PREVENTING CYBER RISKS FOR CHILDREN USING MACHINE LEARNING (E.C.B)

[1]BATOOL ALBADAN, [2]DR. SYED MUJAHID

[1,2]Department of Systems Engineering King Fahd University of Petroleum and Minerals
E-mail: batoulalbadin@gmail.com

Abstract - With the growing use of mobile devices among children, ensuring safe access to age-appropriate applications has become essential. Traditional parental controls and age-gating mechanisms are often either too restrictive or easily bypassed. This study explores an innovative approach to child protection in digital spaces by implementing image-based age detection, using facial recognition and classification to estimate a child's age when they attempt to access certain applications. One of the primary challenges in this approach lies in balancing the misclassification costs associated with age estimation errors, which can result in underestimating or overestimating a child's age. Each misclassification has distinct consequences: underestimation could limit children's access to suitable content, while overestimation could expose them to inappropriate material. To address these challenges, this research utilizes a dual-model framework, combining a global model pre-trained on extensive datasets with a localized model on the user's device, tailored for the specific user environment. The global model benefits from broader generalization, while the local model adapts to individual conditions, allowing for enhanced accuracy in age estimation. Privacy-preserving techniques, such as data encryption and anonymization, ensure that user data is protected, with parents given the option to control data sharing preferences. This study's findings suggest that the methodology effectively balances age classification accuracy, privacy, and parental control, offering a scalable solution to protect children in digital environments.

Keywords - Image Classification, Misclassification Cost, Machine Learning Models, Cost-Sensitive Framework, Privacy Protection, Misclassification Errors, Privacy-Preserving Techniques, Cost-Sensitive Learning, Digital Content Access, Child Protection Systems, Deep Learning

I. INTRODUCTION

In today's digital landscape, children are increasingly exposed to various mobile applications, games, and online content. While this access can be educational and entertaining, it also poses significant risks, as not all content is suitable for every age group. Traditionally, methods such as parental controls or age-gating have been used to restrict access, but these approaches are often either too rigid or easily circumvented. Recent advances in artificial intelligence and computer vision offer an innovative alternative: using real-time image recognition to estimate a child's age from their appearance at the moment they open an application, and using that estimate to decide what content or features are appropriate for them [1].

Classification, a fundamental task in machine learning, assigns data points to predefined categories based on input features; in image-based age estimation, models are trained to recognize patterns and predict characteristics such as age and gender. Age estimation from images is inherently imprecise because of individual differences in appearance, environmental conditions, and model limitations, so misclassification is a significant risk. If a child is estimated as older than they are, they may gain access to unsuitable or harmful content (a false positive); if a child is misclassified as younger, they may be unfairly restricted from appropriate content (a false negative) [2]. Because the consequences of these two error types are not symmetric — false positives carry the more serious safety risk — balancing their costs is central to designing an effective and usable age-classification system [3].

Privacy preservation is equally important, since the system handles sensitive images of children. Core safeguards include data minimization (collecting and retaining only what is necessary), encryption of image data in transit and at rest, edge computing (processing images locally on-device to avoid cloud exposure), anonymization, and obtaining parental consent. The system must also comply with relevant regulations such as the General Data Protection Regulation (GDPR) and the Children's Online Privacy Protection Act (COPPA) [4]. This research explores the trade-offs between misclassification types, aiming to protect children from unsuitable content without unnecessarily restricting their access to age-appropriate applications, by investigating classification models, analyzing their error rates, and applying cost-sensitive and privacy-preserving techniques [3].

A. Research Problem

UNICEF's "The State of the World's Children 2017: Children in a Digital World" documents the risks children face online, including exposure to explicit content, cyberbullying, and misinformation, along with the psychological and behavioral impacts of such exposure [10]. In the UK in 2024, mobile-phone ownership was reported at roughly 98%, 25%, 59%, 24%, and 25% for the 16–17, 12–15, 8–11, 5–7, and 3–4 age groups respectively, increasing the risk that

children access age-inappropriate apps; this underscores the need for tools that can monitor children's phone use even when parents are not present [11].

B. Problem Statement

Parents face significant challenges in monitoring their children's app usage given the vast number of applications available on mobile devices. Traditional parental-control tools often lack the flexibility and granularity needed to adequately protect children from inappropriate content. This project aims to empower parents by letting them select which apps should be monitored, enabling access control based on age-appropriate criteria.

II. LITERATURE REVIEW

Several lines of work address the cost of misclassification in image-based age detection for child protection. Research on dynamic threshold adjustment argues that, from a child-protection standpoint, false positives are riskier than false negatives because they expose children to unsuitable content, and advocates prioritizing recall over precision through thresholds that respond to real-time contextual factors [5].

Work on privacy preservation in child age-detection models highlights techniques such as federated learning and anonymization protocols that allow a system to classify images without retaining or transmitting sensitive data, aligning with ethical standards for protecting minors' data [6].

Studies of cost-sensitive learning show that assigning higher costs to specific, more harmful misclassifications — for example, falsely classifying a younger child as older — and using class-weight adjustments improves outcomes in high-risk categories, enhancing child safety [7].

Research on misclassification information for fine-grained image classification investigates the use of "confusion information" among visually similar classes, employing specialized classifiers to distinguish closely related age groups and reduce errors in borderline cases [8].

A survey of age estimation from facial images reviews traditional approaches — such as the Active Appearance Model and Support Vector Machines — alongside modern deep-learning methods using Convolutional Neural Networks (CNNs). CNN-based models generally achieve superior performance by capturing subtle age-related features, but still face unique challenges in estimating children's ages because children's faces change rapidly over short periods [9].

III. OBJECTIVES

A. Research Questions

- How accurate are current image-classification models in determining a child's age from facial images?
- What are the primary factors contributing to misclassification errors in these models?
- How can the system balance the risk of false positives (misclassifying younger children as older) and false negatives (misclassifying older children as younger) while optimizing for child safety and user experience?

B. Hypotheses

Main hypothesis: E.C.B reduces classification errors on websites for the 3–16 year age group with high efficiency.

Null hypothesis: E.C.B does not affect classification errors on websites for the 3–16 year age group.

- False positives (incorrectly identifying a child as older) have more severe consequences than false negatives, making it essential to focus on minimizing them.
- Cost-sensitive learning approaches can significantly improve the balance between false positives and false negatives by prioritizing more harmful misclassifications.
- Incorporating privacy-preserving techniques such as edge computing and encryption will enhance the trustworthiness of the age-detection system without compromising classification performance.

C. Purpose

The overarching purpose is to ensure children are protected from inappropriate digital content by developing a robust age-classification system that accurately determines a child's age at the moment of application access, while balancing safety against minimizing unnecessary restrictions on age-appropriate content. The study's variables fall broadly into independent and dependent categories, with mediating or control variables accounting for additional influencing factors.

D. Novelty

- An app that lets parents designate which other apps on the phone are monitored, removing the need for a separate "kids mode."
- A global model learned via deep learning with an encoder; computationally expensive to train, but lightweight to run once deployed to phones.
- A local model that combines the trained deep-learning model with personalized shallow learning on-device, with an option for the user not to share data with the global model.

IV. METHODOLOGY

A. Variables

- Independent variable: number of images.
- Dependent variable: classification accuracy.
- Fixed variables: programming language (Python), deep-learning framework, global model, local model, personal learning.

B. Materials

- Image dataset: high-quality, annotated facial images of children spanning a wide range of ages (e.g., UTKFace or IMDB-WIKI), or a proprietary dataset meeting age-specific permission and privacy requirements.
- Ethics and consent forms: given the use of images of minors, ethical clearance, parental consent, and data-protection protocols (e.g., GDPR compliance) are essential.
- Annotated labels: each image carries a validated age label or age range to support accurate model training.

C. Tools

- Deep-learning frameworks: Tensor Flow or PyTorch for building and training the neural network models.
- Pre-trained models: ResNet, VGG, or Efficient Net as a baseline, fine-tuned for age-specific classification.
- Cost-sensitive learning algorithms: assign different weights to misclassifications to minimize the cost of false positives (inappropriate access) versus false negatives (unnecessary restriction).
- Privacy-preserving techniques: federated learning or differential-privacy tools (e.g., Tensor Flow Privacy) to build models without transferring sensitive data.

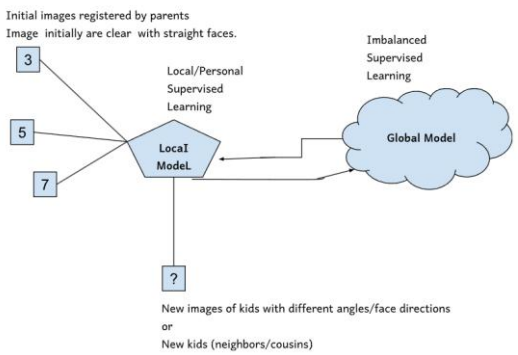


Fig. 1. Local/global model architecture: a local model trained on parent-registered images interacts with an imbalanced, supervised global model to classify new images.

D. Application Setup

Upon installation, parents choose which applications on their child's device should be monitored for age appropriateness, and select whether to enable data sharing with the global model or keep data local to enhance privacy.

E. Programming Pipeline

- Import required libraries for data manipulation, image processing, model training, privacy-preserving noise addition, and evaluation.
- Set parameters for decentralized learning, differential privacy, and image processing.
- Load images and labels from a folder structure (one subfolder per class); resize, normalize, and one-hot encode labels.
- Exploratory data analysis: class distribution, original image dimensions, sample images per class, and pixel-intensity distribution.
- Define a CNN architecture sized to the input shape and number of classes.
- Add differential privacy to gradients via Laplace noise during training.
- Run decentralized training across multiple agents, each updating models using differentially private gradients shared with neighbors.
- Split data into training and testing sets and train the decentralized models, returning trained models per agent.

F. Model Training

The application uses a pre-trained global model developed with imbalanced supervised learning on a diverse image dataset, trained to recognize age groups and generalize across facial features. Global Model-A resides on the server and is periodically updated with new data; Global Model-B is a copy residing on the user's device, obtained at install time. A local shallow-learning model is also initialized on-device at install time, working alongside Global Model-B; it is refined from photos that parents upload or capture of their children, stored in the local model's on-device database.

G. Image Capture and Decision Making

When a child attempts to access a restricted app, the local model captures an image of the child's face. The image is processed locally first, using an ensemble of Global Model-B and the Local Model, to maintain privacy; if the parent has opted in to data sharing, federated learning with Global Model-A is also performed. Based on the resulting age estimate, the application determines whether the child may access the requested app, denying access if the child is deemed too young.

H. Privacy Preservation

All images sent to the global model are encrypted in transit, keeping sensitive information confidential. Parents retain control over data sharing and can choose to keep their child's data local rather than sharing it with the global model.

I. Continuous Learning

Parental feedback on access decisions is used to refine the local model over time, while the global model can be updated with aggregated, anonymized data to improve performance without compromising user privacy.

V. PRIMARY RESULTS

A. Improved Age-Classification Accuracy

Combining global and local models improves the precision of age estimates, particularly around critical age thresholds where misclassification risk is highest, by letting the local model interpret environmental factors while the global model contributes broader generalization. Parental feedback further refines the local model, progressively improving its accuracy for individual usage patterns.

B. Strong Privacy Protection

Encryption and user-controlled data sharing keep sensitive information protected throughout the process; data is encrypted and anonymized before any interaction with the global model. Letting parents choose if and when to share data upholds privacy preferences and supports user trust in data handling.

C. Effective Parental Control Options

Parents can tailor monitoring to their child's needs by specifying which apps require monitoring, offering a more precise approach than traditional parental controls. Responsiveness to parental feedback allows the app to remain aligned with parental preferences and to grow more accurate over time.

D. Balanced Misclassification Costs

By optimizing for recall, the methodology prioritizes safety in classification, reducing cases where children access inappropriate content while minimizing both over-restriction and inappropriate access — balancing protection with user experience.

Future work: increase the number of training images to improve classification accuracy, and further develop the underlying cost equation.

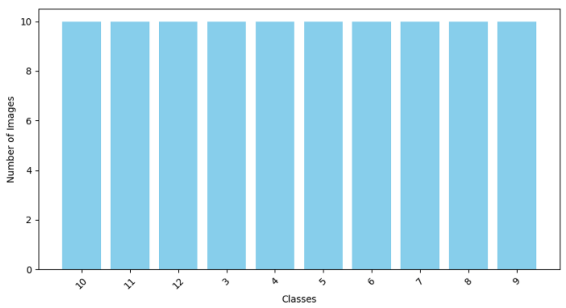


Fig. 2. Class distribution of the training dataset used in the program experience, showing an equal number of images per age class.

VI. INTERPRETATION & CONCLUSIONS

A. Enhanced Accuracy in Age Classification

The dual-model structure, combining a global model trained on extensive, demographically broad data with a local model adapted to specific user patterns, helps mitigate errors around critical age boundaries, keeping content restriction effective and age-appropriate.

B. Privacy and Security Protocols

Embedding privacy-preserving techniques such as federated learning and data encryption lets the local model estimate age without sending identifiable images to a central server, aligning with regulatory and ethical standards for child data protection and giving parents more control over their child's digital experience.

C. Parental Customization and Flexibility

Allowing parents to specify which apps are monitored strengthens the system's practicality, providing tailored, adaptable monitoring that reflects individual child needs while minimizing unnecessarily restrictive measures.

D. Cost-Sensitive Learning to Balance Misclassification Risks

Cost-sensitive learning lets the system better handle the unequal costs of misclassification, dynamically adjusting thresholds based on error feedback to reduce both false positives (inappropriate access) and false negatives (unnecessary restriction), contributing to an optimal balance between safety and accessibility.

E. Real-Time Adaptability and Self-Learning

The local model's ability to update itself from real-time usage data refines age estimation over time, improving accuracy and responsiveness as children's facial features change, and keeping the system relevant over the long term.

F. Conclusion

The study demonstrates that combining enhanced classification accuracy, robust privacy protection, parental customization, and cost-sensitive learning offers a comprehensive solution to age-appropriate content restriction. The approach safeguards children from inappropriate content while giving parents adaptable tools to manage their child's digital experience. With further refinement — particularly in model sensitivity and error management — this methodology holds potential as a standard approach for age-restricted content across mobile applications.

REFERENCES

[1] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2015, pp. 34–42.

[2] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5810–5818.
[3] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.
[4] J. Lin, L. Cheng, and Y. Xu, "Webly supervised fine-grained image recognition using graph representation and metric learning frameworks," Electronics, vol. 11, no. 24, p. 3942, 2022.
[5] J. Xu et al., "Dynamic threshold adjustments for cost-effective classification in sensitive contexts," 2023.
[6] Y. Chen et al., "Privacy preservation in child age detection models," 2022.
[7] Train in Data, "Cost-sensitive learning: Beyond the accuracy in imbalanced classification," 2022.
[8] D.-H. Wang et al., "Exploring misclassification information for fine-grained image classification," 2021.
[9] H. Sun and Q. Li, "Age estimation from facial images: A survey," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1279–1297, 2020.
[10] UNICEF, "The state of the world's children 2017: Children in a digital world," 2017.
[11] F. Laricchia, "UK: Children owning mobile phones by age," 2021, accessed Aug. 24, 2023.

★★★