What is the impact of makeup on automated face recognition and age estimation models across everyday, syntethic and extreme conditions?

Frank van der Velde
Faculty Tech
Hogeschool van Amsterdam
Amsterdam, Nederland
frankvdv1998@gmail.com

***Abstract*— Self-checkout systems in supermarkets face challenges with age verification for alcohol sales. Automated face recognition and age estimation models offer a potential solution but may be unreliable when facial appearance is altered by makeup. This study examines the impact of everyday, synthetic, and extreme makeup styles on model performance. InsightFace is used for recognition and DEX (Deep EXpectation of Apparent Age) for age estimation, selected for their widespread use, reproducibility, and benchmark relevance. Three datasets are employed: the Kaggle Makeup Detection Dataset with real paired before-and-after images, the FFHQ-Makeup Dataset with GAN-generated synthetic makeup, and a curated set of drag and horror makeup representing extreme conditions. Performance is evaluated through recognition certainty and mean absolute error (MAE), using ground-truth labels where available and prediction disparities otherwise. The research addresses gaps in understanding how makeup particularly extreme styles affects automated systems, providing insights into the vulnerabilities of AI-based age verification in supermarket compliance.**

Keywords*— *Age Estimation, Facial Recognition, Makeup Bias, Computer Vision, Alcohol Compliance, Apparent Age

I. Introduction

Self-checkout registers are becoming increasingly common in Dutch supermarkets, with Albert Heijn leading in their adoption to improve efficiency and reduce waiting times [1]. For customers this offers convenience, while for employees it reduces the need to staff traditional checkout counters. However, the sale of age-restricted products such as alcohol introduces challenges. Dutch law requires strict age verification, and government inspections show that compliance is still problematic. Between July 2023 and June 2024, the Dutch Food and Consumer Product Safety Authority (NVWA) reported widespread failures in age checks across retail, with many stores making mistakes when selling alcohol [2]. The fines for violations are high, making compliance a priority for supermarkets. Albert Heijn publicly states its commitment to responsible alcohol sales, for example through its *NIX18* policy [3].Yet research shows that in practice, checks often fail or are incomplete, both in stores and in online deliveries [4].

Because manual verification at self-checkouts is time-consuming and error-prone, automated solutions are worth considering. Facial recognition combined with age estimation algorithms could theoretically streamline alcohol sales by reducing reliance on human judgment. However, such systems are not yet widely used in Dutch retail, in part due to strict privacy laws and ethical concerns about biometric surveillance [1]. Nonetheless, the technology is under active development worldwide, and it is realistic to expect supermarkets to consider its implementation in the future. This raises important questions about fairness, accuracy, and potential vulnerabilities of automated age estimation.

One overlooked factor is the impact of makeup. Makeup can significantly alter the apparent age of individuals by concealing wrinkles, enhancing facial contrast, or otherwise changing visual cues used for age estimation. Research shows that cosmetics can not only shift perceived age but also reduce the reliability of computer vision algorithms [5]. While this has been documented in controlled biometric settings, its implications for supermarket age verification have not yet been studied. From an ethical perspective, it is important to ensure that people wearing makeup are treated fairly and not systematically misclassified. From a security perspective, it is equally important to investigate whether makeup could be used deliberately to mislead AI models, for example by making underage individuals appear older.

This study therefore examines the research question: *"What is the impact of makeup on automated face recognition and age estimation models across everyday, syntethic and extreme conditions?"* To answer this, the project combines a literature review with experiments

testing an age estimation model and a face recognition model on images of the same individuals with and without makeup. The goal is to determine whether makeup increases prediction errors or introduces systematic bias, and to assess the potential consequences for the reliability and fairness of future AI-based age verification systems in the Dutch retail context.

II. Literature Review

A. Impact of Makeup on Facial Recognition and Age Estimation

The influence of makeup on facial recognition and age estimation has been a subject of interest in computer vision research. Studies have demonstrated that makeup can significantly alter the perception of facial features, affecting the accuracy of automated systems like has been demonstrated in the case of age-induced makeup attacks [6]. For instance, research indicates that facial cosmetics can confound automated gender and age estimation algorithms by modifying facial features such as skin texture, wrinkle concealment, and facial contouring [5]. Additionally, makeup can lead to overestimation of age in younger individuals and underestimation in older individuals, as it enhances features associated with youthfulness while masking signs of aging [7].

B. Gaps in Existing Literature

While existing studies provide valuable insights into the effects of makeup on facial recognition, several gaps remain. Notably, there is limited research on the impact of extreme makeup styles, such as drag makeup, grime and horror makeup on automated age estimation systems. These makeup styles often involve exaggerated features, completely changing and hiding features and stylized applications that may significantly differ from conventional cosmetic practices. Existing literature does not comprehensively address how such extreme makeup styles affect the performance of age estimation models.

Furthermore, the specific context of alcohol sales in supermarkets has not been extensively studied in relation to makeup's impact on age estimation. While general studies on makeup and facial recognition exist, there is a lack of research focusing on the implications for compliance in age-restricted product sales, particularly in automated retail environments.

C. Need for Current Research

This research is essential to bridge the existing gaps and address the evolving challenges in automated age verification systems. As supermarkets increasingly adopt self-checkout systems, ensuring accurate and fair age estimation becomes critical, especially for age-restricted products like alcohol. Understanding how makeup, including extreme styles like drag makeup, influences automated age estimation will inform the development of more robust and equitable systems. Moreover, this study will contribute to the broader discourse on demographic biases in AI systems and their implications for privacy and fairness in automated retail settings. While also giving an insight in the inclusivity in the case of people with heavy make up styles like drag makeup.

III. Research Methodology

A. Aim and Research Design

The aim of this study is to investigate how different types of makeup ranging from everyday cosmetics to extreme applications such as drag and theatrical horror makeup affect the reliability of automated face recognition and age estimation models in supermarket self-checkout contexts. The study addresses a knowledge gap by extending research beyond everyday cosmetics into extreme styles, and an empirical gap by applying the findings to the practical problem of alcohol sales compliance.

The research design combines two approaches: a literature review of existing work on makeup-induced biases in computer vision models, and an experimental analysis using publicly available and self-compiled datasets. Particular attention will be paid to whether makeup systematically reduces recognition confidence or increases age estimation error, and whether extreme styles such as drag and horror makeup can either deliberately or unintentionally mislead models.

Two pre-trained models will be employed. For face recognition, the study uses InsightFace [8], a state-of-the-art framework widely adopted in both academic and applied biometric systems. InsightFace is chosen because of its proven robustness, industry relevance, and strong benchmark performance, making the findings transferable to real-world contexts such as supermarket self-checkouts. For age estimation, the study adopts DEX (Deep EXpectation of Apparent Age) [9], introduced by Rothe et al. and trained on large-scale datasets such as IMDB-WIKI and APPA-REAL. DEX it is a proven and widely cited and reproduced age estimation model, offering strong comparability with prior research. Its ease of implementation, availability of pre-trained weights, and well-documented training datasets make it suitable for reproducible evaluation, despite its older VGG-16 backbone. This choice balances methodological rigor with practical feasibility.

B. Data collection and Datasets

The experimental analysis will draw from three complementary sources of data, reflecting the challenges of makeup-induced bias in both everyday and extreme conditions.

The first is the Kaggle Makeup Detection Dataset [10], a small paired dataset containing approximately 25 individuals photographed before and after applying everyday makeup. Although limited in size, this dataset provides rare real-world paired examples that enable direct person-level comparisons under natural conditions.

The second is the FFHQ-Makeup Dataset [11], based on the Flickr-Faces-HQ collection. In this dataset, makeup is synthetically applied using generative adversarial networks (GANs). Synthetic makeup in this context refers to algorithmically generated cosmetic effects digitally overlaid on real base faces. This process ensures that the only systematic difference between paired images is the presence of makeup, thereby allowing controlled large-scale testing across thousands of identities and styles. While synthetic, this dataset compensates for the scarcity of large real-world paired makeup data and provides a statistically robust basis for experimentation. It is especially useful because it has makeup examples on groups that are harder to get examples of makeup of such as children and men.

Finally, additional curated data will be collected from publicly available and fair-use sources, specifically before-and-after images of drag makeup and horror effects makeup from media such as *RuPaul's Drag Race* and *The Boulet Brothers' Dragula*. These examples represent extreme makeup styles, where facial features are deliberately resculpted, concealed, or exaggerated, such as eyebrow blocking, prosthetics, or contouring to alter perceived facial structure. Because no standardized dataset of such transformations exists, this set must be manually compiled. Because the data is part of the free domain and matching the high quality images with ages it provides unique insight into whether models remain reliable under adversarial or transformative conditions.

All datasets will be pre-processed through normalization and resizing to meet model input requirements. Where applicable, care will be taken to balance conditions between makeup and non-makeup samples for fair evaluation. The data sets do not inherently have the labels for things such as race and gender which could provide an problem in insuring balance.

C. Research Approach and Analysis

The methodology consists of three phases.

First, a literature review will establish the theoretical foundation by analyzing prior studies on cosmetics and bias in computer vision. Works such as Chen et al. on the impact of facial cosmetics on automated age and gender estimation ***[5]*** and Clapes et al. on demographic bias in apparent age estimation ***[12]*** provide essential background.

Second, experimental testing will evaluate both recognition and age estimation models. For recognition, InsightFace will be applied to paired before-and-after makeup images to measure whether recognition certainty (confidence scores) decreases under cosmetic conditions. For age estimation, DEX will be used to predict apparent ages across the datasets. Performance will be evaluated using Mean Absolute Error (MAE). In cases where ground-truth ages could be available such as with the drag makeup data or subsets of Kaggle Makeup Detection, MAE will be calculated against true labels. In cases without age labels, such as FFHQ-Makeup and curated drag/horror sets, the disparity between paired before-and-after predictions will be measured as a proxy for error. Both macro-averaging (per-person averaging before aggregation) and micro-averaging (per-image averaging) will be reported, with emphasis on macro-averaging to ensure fairness at the individual level.

Third, supplementary tests will be conducted on curated extreme makeup examples. These analyses will explore whether such applications significantly distort recognition certainty and age predictions, and whether they can be interpreted as adversarial conditions capable of bypassing automated age verification systems.

By comparing baseline performance on natural faces with results under everyday, synthetic, and extreme makeup conditions, the study will generate both quantitative findings (error rates, recognition confidence, prediction disparities) and qualitative insights (potential vulnerabilities, fairness concerns). The results will be interpreted in relation to the ethical and legal challenges of deploying AI-based age verification in supermarkets, where both underestimation and overestimation of age can have serious compliance consequences.

IV. PLANNING

TABLE I. PLANING FOR RESEARCH PROPOSAL

Week	***Date***	***Tasks***
2	8 sept – 14 sept	- Choosing research topic - Formulating research questions - Write introduction - Choose research method - Find useful literature and data
3	15 sept – 21 sept	- Start of literature research - Finish research proposal ***19 september: Submit research proposal***
4	22 sept – 28 sept	- Prepare data sets for testing research question (various types of makeup)
5	29 sept – 5 okt	- Setup face detection model - Setup age estimation model
6	6 okt – 12 okt	- Start experiments to gather results
7	13 okt – 19 okt	- Evaluate results
8	20 okt – 26 okt	- Analyse results and add conclusion to research
9	27 okt – 2 nov	- Gather feedback on research and finalize
10	3 nov – 9 nov	- Add relevant parts of research paper to group rapport

Fig. 1. Planning with which tasks I would like to do each week to finish the research before the deadline.

V. Conclusie / Hypothesis

This study hypothesizes that everyday makeup will cause modest deviations in automated age estimation, typically shifting predictions by a few years, while having limited impact on recognition certainty. In contrast, extreme styles such as drag and theatrical horror makeup are expected to produce substantial prediction disparities and reductions in recognition confidence, as these applications deliberately resculpt or conceal key facial features. While InsightFace is anticipated to maintain reliable identification under most everyday cosmetic conditions, heavily transformative makeup may introduce uncertainty and misclassification in both recognition and age estimation. By testing across real, synthetic, and curated datasets, the study further assumes that systematic disparities will emerge between before-and-after conditions, even in the absence of ground-truth ages. These findings will provide insight into whether makeup poses a practical vulnerability for supermarket compliance. In the longer term, the study suggests that integrating a complementary makeup detection mechanism could serve as a safeguard by triggering manual checks under conditions of high deviation, thereby improving fairness and reducing compliance risks in self-checkout systems.

VI. References

[1] Avrotros, ""Mag dit zomaar?" Vragen over camera's bij de zelfscankassa," 23 05 2025. [Online]. Available: https://radar.avrotros.nl/artikel/mag-dit-zomaar-vragen-over-cameras-bij-de-zelfscankassa-61602.

[2] Nederlandse Voedsel- en Warenautoriteit, "Inspectieresultaten Alcoholwet juli 2023 - juni 2024," 2024. [Online]. Available: https://www.nvwa.nl/onderwerpen/alcoholverkoop/inspectieresultaten/2024/alcoholwet-juli-2023-juni-2024.

[3] Albert Heijn, "Duurzaamheidsverslag," 2022. [Online]. Available: https://duurzaamheidsverslag.ah.nl/2022/nix18.

[4] Nederlandse Voedsel- en Warenautoriteit (NVWA), "Factsheet Online alcohol bestellen door jongeren 2024," Breuer & Intraval / NVWA, 2025.

[5] C. a. D. A. a. R. A. Chen, "Impact of facial cosmetics on automatic gender and age estimation algorithms," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, Lisbon, Portugal, 2014.

[6] K. a. M. Z. a. M. S. Kotwal, "Detection of Age-Induced Makeup Attacks on Face Recognition Systems Using Multi-Layer Deep Features," IEEE, 2019.

[7] E. B. Gavas, C. Hegde, N. Memon and S. Banerjee, "DiffClean: Diffusion-based Makeup Removal for Accurate Age Estimation," arXiv, 2025.

[8] InsightFace, "InsightFace: An open source 2D & 3D deep face analysis library," [Online]. Available: https://insightface.ai/. [Accessed 19 9 2025].

[9] siriusdemon, "pytorch-DEX: Pytorch implementation of DEX: Deep EXpectation of apparent age from a single image," GitHub, [Online]. Available: https://github.com/siriusdemon/pytorch-DEX. [Accessed 19 9 2025].

[10] tapakah68, "Makeup Detection Face Dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/tapakah68/makeup-detection-dataset. [Accessed 19 9 2025].

[11] X. Yang, S. Ueda, Y. Huang and T. Akiyama, "FFHQ-Makeup: Paired Synthetic Makeup Dataset with Facial Consistency Across Multiple Styles," CyberAgent AI Lab; Keio University, 2022.

[12] A. Clapes, O. Bilici, D. Temirova, E. Avots, G. Anbarjafari and S. Escalera, "From Apparent to Real Age: Gender, Age, Ethnic, Makeup, and Expression Bias Analysis in Real Age Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Salt Lake City, UT, USA (CVPR 2018 location), 2018.

[13] BNNVARA, "Kwart consumenten zelfscan vergeet producten af te rekenen," 17 02 2024. [Online]. Available: https://www.bnnvara.nl/kassa/artikelen/kwart-consumenten-zelfscan-vergeet-producten-af-te-rekenen.

[14] Q&A Retail, "1-op-5 consumenten vindt diefstal in supermarkt onder omstandigheden toelaatbaar," 4 2024. [Online]. Available: https://www.qanda.nl/publications/1-op-5-consumenten-vindt-diefstal-in-supermarkt-onder-omstandigheden-toelaatbaar#:~:text=Zelfscankassa%20heeft%20bij%20meerderheid%20de,voorkeur%20boven%20de%20gewone%20kassa..