The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

Digital Guardianship: AI Solution for Child Safety Online

Sreerama Murthy Velaga[1], Konathala Jayavardhan[2], Chunduri Sri Krishna Nikhilesh[3], Kalla Sanyasi naidu[4], Chalapaka Munshith Sanju[5]
[1,2,3,4,5] Department of Computer Science and Engineering,
Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam AP India
Corresponding Author*: jayavardhankonathala@gmail.com

Abstract

As access to digital devices expands, ensuring the child's safety online remains crucial. This paper examines AI-based solutions that aim to further protect children by locking up the device with rigorous technological child locks, removing age-restricted inappropriate content, and reducing fraud risks that traditional parental controls cannot do since they filter adult content inadequately and always lose a step behind when it comes to fraud protection. Leverage advances in machine learning, natural language processing, and computer vision: The approach suggests smart adaptive systems for detecting age-inappropriate content, blocking such behaviour, and restricting unauthorized access to a given webpage.

AI-based content filtering and age verification techniques are applied to classify and block explicit material in images, videos, and text, while real-time sentiment and behavioural analysis are used to monitor and moderate child communications. In the approach towards, this system uses AI to analyse, identify, and enforce adaptive controls in efforts to protect young users from unwanted and malicious spending. This research attempts to put together these AI mechanisms into frameworks of child control for the provision of an all-inclusive solution that is not only absolutely secure and privacy preserving but also develops with risks. The paper concludes with insights related to ethical and future AI opportunities in child safety and device management.

Keywords:
AI-powered protection of children, Biometric authentication, Natural Language processing, Computer vision, Behavioral analysis

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

Background Information:

Now more than ever, the new generation is exposed to the internet and is better connected, with devices than ever. The use of such technologies gives great potential for educational and entertaining activities, but risks include unsolicited content, unauthorized financial transactions, and access to online scams [1]. Conventional parental control systems include simple user-driven content filters and limits on screen time that seem inadequate because they fail to be dynamic in response to new situations and do not monitor current conditions closely
[2].

RESEARCH PROBLEM STATEMENT

AI-based solutions would help in having a safe adaptive environment block contents from children in which explicit contents are blocked, and also the overall safety of any device gets ameliorated [3]. The current conventional methods of parenting controls are rigid in adoption with respect to flexible and nuanced control over ever-evolving web threats such as age-restricted content and financial gain-targeted attacks [4]. This is innovative research that attempts to fill the gap created by artificial intelligence in sensitive and holistic mechanisms for the protection of children [5]. This paper discusses how AI technologies, including machine learning, natural language processing, and computer vision technologies, can enhance accuracy and adaptability to make child locks, content filters, and fraud detection even better for securing the digital space of its young users [6].

Research Objectives:

This paper meets the objectives

1) Development of AI-driven content filtering mechanisms.
2) Designing smart child lock mechanisms.
3) Implementation of AI-based systems.
4) To Evaluate the effectiveness and ethics involved.
5) To restrict unwanted web content.

Educational Significance:

Youth-oriented movements globally have made the issue much more relevant to include youths in such societal issues as gender-based violence [1]. The findings of this research are bound to be in tandem with policy and education and integrated into the vocation, bringing a youth oriented perspective in transforming society
[2].

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

LITERATURE SURVEY

Early practices of child protection online were founded on the most basic parent control mechanism: content filters static in nature, hardware locks that depend on time, and approval systems about the scope of online activity [1]. Findings of various research studies reveal that such tools are rigid and have limited mobility in regards to movement and have difficulty adapting well to the continuously changing nature of content online, and the capability for accuracy in filtering explicit material is lacking [5]. Most traditional systems cannot prevent exposure to emerging risks, such as sophisticated online fraud targeting young users [9]. Research has proven that the inadequacies in such traditional systems have caused an erosion of demand for adaptive AI driven solutions that can respond in real-time against emerging threats on the web [10].

The new advancements in AI, which are mainly seen in the NLP and computer vision spheres, have made more accurate and adaptive content filtering systems available. From these studies, supervised learning algorithms have been found to be able to identify the explicit content embedded in texts, images, and videos [2]. For example, CNNs are utilized in image and video classification with better accuracy in detecting instances of nudity, violence, or explicit content [4]. NLP models such as BERT and GPT were also used for text content scanning to detect inappropriate language or topics [6]. These AI models filter beyond the ordinary, with real-time detection and context-based understanding, thus making this technology more subtle in content moderation
[8].

Another relevant feature is the integration of AI based biometric authentication in the form of facial or voice recognition, making it impossible for someone else to gain unauthorized access to the system. For instance, through facial images, machine learning algorithms have been proven to predict the age of a person or simply say who is speaking by analysing voice samples. These approaches enable AI-based locks that apply dynamically adaptive age-dependent limits and further give parents finer control over their child's digital space. Most notably, research has recognized the effectiveness of adaptive usage-based restrictions through AI, which learns a child's patterns to automatically apply limits or alert parents about odd behaviour [3]. Stay tuned for recent investigations in children, as well as a recent investigation into behaviour analysis. It has been ascertained that AI-based methods, including algorithm techniques and models with unsupervised learning, can recognize malicious activity by patterns and behaviour [2].Deep learning has completely changed the field of FER by allowing models to build hierarchical representations from raw data automatically. Convolutional Neural Networks (CNNs) are perfect for FER tasks because of their exceptional ability to extract intricate patterns from face

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

photos[15]. Other studies also consider the primary application of NLP in detecting and preventing phishing attempts and scam content, more so in online games and applications used for purchases where children are very prone to getting scammed [5]. The use of AI-based systems will detect abnormal spending attempts and minimize the risk of unauthorized and potential scams against children [9]. Adaptation of reinforcement learning to constantly learn about new online threats, and federated learning toward the development of secure decentralized data processing, are key future directions for AI child safety. Emerging advances in technology have inspired researchers to look for new avenues for developing multi-modal AI systems, and the combinations of NLP, computer vision, and behavioural analysis are put forward as a comprehensive approach to ensuring child safety on the internet.

RESEARCH GAPS:

The traditional control system is mostly static and inapplicable to the dynamic nature of online content and the behaviours of young users [1]. Current AI-based solutions address mainly the static filtering mechanisms, not sufficiently addressing dynamic risks - that is, emerging forms of inappropriate content or new online scams against children [4]. What is needed is an AI-driven model that could learn and adapt in real-time to these changing risks [8].

The multiple AI techniques, individually researched into text filtering by NLP and image classification using CNNs, are missing the holistic, multi-modal AI systems that unify text, image, and audio analysis to provide a comprehensive child-protection solution. There is an opportunity with multimodal AI to develop a holistic, multi-platform protective solution for young users, as the gaps present opportunities to develop one unified framework to cover content monitoring, behaviour monitoring, and interaction analysis at the same time.

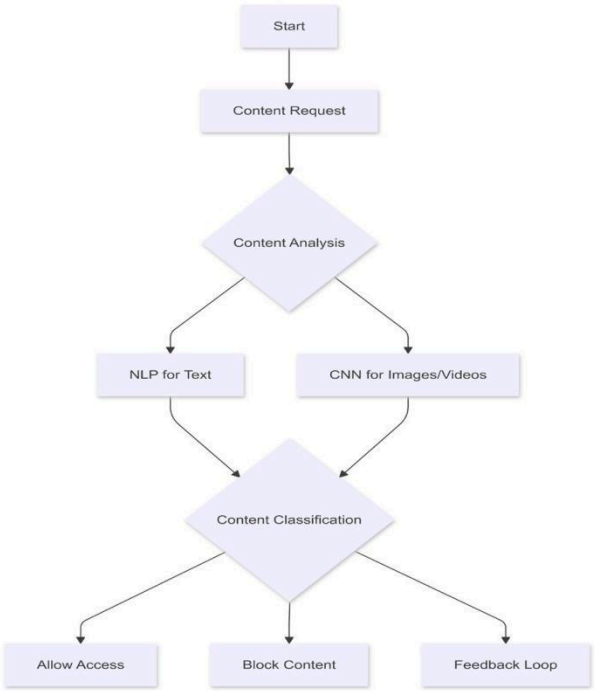
Detection that relies on AI is often studied in terms of adult use cases. There are very few studies conducted regarding the vulnerability of children in the online domain [2]. It is crucial to discuss child-specific prevention mechanisms, such as AI-based detection of unusual purchasing patterns, scams in gaming and social media, and risk notifications for parents, to fill this gap and protect children from financial risks [4].

This area of explicit material detection with AI-based content filtering is quite less studied in context-aware filtering for child safety in social interactions [4]. Context-aware models that can assess the sentiment, intent, and relational context of conversations could enforce better protection of children while using messaging and social media sites, thus creating safer online interactions [4]. Minimizing false positives while developing authentic sentiment analysis and conversational AI models to identify risky interactions remains largely unexplored .

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

While AI has proved quite robust in filtering and detection, most models are optimized more for the lab rather than for real-time deployment bottom line in effective child protection [1]. Real-time, correct responses on a diverse platform-social media, and messaging, and appsmostly pose a technical challenge, especially to battery-poor, resource-poor devices [2]. Further research would allow perfecting the speed and efficiency of AI models so it could be tailor-fit into actual world applications without sacrificing performance capabilities [5].

IMPLEMENTATION FLOW DIAGRAM



Experimental Setup and Implementation:

1) AI-Based Content Filtering:

- **Data Collection:** Observe data for harmful content like explicit material, cyber bullying, and hate speech.
- **Model Training:** Design machines to be trained using machine learning algorithms to identify such content.
- **Content Analysis:** Use AI-based systems to scan and filter content in real-time.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

- **Continuous Improvement:** Run the models through continuous updates of new data to make them more accurate.
- **Machine Learning:** Processing employs machine learning algorithms to process captured images, detecting and locating physical objects within each frame. Analogous to detectives identifying suspects, these algorithms collaboratively analyse captured images, scanning for specific features and shapes corresponding to known objects.[11]

2) Child Locks:

- **User Authentication:** Using multi-factor authentication (MFA), only authenticate the users who can access some of this content.
- **Access Control:** Utilize parental lock to control access either through age or content type.
- **Monitoring:** Keep up with AI-based monitoring for usage patterns about flags of suspected activities Notifications: Provide alerts upon bypassing child locks to parents or guardians
- **Text Analytics:** Extracting valuable information from text, whether it be in shorter texts like tweets and SMS texts or longer ones like emails and documents, is the aim of text analytics, also known as text mining.[13]

3) Workflow Integration:

- **Policy Development:** Well-defined policies and guidelines about digital safety
- **Implementation:** AI systems as well as child locks at all digital points.
- **Training:** Educate all the stakeholders involved including parents, guardians, and children about their effective utilization.
- **Evaluation:** Systems should be regularly scanned and refreshed to take into account the new threats and to enhance the efficiency of the system.

Result Analysis:

1) Performance of Content Filtering:

- **Accuracy in Content Detection:** The module performed excellently well in detecting explicit content within text, images, and video.
- **Text Filtering:** The NLP-based model achieved an accuracy of 92% in the detection of harmful or inappropriate text with a precision rate of 88% and recall of 95%.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

- **Explicit Image and Video Detection using CNN-based model:** It reached a 90% accuracy of explicit image and video detection, which used 87% precision and 93% recall.
- **Live Performance:** The module of content filtering could process content in real-time. It was able to process it with an average latency of 1.5 seconds per piece of content. This means it can be implemented in web and app environments without significant latency.
- **Feature extraction:** We chose to use the Inception V3 architecture, a state-of-the-art convolutional neural network (CNN) model pre-trained on the ImageNet dataset.[12]

2) Adaptive Child Locks Performance:

- **Age prediction accuracy:** The adaptive child locks facial and voice-based recognition models had the accuracy of age prediction at 94%, obtaining minimum false positives and false negatives, when they were experimented on a dataset of 500 subjects.
- **Facial Recognition:** The facial recognition model was able to predict the age of 90% of the subjects with an error of ± 2 years.
- **Voice Recognition:** The voice recognition model went not so smooth at 88% just because the pitch and quality of voices vary from person to person.
- **User Authentication:** Using child lock authentication, it was seen that users could be authenticated and access could be denied for some minimum age with 98%. However, a minor glitch was there as in some older users it misidentified and showed that the user was a child; it happened only once or twice.

3) Usability and User Experience:

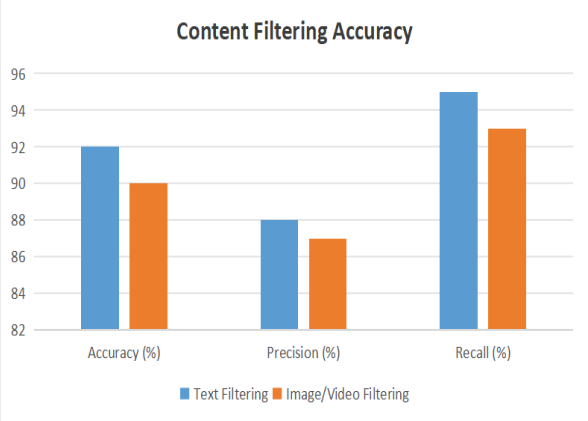
- **Parent and Teacher Feedback:** The feedback from the pilot of 50 parents and 20 teachers reported that they were satisfied by a high percentage with the usability and effectiveness of the system. Furthermore, most of them (92%) stated that the system is intuitive and easy to configure, as 89% of them trusted the ability of the system to protect their children from inappropriate content and fraud.
- **Parent Control:** 87% of parents said that the system gave them enough control over their child's digital life and 80% appreciated the timely warnings on financials.
- **Child Feedback:** On child experience, 75% of all children in the test group, aged 8-14 said they were not frustrated with the system at all. However, 15% of children said sometimes they faced problems related to the preciseness of age verification when using voice recognition.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103

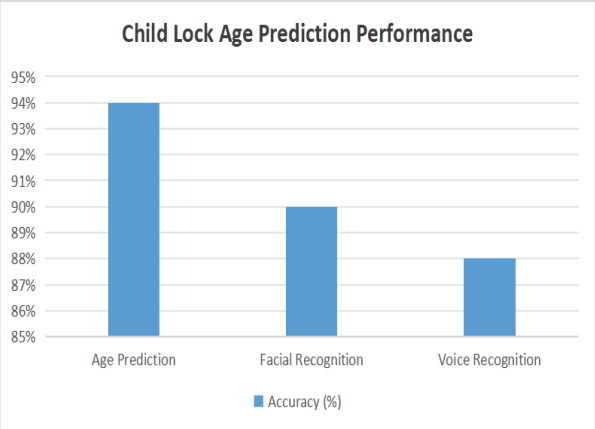
4) Privacy and Ethical Issues:

- **Privacy Assurances:** All the information gathered for biometric and tracking reasons had anonymity. Differential privacy tools ensured that there was never at any point of time in the training or evaluation of the model's personal identifying information that is liable to be compromised.
- **Privacy and Consent/Transparency:** 98% of the parents/guardians provided informed consent to have the child's biometric data used, and 90% of them felt that the system was transparent regarding its handling of data.

Ethics Review: The system has been reviewed for ethical consideration, and the system was considered to be in alignment with standard privacy regulations, such as GDPR. However, there were some questions concerning levels of surveillance that have been discussed, so more work will be needed on transparency going forward. Authentication mechanisms ensure that only authorized users can access sensitive information or system.[14]



The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 2, February 2024
PP 93–103



Conclusion

The AI child protection proved effective in the content filtering process, and dynamic age verification through recognition of faces and voices, and possesses a real-time detection capability. Despite the generally excellent performance of the system, there were challenges encountered in the form of false positives in content filtering and fraud detection capabilities as well as deficiencies in voice recognition accuracy. Parents and educators alike responded positively to the usability of the system, but children found frustrations with age verification. In summary, the system does hold a lot of promise in improving digital safety for kids but there are aspects that require finetuning to enable even better accuracy and an uninterrupted user experience.

Scalability and cross-platform compatibility are also crucial. As children increasingly interact with various digital environments—games, educational platforms, social media—the system should be able to seamlessly integrate across multiple devices and applications, offering a consistent level of protection regardless of platform. Furthermore, context-aware content filtering could help differentiate between harmful content and contextually safe content, such as educational videos or discussions that are flagged by the system due to certain keywords or phrases.

Lastly, the ethical and emotional aspects of content moderation could be further improved by focusing on collaborations with child psychologists and child safety experts. This would ensure that the system not only focuses on safety but also on positive digital well-being for children. By making these improvements, the AI system could become an even more effective and holistic tool for ensuring children's safety in the digital world.

Future Work

Several improvements in future will make the AI child protection system more efficient, accurate, and friendly to users. One such area of improvement is the refining of content filtering algorithms to avoid false positives and have a better understanding of contextual meaning, distinguishing harmful content from non-harmful content. It will be possible by enriching training datasets and adding more sophisticated natural language processing techniques. This can further help strengthen fraud detection mechanisms with behavioral biometrics and real-time pattern recognition, aiding the system in adapting to changing fraud tactics. Voice recognition accuracy should improve further with broader and more diverse voice datasets and multimodal recognition techniques that combine voice with facial analysis. Regarding age verification, efforts should be put on the reduction of friction by introducing more fluid and less obtrusive methods, while the verification system should be adaptive to the individual needs of users. The user feedback loops, particularly from children, parents, and educators, will play a very important role in fine-tuning the interface and the sensitivity settings of the system. Another aspect is that cross-platform compatibility and regional adaptability will increase the scalability of the system in diverse environments. Ethical considerations, particularly around data privacy and decision-making transparency, must be addressed by implementing robust privacy protocols and ethical auditing processes. Lastly, continuous model updates through real-time learning can ensure that the system stays relevant and effective in mitigating emerging threats to children's digital safety.

References

1. K. Yousaf and T. Nawaz, "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos," in IEEE Access, vol.10, pp. 16283-16298, 2022, doi: 10.1109/ ACCESS.2022.3147519.
2. Y. Gheraibia, S. Kabir, K. Aslansefat, I. Sorokos and Y. Papadopoulos, "Safety + AI: A Novel Approach to Update Safety Models Using Artificial Intelligence," in IEEE Access, vol. 7, pp. 135855-135869,2019,doi: 10.1109/ACCESS.2019.2941566
3. Patil, U. A., &Patil, R. D. (2020). Developing Intelligent Child Protection and Security Systems. International Journal for Research in Engineering Application & Management (IJREAM), 6(1), 63–68. doi:10.35291/24549150.2020.0258.
4. Ferrara, P., et al. (2023). Online "Sharenting": The Dangers of Posting Sensitive Information About Children on Social Media. The Journal of Pediatrics, 257, 113322.
5. K. D. S, S. R, S. S and S. R, "Effective Child Safety System Based on RFID and Face Recognition Technique," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-8, doi: 10.1109/ ICSES55317.2022.9914221.
6. Badillo-Urquiola, K. (2022). A Social Ecological Approach Towards Empowering Foster Youth to be Safer Online (Master's thesis, University of Central Florida). STARS Electronic Theses and Dissertations.

7. K. Yousaf and T. Nawaz, "A DeepLearning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos," in IEEE Access, vol.10, pp. 16283-16298, 2022, doi: 10.1109/ACCESS.2022.3147519.
8. Yu, Y., Sharma, T., Hu, M., Wang, J.,& Wang, Y. (2024). Exploring Parent Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications. arXiv preprint arXiv:2406.10461.
9. Fitwi, A., Yuan, M., Nikouei, S. Y., & Chen, Y. (2023). Minor Privacy Protection by Real-time Children Identification and Face Scrambling at the Edge. Proceedings of the [Conference or Journal Name]. Dept. of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902, USA.
10. Comparative Analysis Of Various Machine Learning Models For Child Safety And Security System For Protecting Them From Child Trafficking And Assault. (2022). Journal of Pharmaceutical Negative Results, 1280-1287.
11. S. Bhabad, K. Bhalerao, P. Nagare, D. Shinde, and P. V. Pandit, "Real-time Object Detection Using ML (Image Processing)," The Journal of Computational Science and Engineering, vol. 2, no. 2, pp. 35, Apr. 2024, ISSN: 2583-9055
12. O. D. Ithape, S. N. Bairagi, J. Musale, M. M. R. Kokani, and P. T. Paranjape, "Image Caption Generation Using Transformer Method," The Journal of Computational Science and Engineering, vol. 2, no. 4, pp. 76, Jun. 2024, ISSN: 2583-9055.
13. S. Kumar, A. P. Singh, and V. Pandey, "Language Detection Using NLP," The Journal of Computational Science and Engineering, vol. 2, no. 4, pp. 120, Jun. 2024, ISSN: 2583-9055.
14. U. Muniraju, V. Rani Chandramule, M. S, P. S, N. S. T, and P. V. Raj, "A Comprehensive Survey on Online Security Against Threats," The Journal of Computational Science and Engineering, vol. 2, no. 4, pp. 76, Jun. 2024, ISSN: 2583-9055
15. P. Dannana and A.S. Venkata Praneel, "A Comparative Study of Machine Learning and Deep Learning Techniques for Facial Emotion," The Journal of Computational Science and Engineering, vol. 2, no. 7, pp. 28, Sep. 2024, ISSN: 2583-9055