1 shows the levels of acceptability of the software evaluated. From “1.00” to “1.75,” it is “Totally Unacceptable,” which means the application might have failed to execute its main functions; “1.76” to “2.50” is

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

"Unacceptable," which means the application has executed but lacks the consistency to function well; "2.51" to "3.25" is "Acceptable," which means that the features of the application almost meet its expected result; and lastly, "3.26" to "4.00" is "Perfectly Acceptable," where the overall functions and objectives of the application work smoothly.

RESULTS AND DISCUSSION

The researchers developed a web browser plugin that can detect web images and censor objectionable types of images in real-time. The plugin is available on the iCensr website or Chrome Web Store. The algorithm used in this study is YOLO v8 (You Look Only Once version 8) to utilize the model's high performance.

Figure 8 displays the landing page of iCensr's official website. It provides general information about the iCensr web browser plugin and the development team. Users can interact with the website by selecting "Get Started," "Rate Us," or "Download." The "Get Started" button guides users on installing and using the iCensr plugin. The "Rate Us" button directs users to the website's feedback form section. Finally, the "Download" button redirects users to the Chrome Web Store, where they can find the iCensr browser plugin.

Figure 8

iCensr Official Website Landing Page

Figure 9 shows the interfaces of the iCensr plugin. When the plugin is first launched, the user is prompted to set a password, which is required for every access to the plugin. After successful authorization, the user can enable or disable censoring by toggling a switch in the interface. The interface also includes three buttons for additional interactions: "Change Password," "Give us Feedback," and "Terms of Use." The "Change Password" button allows the user to update their current password. The "Give us Feedback" button redirects the user to the iCensr plugin's website, specifically to the feedback section. Similarly, the "Terms of Use" button takes the user to the terms of use page of the website.

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

IJITGEB

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Figure 9

iCensr Plugin

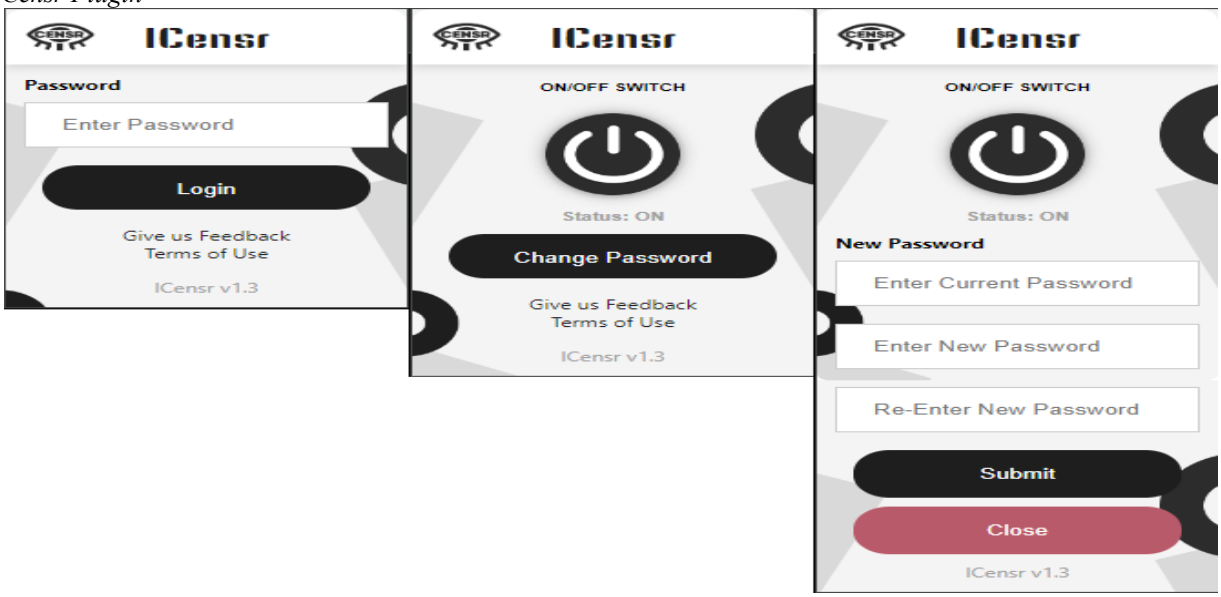


Figure 10

iCensr Plugin Preparation

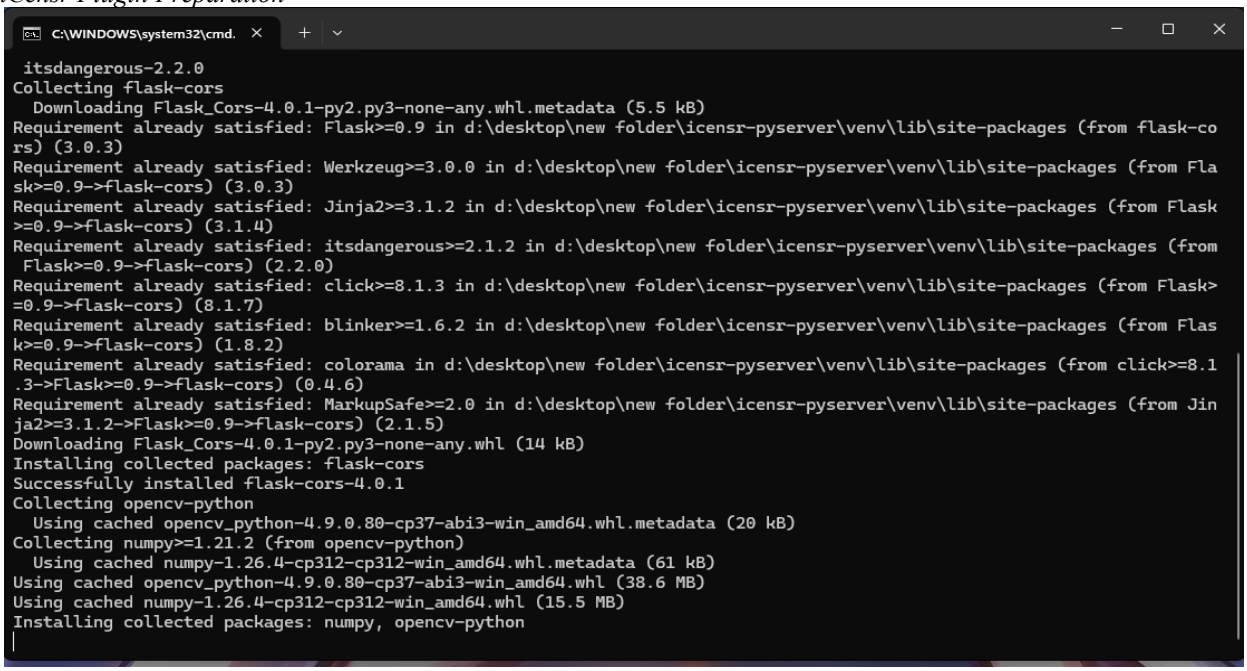


Figure 10 illustrates the setup process for the plugin, which involves installing a local server. This is achieved by executing a .bat script that installs the required Python library into a designated folder for the local server script. The download and installation of these Python libraries consume 2GB of the user's storage. Subsequently, running the start-terminal.bat file is necessary to activate an effective censoring and detection model when initiating the plugin.

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Figure 11

Screenshot of iCensr Plugin

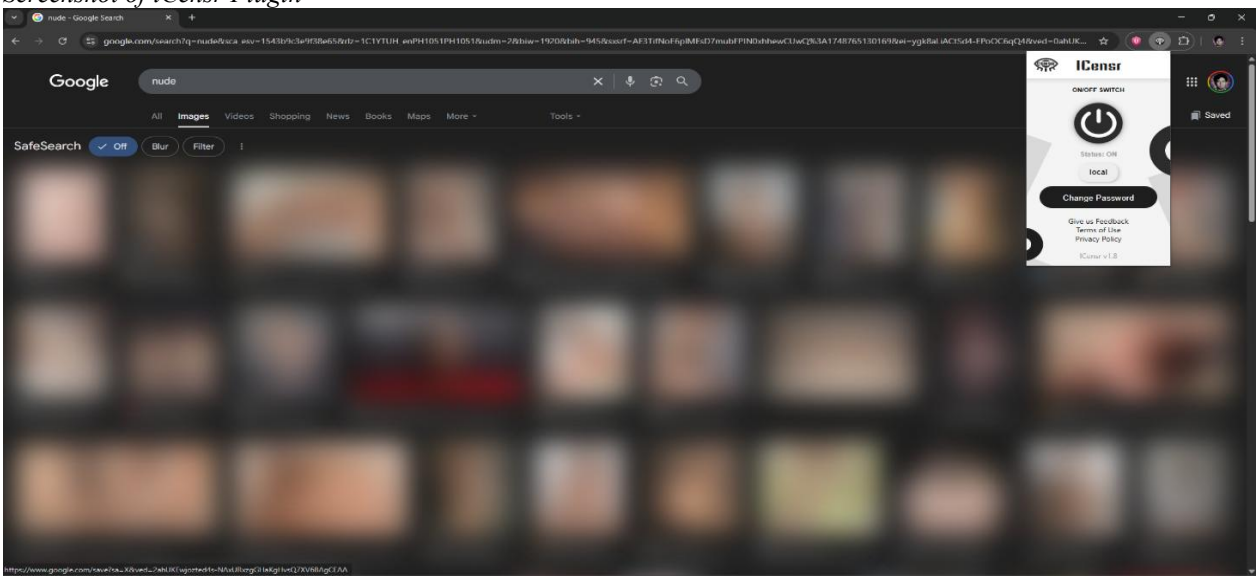
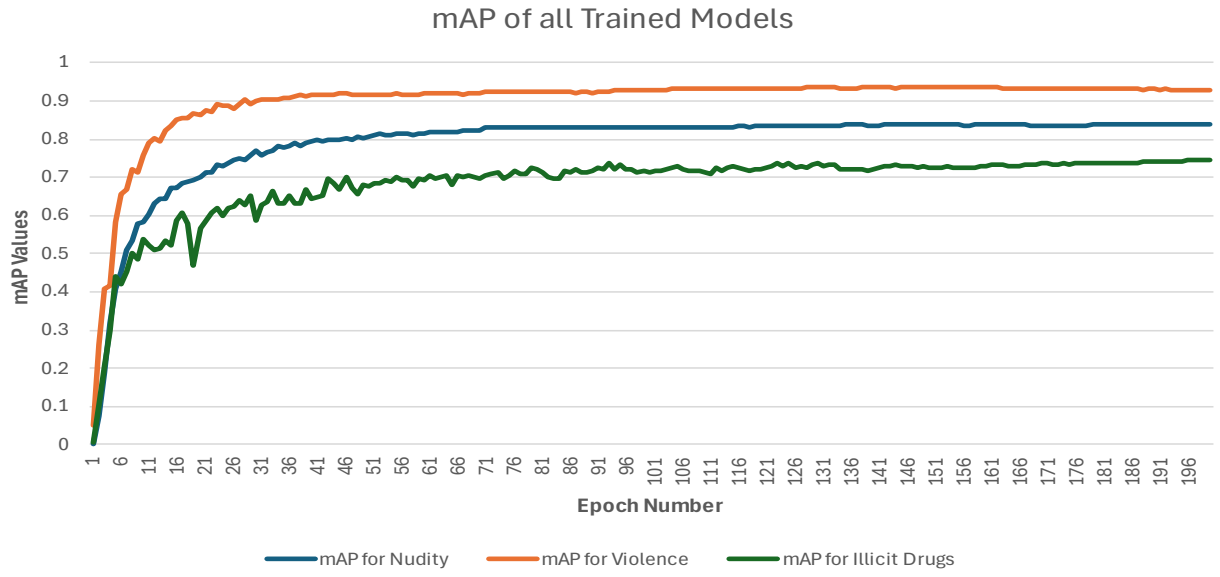


Figure 11 shows the iCensr plugin censoring and blocking images that are deemed objectionable by the object detection models in real-time. The plugin is designed to blur all the images on a website first before running the object detection models on each image. It is designed this way to avoid accidental viewing of undesired images.

The researchers evaluated the performance of the YOLOv8 model integrated with the plugin using the Ultralytics YOLOv8 Library. Specifically, acquiring the mean average precision (mAP) for three categories: Nudity, Violence, and Illicit Drugs during the training process. The Models trained on each corresponding dataset provided from Roboflow and the training set to 200 epochs. This data is to show the performance of each model used for the ICensr Plugin in detecting images.

Figure 12

mAP of All Models: Illicit Drugs, Nudity & Violence



Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

In the Figure 12, the relationship between the mAP and the epoch number and how the mAP Value increases as the epoch number also increases is displayed. It shows that throughout the training of the Nudity model, it achieved an mAP value of 0.8382 or 83.82%, while the Violence model achieved an mAP value of 0.9271 or 92.71%, indicating that the Nudity and Violence Models have a good performance. The Illicit Drugs model, however, achieved an mAP value of 0.7459 or 74.59%, even though the lowest, the Violence Model, also has an acceptable performance. This suggests that there may be improvements to be made for the model of Illicit Drugs.

Table 2

Evaluation Results from IT Experts

Criteria	M	Interpretation
Functional Suitability	3.29	Highly Acceptable
Performance Efficiency (Plugin)	3.17	Acceptable
Performance Efficiency (Website)	3.20	Acceptable
Compatibility	3.40	Highly Acceptable
Interaction Capability (Plugin)	3.35	Highly Acceptable
Interaction Capability (Website)	3.41	Highly Acceptable
Maintainability	3.31	Highly Acceptable
Reliability (Plugin)	3.30	Highly Acceptable
Reliability (Website)	3.30	Highly Acceptable
Flexibility	3.24	Acceptable
Security	3.43	Highly Acceptable
Safety	3.40	Highly Acceptable
Average Mean	3.32	Highly Acceptable

Note. *This table presents the evaluation results from IT experts based on various software quality criteria. The interpretation of mean scores follows a 4-point Likert scale: 3.26–4.00 = Highly Acceptable, 2.51–3.25 = Acceptable, 1.76–2.50 = Unacceptable, 1.00–1.75 = Highly Unacceptable.*

Table 2 shows the summary of the results of the evaluation, which involves a total of 15 IT Experts. It shows that the IT Experts deemed the plugin, iCensr acceptable based on ISO/IEC 25010:2023. The evaluation of the iCensr plugin shows that the IT Experts rated Security as the highest mean, indicating that the IT Experts highly agree that iCensr excels at having confidentiality in terms of personal data and integrity. That is because, in terms of confidentiality, the IT Experts agree that the data or images that are being processed by the plugin are not being stored, preventing outside threats, as the whole process happens on the user's device. In integrity, the plugin provides a password input to be able to access and modify controls. Moreover, the website's interaction capabilities were highly rated as well. The experts found the website's features and functionalities to be sufficient, user-friendly, visually appealing, and fully operational. On the other hand, the lowest ratings among the factors of the evaluation are the Performance Efficiency of both the Plugin and the Website. Although the means of both Performance Efficiency are the lowest, it is still interpreted as Highly Acceptable, being in the lowest rating suggests that there can be some improvements that can be made. The cause is that the plugin does not have 100% mAP and sometimes makes mistakes that produce false positives. The website's performance efficiency is rated lower than others, probably due to being hosted on a free platform and containing too many animations, which impacts its overall functionality.

Table 3 shows the summary of the results of the evaluation, which involves a total of 43 respondents. It shows that the end users deemed the plugin, iCensr, acceptable based on ISO/IEC 25010:2023. The evaluation by the end users of the project shows that Safety and Compatibility scored the highest among the factors within the software evaluation. Results indicate that most end users agreed that iCensr can successfully protect users from objectionable images from the web, as it can effectively censor. It also shows the end users agreed that the plugin is compatible with working alongside other software such as a browser application, websites, and other plugins.

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Table 3

Evaluation of Results of End Users

Criteria	M	Interpretation
Functional Suitability	3.33	Highly Acceptable
Performance Efficiency (Plugin)	3.21	Acceptable
Performance Efficiency (Website)	3.19	Acceptable
Compatibility	3.42	Highly Acceptable
Interaction Capability (Plugin)	3.34	Highly Acceptable
Interaction Capability (Website)	3.41	Highly Acceptable
Reliability (Plugin)	3.28	Highly Acceptable
Reliability (Website)	3.29	Highly Acceptable
Flexibility	3.26	Highly Acceptable
Security	3.40	Highly Acceptable
Safety	3.42	Highly Acceptable
Average Mean	3.33	Highly Acceptable

Note. *This table presents the evaluation results from IT experts based on various software quality criteria. The interpretation of mean scores follows a 4-point Likert scale: 3.26–4.00 = Highly Acceptable, 2.51–3.25 = Acceptable, 1.76–2.50 = Unacceptable, 1.00–1.75 = Highly Unacceptable.*

Evaluation results also show that the factors that are rated the lowest are also the Performance Efficiency of both the Plugin and Website. While still being interpreted as Highly Accepted, being in the lowest rating suggests that, like in IT Expert, there can be some improvements that can be made to the performance of the plugin, such as further increasing the accuracy of the object detection models, and to the website, such as hosting the website on a better website hosting platform.

CONCLUSION

The iCensr plugin, developed for real-time web image detection and censorship, demonstrates a promising solution for enhancing online safety by identifying and blocking objectionable content such as nudity, violence, and illicit drugs. The integration of the YOLO v8 image detection algorithm proved to be effective in detecting such content, with varying levels of performance across the three categories. The results found that the YOLOv8 Algorithm is an effective option for an Image Censoring Plugin. Data collected using the YOLOv8 object detection model, with datasets provided by various users through Roboflow, demonstrated satisfactory quality. Specifically, the model achieved a mean Average Precision (mAP) of 83.82% for detecting nudity, 92.71% for violence, and 74.59% for illicit drugs. The evaluation results according to the ISO/IEC 25010:2023 standard showed that the project is considered Highly Acceptable by both groups. IT experts rated it with a grand mean of 3.32, indicating it as "Highly Acceptable," while end users rated it slightly higher at 3.33, which also falls under the "Highly Acceptable" category.

RECOMMENDATION

Based on the findings of this study, it is recommended that the iCensr plugin be further developed and expanded to enhance its functionality and coverage. Future efforts should focus on incorporating additional categories of objectionable content, such as hate speech and graphic violence, to broaden its protective capabilities. Additionally, it would be beneficial to integrate improvements in the model's performance, specifically targeting lower Confidence Scores for more accurate image detection and censorship.

To maximize its impact, a mobile version of the iCensr plugin could be developed, extending its utility to users beyond desktop platforms. Additionally, refining the plugin's usability, based on feedback from end users, would help in ensuring its accessibility and ease of use for a broader audience. As online content continues to evolve,

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

continuous updates to the datasets and the underlying YOLO v8 model are necessary to maintain the plugin's effectiveness in addressing emerging forms of objectionable content.

Finally, further research could explore integrating the iCensr plugin with other online platforms, such as social media and content-sharing websites, to provide comprehensive content moderation and create safer online spaces across the internet.

REFERENCES

Bhalerao, C. (2023). YOLO V8: The real state-of-the-art? *Medium.* https://medium.com/mlearning-ai/yolo-v8-the-real-state-of-the-art-eda6c86a1b90

Bhatti, A., Umer, M., Adil, S., Ebrahim, M., Nawaz, D., & Ahmed, F. (2018). Explicit content detection system: An approach toward a safe and ethical environment. *Applied Computational Intelligence and Soft Computing, 2018,* 1–13. https://doi.org/10.1155/2018/1463546

Bicho, D., Ferreira, A., & Datia, N. (2020). A deep learning approach to identify not suitable for work images. Retrieved from https://repositorio.ipl.pt/bitstream/10400.21/12354/1/A%20deep_AFerreira.pdf

Boesch, G. (2023). Object detection in 2023: The definitive guide. *Viso AI.* https://viso.ai/deep-learning/object-detection/

Dixon, S. J. (2023). Global daily social media usage 2023. *Statista.* https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Facing History & Ourselves. (2016). Visual essay: The impact of propaganda. *Facing History & Ourselves.* https://www.facinghistory.org/resource-library/visual-essay-impact-propaganda

Gillis, A. S., Burns, E., & Brush, K. (2023). What is deep learning and how does it work? *TechTarget.* https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network

Hussain, M. (2023). YOLO-v1 to YOLO-v8: The rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines, 11*(7), 677. https://doi.org/10.3390/machines11070677

Internet Matters. (2018). What is inappropriate content? *Internet Matters.* https://www.internetmatters.org/issues/inappropriate-content/learn-about-it/

Izzah, N., Budi, I., & Louvan, S. (2018). Classification of pornographic content on Twitter using support vector machine and Naive Bayes. In *2018 4th International Conference on Computer and Technology Applications, ICCTA 2018* (pp. 156–160). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CATA.2018.8398674

Khan, M. A., Rawan, B., & Ullah, A. (2020). Growing up with media violence and psychological trauma among youth in Pakistan. *Pakistan Journal of Criminology, 12*(1), 79–88.

Kim, Y., Kim, T., & Yoo, S. E. (2022). TsCNNs-based inappropriate image and video detection system for a social network. *Journal of Information Processing Systems, 18*(5), 677–687.

Kundu, R. (2023). YOLO algorithm for object detection explained [+examples]. *V7 Labs.* https://www.v7labs.com/blog/yolo-object-detection#:~:text=a%20negative%20prediction.-,What%20is%20YOLO%3F,repurposed%20classifiers%20to%20perform%20detection

Lin, W. H., Liu, C. H., & Yi, C. C. (2020). Exposure to sexually explicit media in early adolescence is related to risky sexual behavior in emerging adulthood. *PLOS ONE, 15*(4), e0230242. https://doi.org/10.1371/journal.pone.0230242

Rath, S. (2023). YOLOv8: Comprehensive guide to state-of-the-art object detection. *LearnOpenCV.* https://learnopencv.com/ultralytics-yolov8/

Severen, B. V. (2022). Top 10 dangers of the internet for children: *MVS Legal.* https://milwaukee-criminal-lawyer.com/what-are-the-dangers-of-using-the-internet-for-kids/

Stubbs, J., Nicklin, L., Wilsdon, L., & Lloyd, J. (2022). Investigating the experience of viewing extreme real-world violence online: Naturalistic evidence from an online discussion forum. *New Media and Society.*

Vinney, C. (2023). Here's how violent media can impact your mental health. *Verywell Mind.* https://www.verywellmind.com/what-is-the-impact-of-violent-media-on-mental-health-5270512

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025

Wu, D., Lv, S., Jiang, M., & Song, H. (2020). Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Computers and Electronics in Agriculture, 178,* 105742. https://doi.org/10.1016/j.compag.2020.105742

Zulfiqar, S. H. (2021). Does media violence cause violence? Can exposure to violent TV shows, movies, and video games turn people aggressive and desensitized to violence? *Eliva Press.* https://books.google.com.ph/books?id=14w2EAAAQBAJ

Received Date: April 8, 2025 / **Revised Date:** May 2, 2025 / **Accepted Date:** June 7, 2025