Enhancing Efficiency in Self-Checkout: An AI-Based Age Estimation Model for Alcohol Purchase Verification in Supermarkets

MAAI Project Vision, Group 9

Alex Cheng
Master Applied Artificial Intelligence
Hogeschool van Amsterdam
Amsterdam, The Netherlands
alex.cheng@hva.nl

Bart Combee
Master Applied Artificial Intelligence
Hogeschool van Amsterdam
Amsterdam, The Netherlands
bart.combee@hva.nl

Jim Mekkelholt
Master Applied Artificial Intelligence
Hogeschool van Amsterdam
Amsterdam, The Netherlands
jim.mekkelholt@hva.nl

Rick Steenhorst
Master Applied Artificial Intelligence
Hogeschool van Amsterdam
Amsterdam, The Netherlands
rick.steenhorst@hva.nl

Tijn W. Kahmann
Master Applied Artificial Intelligence
Hogeschool van Amsterdam
Amsterdam, The Netherlands
tijn.kahmann@hva.nl

***Abstract*—This research presents the development of an age estimation model specifically designed for self-checkout systems in supermarkets, aimed at improving the process of age verification for alcohol purchases. The model predicts customer age to streamline the often inefficient self-checkout alcohol sales process. The objective is to improve efficiency by reducing customer delays while ensuring age verification compliance. The study follows a structured methodology, resulting in a fully functional application that integrates the age estimation model into the self-checkout process. Findings show a reduced need for manual intervention, enhancing retail operational efficiency.**

***Index Terms*—Machine learning, Face detection, Age estimation, Process improvement, Retail AI application.**

I. INTRODUCTION

A. Context

In the Netherlands, the legal age for purchasing alcohol is 18 years [1]. The NIX18 campaign [2] actively aims to prevent underage drinking. In collaboration with CBL, the food trade agency, it supports the objective of achieving a 100 percent compliance rate in alcohol sales [3], [4].

Underage drinking is a serious issue. A 2023 Trimbos study found nearly 40% of 12- to 16-year-olds have consumed alcohol [5], with 14% of young boys reporting binge drinking (defined as consuming at least six glasses) in the month preceding the research. Dutch alcohol laws and the objectives set by CBL place significant responsibility on supermarkets, particularly on checkout staff, to ensure compliance by verifying customer ages. Interviews with staff [6] highlighted the difficulties of age verification due to high workloads, social pressures, and occasional confrontational situations. Despite CBL's efforts to enhance compliance, the latest 2022 statistics show that supermarket compliance rates remain at only 63% [7]. This persistent issue of low compliance may be linked to the challenges faced by employees of self-scan registers.

B. Problem

Employees experience time pressure, particularly during busy periods, as the age verification process is often time-consuming and carries the risk of errors, such as selling alcohol to minors. Frustrated customers often complain about delays, adding to employee stress [8]. Considering the points mentioned above, as well as the interviews and empathy map in Appendix A, employees express a desire for a more efficient process with less manual interventions. They seek a standardised system that simplifies the process and reduces waiting times.

C. Proposal

Several approaches were considered to enhance the age verification process. One option is to scan government-issued IDs at the checkout, automatically verifying age without requiring staff involvement. Another possibility is linking age verification to bank cards, by using financial data to confirm eligibility. However, both approaches present limitations. They raise privacy concerns and rely on external government systems and data, creating compliance and technical issues. Therefore, the chosen solution is an age verification system using AI-driven facial detection and machine learning for age estimation. The aim is to improve efficiency by streamlining the operational workflow, assisting supermarket staff while maintaining a human in the loop.

D. Current Work

In contrast to Dutch supermarkets [6], five UK-based supermarkets have tested facial detection technology to verify the

age of alcohol buyers at self-service checkouts [9]. Additionally, the UK government is adjusting laws to accommodate this change [10]. An example AI model used for this purpose was developed by Yoti and showed promising results in terms of both performance and robustness. The Yoti regression model achieved a Mean Absolute Error (MAE) of 2.9 years for individuals aged 6 to 70. Furthermore, the model displayed no discernible bias across gender and skin tones [11]. The topic of bias across different demographies was introduced in the ethics workshops. This work applies the FairLearn framework to address the impact of this topic [12].

E. Gap

As explored in the State of the Art section, face age estimation models have demonstrated promising performance for robust age verification using facial images. The implementation of these systems for monitoring age verification in alcohol sales could be beneficial for multiple stakeholders. However, the lack of exploration of this technology in Dutch supermarkets highlights a clear gap, which this research aims to address.

This paper investigates the development of a prototype system designed to guide customers through an AI-driven checkout process in supermarkets, including age verification. The primary aim of the prototype is to enhance the workflow for employees, our main stakeholders, while ensuring regulatory compliance and offering customers a transparent choice to protect their privacy. The AI model must be capable of detecting faces and making age predictions based on a specified age threshold.

The development of out prototype is based on the CRISP-ML(Q) methodology [13], with adjustments to fit our specific context, utilizing the DeepFace Framework for model development.

In summary, the paper proposes an AI-system for automatic age verification, which contains:

- Age estimation model with confidence scoring.
- Privacy-focused user interface.
- Human-in-the-loop verification.

II. BACKGROUND

A. Basic Knowledge

Self-checkout systems allow customers to scan and pay for their items without assistance from store employees. In the case of purchasing alcohol, regulations in the Netherlands require anyone appearing under the age of 25 to undergo age verification [1]. Currently, this process is carried out manually. When alcohol is scanned at the self-checkout, employees are alerted by the system and must physically verify the customer's age. If the employee estimates that the customer appears under 25, they must request an ID and manually enter the date of birth into the system. This process could be enhanced through an automated age verification system using facial detection. At present, Dutch supermarkets do not have automated age estimation systems in place, meaning that the age verification process remains fully manual and entirely dependent on human intervention, effectively creating a baseline model with 100% manual checks.

B. State of the Art

Recent advances in face recognition technology focus on achieving high accuracy through four key stages: detection, alignment, representation, and verification [14]. Age estimation models primarily build upon face recognition frameworks. State-of-the-art methods leverage Convolutional Neural Networks (CNNs) [15], trained on large public datasets such as LFW [16], IMDB-WIKI [17], and UTK-Face [18], to enhance performance. These approaches yield competitive results in both identity verification and age estimation tasks [19][20]. For instance, the age estimation model in the DeepFace framework achieves a mean absolute error (MAE) of 4.65 years on the IMDB-WIKI dataset [19]. This demonstrates promising outcomes compared to human age estimation abilities based on facial images, as evidenced by one study where human predictions on FG-NET and MORPH II datasets had MAEs of 4.7 and 6.3 years, respectively [21].

C. Stakeholder-analysis

Based on multiple observations and interviews, we have identified three primary stakeholders:

- Customers
- Employees
- Supermarket chains

These stakeholders are visualised in a stakeholder map (Figure 1), ensuring that the needs and influences of all relevant parties are clearly understood and effectively managed [22].

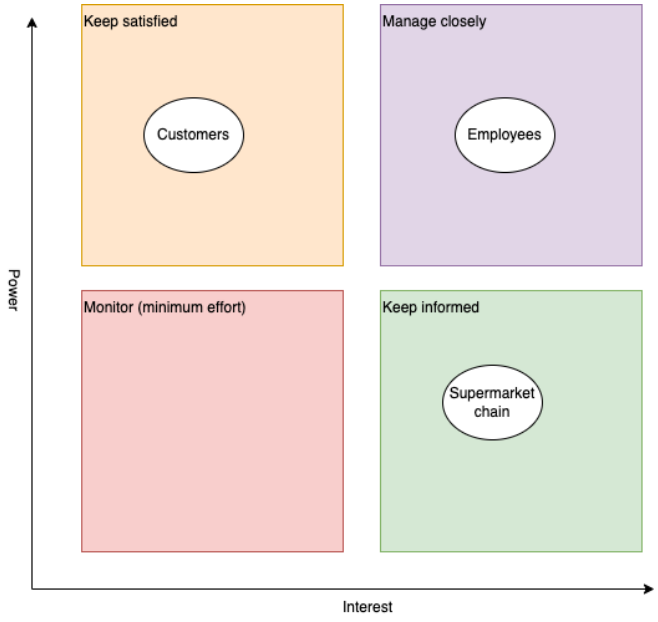

Fig. 1. Stakeholder Map

We focus on one stakeholder in particular, the employees. This is because based on our research, we believe that there is

the most room of improvement that could be made for them in terms of less workload and more efficient monitoring of alcohol purchases.

Furthermore, we have the goal to keep the customers satisfied by introducing a new system without increasing the wait time or extending the process duration compared to the current system.

And lastly, we will keep the supermarket chains informed when this product is ready to be released for the market.

III. REQUIREMENTS

The requirements are structured using the MoSCoW prioritization technique, which organizes requirements into four categories: Must have, Should have, Could have, and Won't have. The requirements are categorised into product requirements and model-specific requirements. This section only contains the must-have requirements as these are the most important and in scope for the research, see Appendix B for the full MoSCoW requirements.

A. Must Have Product Requirements

The following must have product requirements have been identified:

- **Single Register Interface:** The interface must be designed to allow employees to manage a single register without interruptions.
- **Human-in-the-Loop Option:** The system must include a manual verification option.
- **User Data Handling:** The system must inform users about data handling practices, explicitly stating that photos will not be saved.

B. Must Have Model Requirements

The following must have model requirements have been identified:

- **User Interface:** The model must be able to receive an image and return an age prediction through a user interface.
- **Face Detection:** The model must detect faces before performing any prediction. If no face is detected, no age estimation should be done.
- **Age Estimation:** The model must estimate the customer's age achieving a Mean Absolute Error (MAE) of less than 6.0 [23].
- **Response Time:** The model must provide predictions within 10 seconds.

IV. PROTOTYPE

This section details the prototype, covering its value proposition, flow diagrams, AI breakdown, design patterns, the final prototype, levels of automation, and ethical considerations. The prototype was refined through several iterations, incorporating feedback from stakeholders and users, ensuring alignment with the project's objectives and improving overall usability.

A. Value proposition

The value proposition communicates the advantages and benefits of the AI age estimation model for supermarket employees [24]. It explains how the model enhances efficiency at self-checkouts when purchasing alcohol. In [6] we created multiple value proposition and eventually created this final value proposition:

- **Concept name**: Age Identification System
- **Using**: AI facial detection and machine learning to determine customer age.
- **To**: Provide advice to support staff in the age verification process.
- **We can help**: Supermarket employees conducting alcohol age checks.
- **With a better way to**: Support and automate age verification by reducing manual checks and improving the efficiency of the process.
- **Because / so that**: The time required for alcohol age checks is shortened, creating a simpler, uniform process.
- **With / without**: Faster age verification without delays and underaged alcohol sales

This value proposition is based on discussions with supermarket employees, team leaders, and customers [6]. Observations were also conducted at self-checkouts in different supermarkets, where the specific flow of each supermarket is visualised.

B. Flow diagrams

To visualise the information gathered from various sections of this paper regarding the identified problem, we developed an initial draft of the flow diagram, which is presented in Figure 2.

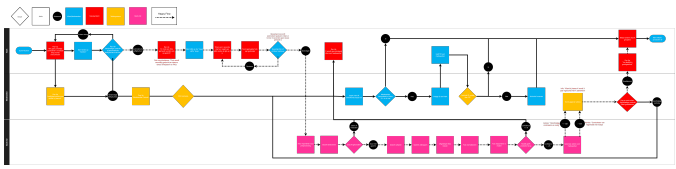

Fig. 2. Draft flow diagram

This draft was refined through several iterations of prototype testing, where we gathered feedback from Communication and Multimedia Design (CMD) students, as documented in the logbook [6]. Additionally, we held critical discussions within our team, with fellow students, and in class settings. These prototype tests and collaborative sessions resulted in the following improvements to the flow:

- Removal of the manual override for employees, streamlining the process and reducing customer confusion.
- Simplified employee interface with fewer options.
- Elimination of customer options to retake or confirm their image.
- Removal of unnecessary customer choices for a smoother process.

The final flow diagram, incorporating these enhancements, is presented in figure 3.

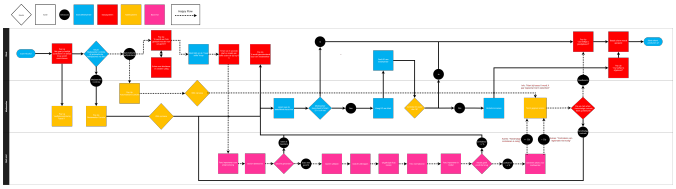

Fig. 3. Draft flow diagram

C. Interface design and design patterns

The interface design was also improved through the same iterative process mentioned in the previous section. For reference, both the old and new interface screens are available in the GitLab repository [25]. The most important changes can be summarised as follows:

- Enhanced button colours and design to guide customers to AI-driven verification.
- Fewer screens for a streamlined process.
- Added privacy and compliance button to explain system operations.
- Improved iconography for non-Dutch speaking users.
- Renamed buttons for greater clarity.

While designing multiple iterations of our prototype, we also took design patterns into account, including those referenced in [26]. We believe that by implementing these design patterns, our system will be more understandable for our stakeholders. In appendix C an overview of the implemented design patterns can be found.

D. Prototype

The flow diagram and design enhancements led to the development of our final prototype system, which aligns with our objectives and incorporates the following key aspects:

- A simple, intuitive interface for employees.
- Customer choice in preferred age verification method.
- Minimal interruption, allowing customers to scan items while the alcohol check proceeds.
- Clear instructions for customers using automated age verification.

The full prototype, will be on Gitlab [25] and shown at the presentation of the paper.

E. AI Breakdown

The Age Identification System uses AI for age estimation and face detection. The system predicts whether a customer is of legal age, but errors like false positives (underage customers classified as adults) or false negatives (adults classified as underage) can occur, as shown in Figure 4. Face detection may also fail if the face is cropped or not fully visible. To address these errors, the system relies on manual checks by employees and user education, ensuring minimal disruption.

In	Task	Out	Possible errors	Impact of errors	% errors yet still usefull	Possible solutions
Photo customer	Model estimates age	Estimated age (binary)	Underage customer is estimated as of legal age (False Positive, FP) Of-age customer is estimated as underage (False Negative, FN) Model cannot read the photo (image format)	- High - Low	- 5% - 15%	Manual check by employee
Photo	Face detection	Cropped face	A fake face No face detected A partial face or a missing part of the face	- Mid - Mid - Mid	- 10% - 50% - 5%	Warn the user that this is not allowed Graceful handoff

Fig. 4. AI Breakdown

F. Levels of automation

The system will be partially automated, with certain processes supported by automation. We estimate level 5 for our application based on the categorization in Figure 5. Levels 6 and 7 present risks, as automatic approval of customers without human intervention could violate Article 22 of the GDPR [27], which protects individuals from decisions solely based on automated processing that have significant effects, such as age-restricted purchases [3][4]. Therefore, level 5, which retains 'human in the loop,' is the most appropriate.

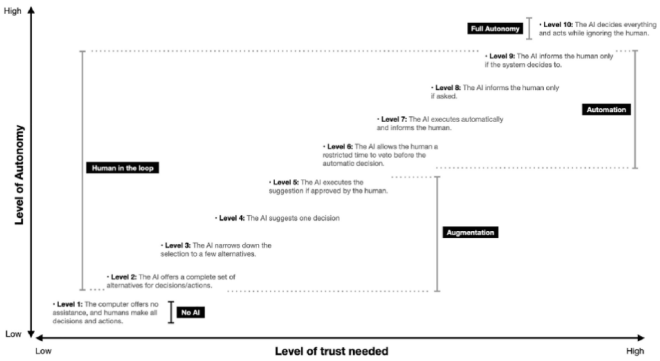

Fig. 5. Levels of Automation

G. Ethical considerations

Developing this AI product has raised several ethical concerns.

First, user privacy and consent are essential. Our prototype seeks permission before capturing images, includes a manual check to address privacy concerns, and immediately deletes images, allowing users to control their privacy and the option to opt out.

Bias in face analysis, especially by race, gender, and ethnicity, is also significant. A study highlighted such biases using datasets like UTK-Face [28], known for diversity but overrepresenting white ethnicity. Observed accuracy disparities showed a mean absolute error (MAE) of 5.05 for Asians

versus 5.7 for whites, alongside gender differences. Attempts to balance ethnicity through under- and oversampling in [28] proved ineffective for improving fairness. Recognizing these limitations in public datasets, this work employs FairLearn to assess and address bias in face analysis [12].

Finally, using face recognition in public areas like supermarkets raises privacy concerns [29]. Although our system is for age verification only, it may impact public trust. To address this, we provide a clear privacy policy explaining that images are not stored and indicate when the camera is active, fostering transparency.

V. MODEL

A. Methodology

This section details the project's methodology, covering dataset selection and exploration, data preprocessing, model architecture and fine-tuning, and evaluation metrics. These steps collectively aim to develop an accurate age estimation model tailored for age verification in supermarket alcohol purchases.

1) Data: In this project, a dataset of facial images labeled with age attributes was required. A search on Papers with Code using the keyword "age estimation" yielded 19 potential datasets. After evaluation, the UTKFace dataset was chosen due to its comprehensive annotations, diverse images, everyday people in real world situations and its wide age-range.

The UTKFace images are sourced from the internet, including the CACD and MORPH datasets. Age, gender, and race labels are initially estimated using the DEX algorithm [30] and then verified by human annotators.

a) UTKFace overview: The UTKFace data is explored by analysing the target variable, age, segmented by gender. The distribution reveals a high representation of infants and individuals around age 25 (the target demographic), with slight gender-based differences, as shown in Figure 6. Given this imbalance, a stratified split by age and gender is necessary.

To prevent data leakage and maintain model integrity, the dataset is split into training, validation, and test sets immediately after analysing the target variable. Features, specifically the images, are only examined post-split.

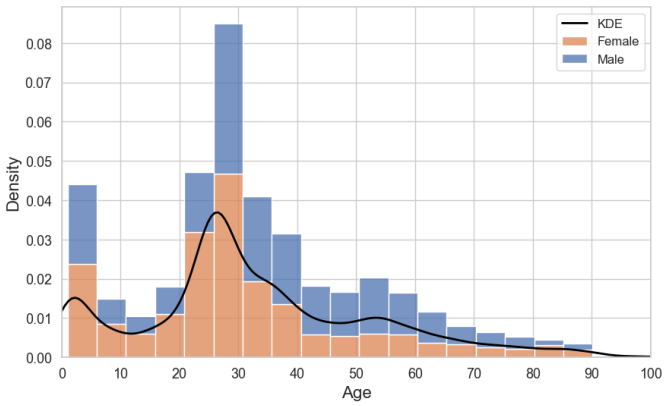

Fig. 6. Distribution of ages with gender split.

b) Train-Test-Validation Split: A 70/15/15 split for training, testing, and validation sets was implemented using a stratified approach based on age and gender. Minor adjustments were made to refine the split, such as removing images with ages over 100 (as the pre-trained model's limit is 100) and excluding files with missing metadata. Due to limited observations in certain age-gender combinations, these images were allocated to the training set. Ethnicity was excluded as a stratification criterion, as it produced too many one-observation groups, making stratification ineffective.

c) Data Exploration: We manually inspected the training images for potential issues. Key concerns included excessive zoom on facial features (e.g., eyes), multiple faces adding noise, black or coloured borders distorting data, and mislabelling, such as children labelled as older. These inconsistencies could impact training accuracy and reliability.

d) Data Cleaning: Data for ages 0 to 5 years was excluded, as this demographic is irrelevant for self-checkout scenarios in supermarkets (e.g., alcohol purchases). Figure 6, shows a skewed age distribution, which could bias the model toward overrepresented ages and reduce accuracy for underrepresented groups. Removing these ages helps reduce skew, allowing the model to focus on a more relevant age range.

The UTK dataset lacks information on repeated appearances of individuals, though manual inspection confirmed duplication. We opted not to address this issue in our study, recognizing it may affect the model's performance [31].

We acknowledge the dataset's limitations, such as inconsistencies and unclean images, as noted earlier. While comprehensive data cleaning could improve model performance, the project's time constraints and dataset size make this impractical. Therefore, we accept these imperfections and their impact on model performance.

2) Models: This subsection outlines the model framework, including baseline selection, backend performance assessment, and training setup. It describes the use of the DeepFace framework for pre-processing and face alignment, the customization of models for age estimation, and the evaluation of multiple backend models to ensure efficient face detection. Finally, it presents training configurations and the criteria for assessing age verification accuracy.

a) Model Framework: DeepFace offers a streamlined pipeline for pre-processing, recognition, and attribute analysis, including automatic face alignment to a frontal pose, resolving model inconsistencies [14]. This simplifies model testing compared to models like MiVOLO, which require custom preparation.

"DeepFace" refers to both Facebook's original face verification model [20] and an open-source framework that supports multiple facial recognition models, including its own age estimation model. This project specifically uses the open-source DeepFace framework and its age estimation model.

Originally developed for facial recognition, the DeepFace framework includes high-accuracy models such as FaceNet and VGGFace, which have achieved 99.63% and 98.95% accuracy on the LFW dataset [32], [15]. Expanding on these

TABLE I
RESULTS TESTING BACKENDS

Backend models	Invalid images	Time in Seconds
OpenCV	274 (8.51%)	0.493347
SSD	192 (5.97%)	0.207688
MTCNN	7 (0.22%)	1.088925
Fast-MTCNN	1 (0.03%)	0.305454
YOLOv8	0 (0.00%)	0.158711
YuNet	81 (2.52%)	0.131494

models, researchers incorporated additional capabilities into the DeepFace framework, enabling predictions of age, gender, race, and emotions from facial images [14], [19].

The DeepFace framework's extract_faces method aligns faces in input images, enhancing model accuracy, as demonstrated by Google's increase from 98.87% to 99.63% with face alignment [15]. Although extract_faces also offers anti-spoofing capabilities, this feature will not be used, as deceit detection is beyond the scope of requirements. To minimise artifacts from alignment, images will be cropped to retain only relevant facial features.

b) Baseline model: The decision not to train a model from scratch in this project is primarily due to the impracticality of such an approach. Deep learning models require vast amounts of data to achieve high performance. For example, Facebook's original DeepFace work used a dataset of 4 million images to train its CNN face feature extractor [20]. and VGG 2 milion. Training with these datasets would require resources far beyond those available for this project. By using pre-trained models, this approach mitigates the need for extensive data and computational resources.

The chosen baselinemodel is the VGGFace model adapted for age estimation [19]. This is a model that was transfer learned from VGGFace and given a new output head to predict ages. This custom head returns a probability distribution across the ages as its prediction when can then be manually converted to the corresponding ages. This model was developed on the IMDB-WIKI dataset and achieved a Mean Absolute Error (MAE) of 4.65 on this dataset.

c) Backend models and selection: The backend model is responsible for detecting and cropping faces in images. The DeepFace framework provides multiple backends for this task [33]. However, Dlib, Mediapipe, and CenterFace were found to be deprecated, and RetinaFace encountered frequent errors. The remaining functional backends were tested on a validation dataset containing 3,218 facial images. These backends were evaluated based on the number of invalid images (cases where the model failed to detect a face) and the average time taken to crop each face. The results are presented in Table I. It was concluded that YOLOv8 is the best-performing model, with zero invalid images and an average crop time of approximately 0.16 seconds. While YuNet is slightly faster, it failed to detect a face in 81 images. Consequently, YOLOv8 has been selected as the backend model.

d) Estimation models and selection: As previously discussed, the DeepFace framework offers several model options for facial recognition. However, due to technical limitations, incompatibility with specific TensorFlow versions, discontinued support, and limited flexibility in adjusting layers, some models were excluded from evaluation. The models excluded for these reasons include DeepFace, DeepID, Dlib, GhostFaceNet, and SFace. The remaining models—VGG-Face, OpenFace, FaceNet, and Arcface—were used in this project. Table II provides an overview of these models, indicating their input and output shapes, with the VGG-Face (age estimation) model serving as the baseline.

TABLE II
SUMMARY OF EVALUATED MODELS WITH INPUT AND OUTPUT SHAPES

Model	Input Shape	Output Shape
VGG-Face (age prediction)	224 x 224 x 3	Scalar (age)
FaceNet	160 x 160 x 3	128-d embedding
OpenFace	96 x 96 x 3	128-d embedding
Arcface	112 x 112 x 3	512-d embedding

TABLE III
KERNEL SIZES FOR CONVOLUTIONAL OUTPUT LAYER PER MODEL FOR AGE ESTIMATION.

Model	Kernel Size
FaceNet	3
OpenFace	1
ArcFace	7

The baseline model, VGG-Age, is the only of the four models that is able to estimate ages. To compare all models with each other the paper [19] was followed. Here they added a custom head to the VGG-Face model consisting of a convolutional 2D layer with an output of 1, 1, 100 followed by a flatten layer which is then activated with a Softmax activation layer. This exact layer architecture was thus also applied to the other models. However, due to their own unique architectures, different kernel sizes are used for the convolutional layer, which are shown in Table III.

Our initial goal was to evaluate all available models in DeepFace and rank them based on performance metrics relevant to age estimation, selecting the three best-performing models. However, due to technical limitations, several models had to be excluded. Consequently, only three models, along with the baseline model, remain, as shown in Table II. These models: VGG-Face (age estimation), FaceNet, OpenFace, and ArcFace are used in this study.

e) Model Training and Hyperparameter Tuning: The models were trained with different combinations of loss functions and class weights. The used loss functions were the built-in sparse cross entropy of Tensorflow, a self made sparse cross entropy, mean squared error and Mean-Variance + sparse categorical cross entropy [34]. These are then all tested with weights according to a uniform distribution $U(1)$, a normal distribution $N(20, 5)$ and weights on the density of ages according to [35]. For each combination only the class weights or loss function was changed.

All models were trained for 50 epochs with an early stopping patience of 5 at 0.01 learning rate. For the early

stopping a minimum delta in the validation loss was set to 0.0001 as in earlier iterations of testing the models got stuck at the same loss value and would iterate infinitely. Post-training, the best validation weights were finalized. An initial batch size of 32 was set according to [36] and kept at 32 due to technical limitations.

f) Output Conversion: The output of the model is a softmax distribution of all ages. These are converted to a predicted age using:

$$\text{Predicted Age} = \sum_{i=0}^{100} i \cdot p_i \tag{1}$$

Which in turn is converted to binary using a threshold. This threshold was set to 25 as all supermarkets agreed to verify the age of anyone they deem at or under 25 [3], [4]

$$y_{\text{binary}} = \begin{cases} 1 & \text{if Predicted Age} \leq 25 \\ 0 & \text{if Predicted Age} > 25 \end{cases} \tag{2}$$

The threshold of ≤ 25 is set to 1, reflecting the nature of classification metrics, which are all based on the True Positive count. To determine the number of people below 25 who were misclassified, the True Positive count must represent the number of instances where people below 25 were correctly classified.

A confidence score is also derived from the softmax output. This distribution is then split at 25, with the probabilities of all ages in each split summed to yield the probability, or confidence, that a person is over 25 or not. See 7 as example.

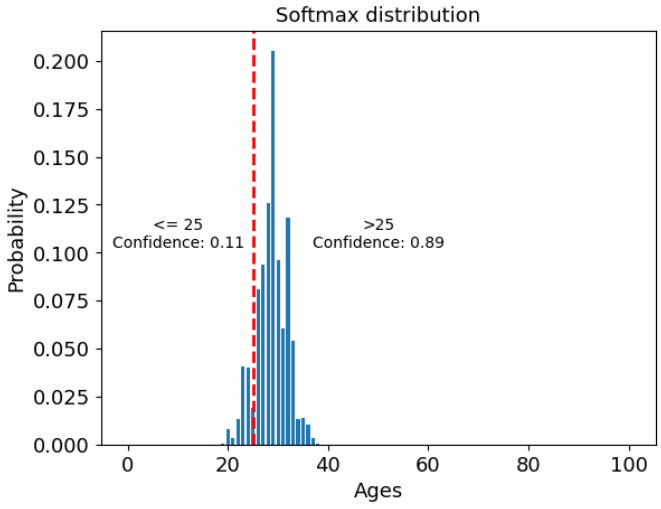

Fig. 7. A softmax prediction with confidence scores for the binary classes.

g) Metric and evaluation: For this work we use two separate metrics. One is more general for the purpose of comparing with related work, e.g. the DeepFace age estimation model [19]. The other revolves around the specific use case of our project: alcohol sales. The key metrics for this problem are recall and precision scores. Recall measures how often a person below 25 is correctly classified, while precision measures how often a person above 25 is incorrectly classified. Both metrics are important for improving efficiency; however, our main priority is the recall score, as it indicates the likelihood of a potential illegal alcohol sale.

A balance between the two could be set as a β in F1 score [37] which results in the following:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{3}$$

Where β indicates which metric is more important and by how much. When $\beta > 1$ the recall is more important and when $\beta < 1$ the precision is more important. We then, for this project, define β as:

$$\beta = \sqrt{\frac{\text{False Positive Tolerance}}{\text{False Negative Tolerance}}} \tag{4}$$

Here the tolerances indicate which of the two metrics is deemed more important, and the square root of the resulting ratio is taken because β is squared in Equation 3.

There is no predefined method to determine these tolerances so we decided to use the class imbalance as the ratio. Which then puts an emphasis on the class of interest. For the UTK dataset, with a class imbalance of $1:3$ for $<= 25$ and > 25, the β corresponds to $\sqrt{3}$ which equals 1.73.

B. Results

Training FaceNet, OpenFace, and Arcface mostly predicted ages over 25, yielding a recall of 0. Due to these results we decided to continue the tests with the model that did work (VGG-Face). The model as is was used as a baseline. The corresponding confusion matrix is seen in 8 with the metrics F_β, recall and precision shows in Table IV.

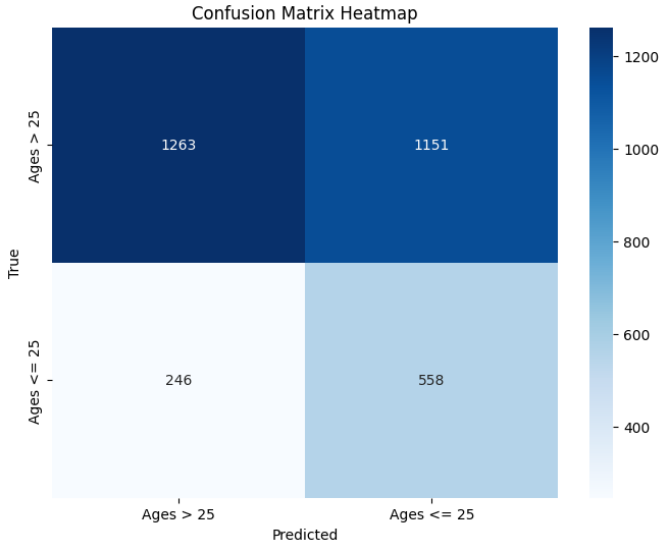

Fig. 8. Baseline confusion matrix.

Training the output layers of this model with the different loss functions and weights results in the following recall and precision tables.

TABLE IV
BASELINE SCORES (ALL METRICS)

F_β	Recall	Precision
0.5417	0.6940	0.3265

TABLE V
RECALL SCORES (ALL COMBINATIONS)

	Uniform $U(1)$	Normal $N(20, 5)$	Density
Built-in Sparse	0.1542	0.3321	0.6704
Custom Sparse	0.6294	0.4453	0.3022
Mean-Variance	0.0286	0.4042	0.1443
MSE	0.0833	0.0050	0.4353

TABLE VI
PRECISION SCORES (ALL COMBINATIONS)

	Uniform $U(1)$	Normal $N(20, 5)$	Density
Built-in Sparse	0.4351	0.4495	0.3697
Custom Sparse	0.3581	0.4197	0.4309
Mean-Variance	0.5111	0.3672	0.5249
MSE	0.5678	0.3077	0.3977

In here we can see that for the recall the version that was trained using the built-in sparse loss and density based weights performed best, followed by the custom sparse loss with uniform weights. When looking at their respective precision scores and comparing them to the baseline we can see that the recall score is worse, but the precision did improve. Calculating the F_β, using Equation 3, from Tables VI and V shows that both the trained models get a higher score than the baseline.

TABLE VII
F_β SCORES (ALL COMBINATIONS)

	Uniform $U(1)$	Normal $N(20, 5)$	Density
Built-in Sparse	0.1771	0.3504	0.5766
Custom Sparse	0.5466	0.4399	0.3214
Mean-Variance	0.0353	0.3962	0.1688
MSE	0.1004	0.0062	0.4272

This shows again that the built-in sparse with density based class weights performs best, closely followed by the custom sparse loss with uniform weights. In Figure 9 is the confusion matrix of the best scoring model shown. Comparing this to the confusion matrix of the baseline shows that the number of people correctly classified below 25 is around the same. But the number of people correctly classified above 25 is increased by over 200.

The resulting scatter plots show that both models still make quite a lot of mistakes for both people under 25 and people above 25.

a) Confidence scores: The confidence scores required for the product can also be determined from the outputs. However, when looking at the output distribution of the models, Figure 12 we can see that this does not represent a probability distribution and thus a confidence score cannot be calculated in the same way.

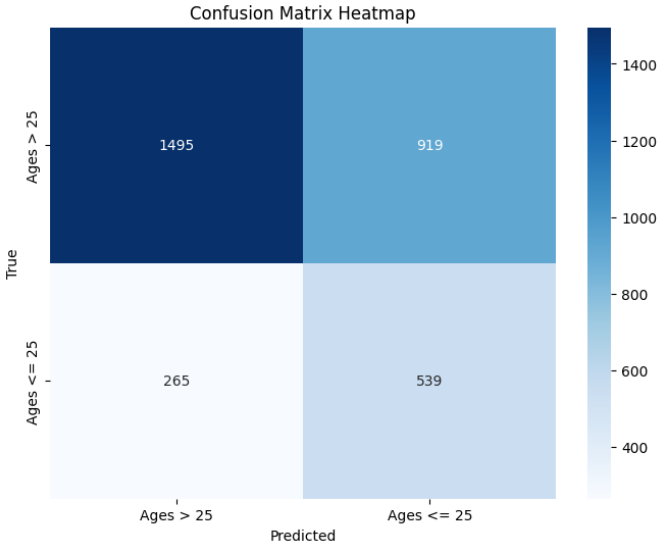

Fig. 9. Confusion matrix of the best performing model.

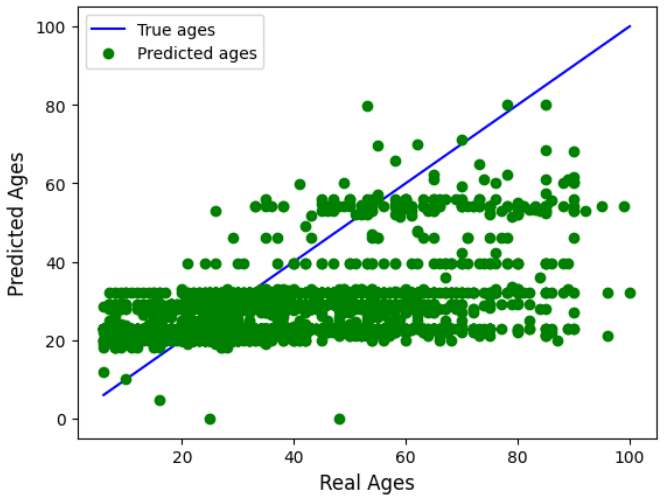

Fig. 10. Baseline predictions compared to the true ages.

The cause of this and the reason why the other models did not work was because we had forgotten to scale the data between 0 and 1 when training the models.

b) Fairness metrics: The impact of bias across different ethnicities and gender can be seen in Fig. 13 and 14. The average difference between 'others' (hispanics and latino's) and 'white' is almost 5 years. The difference between other groups is smaller, but still present. For example, 'Black'(10.35) and 'Asian' (11.27) almost differ 1 year. Furthermore, males were harder to predict than females, which does not agree with the observation of [28], which states that female faces are generally more covered with hair and therefore more difficult to predict.

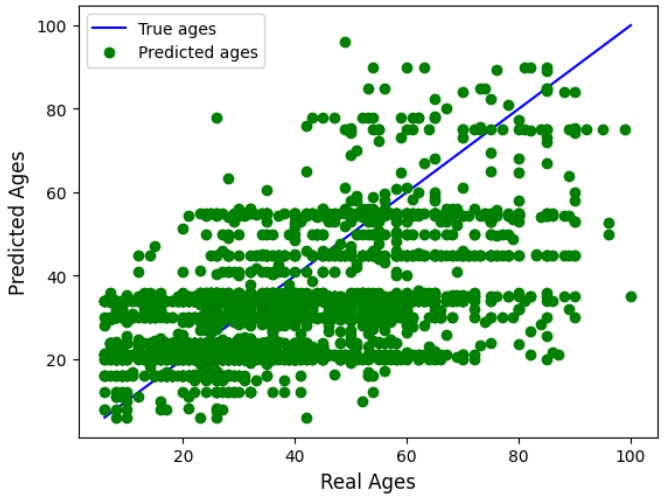

Fig. 11. Best model predictions compared to the true ages

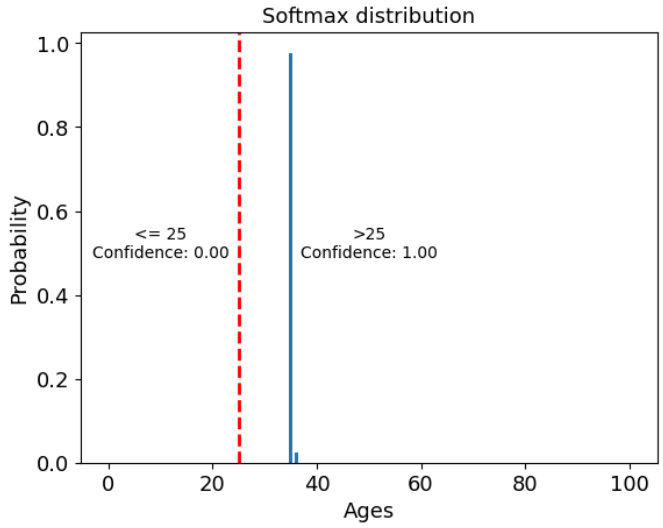

Fig. 12. Output distribution of the best model.

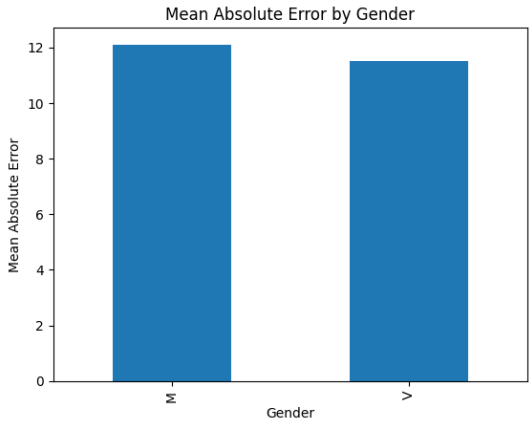

Fig. 13. MAE scores between gender.

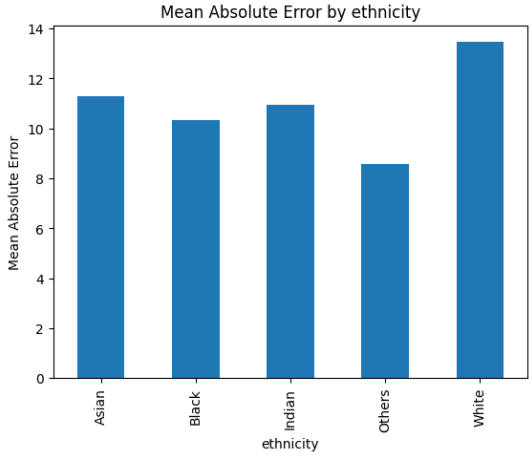

Fig. 14. MAE scores between gender.

C. Conclusion

In this study, we aimed to develop an age prediction model to assist in determining whether a customer purchasing alcohol in a supermarket is under or over the age of 25.

For the development of the model we used the UTKFace dataset on the DeepFace Framework. The DeepFace framework provides a pipeline for facial image pre-processing, recognition, and attribute analysis. On the UTKFace dataset a 70/15/15 split for the training, testing, and validation sets was applied, using a stratified approach based on age and gender.

The baseline model was the pre-trained age estimation model developed by Serengil [19]. As described in the methodology, its output architecture was replicated and applied to three other models to enable performance comparison. The models were trained with different combinations of loss functions and class weights. The key metrics for improving efficiency are the recall and precision scores. Our main priority is the recall score, as described in the methodology the β corresponds to $\sqrt{3}$ which equals 1.73. Also the recall indicates a possible illegal alcohol sale, which is the biggest risk.

Among the configurations, the model using built-in sparse loss with density-based weights emerged as the most effective, achieving the highest F_β score. As mentioned in the results, the model achieved an increase in the correct classification of individuals over 25 by over 200 (TN), while maintaining accuracy for those below 25 (TP). However, optimization in the recall for the individuals above 25 resulted in a slight decrease in precision due to an increase of 19 (265-246) in false positives (FP). This illustrates the trade-off in optimizing recall without compromising precision.

For our choice of model we decided to select the baseline model because it was the only configuration that provided confidence scores, a crucial feature for our application where reliability in age estimation is key. While our custom-trained models showed higher precision and recall compared to the baseline, they lacked the ability to generate confidence scores. If the custom-trained models had included confidence scores,

they would have been preferred over the baseline. However, due to an error in data scaling during training, as described in the results, accurate confidence estimation became impossible. Furthermore, attempts to retrain the models were hindered by server GPU overcapacity, preventing further improvements. As a result, we chose the baseline model, as it best aligns with our project's goals.

Currently everything in this process is manual and therefore 100% of the costumers are checked by the employee. Our model, although not perfect, demonstrated a significant step towards reducing manual checks by employees, contributing to increased efficiency. The model has been designed to meet all the identified must-have requirements. Also three of four should-have requirements are implemented into the product. The could have and won't have requirements where out of scope. Our value proposition to automate the age check for alcohol purchases with an AI-based solution was validated by the improvement in operational efficiency.

VI. DISCUSSION

At the start, we highlighted the problem of underage drinking and the lack of age estimation technology in Dutch supermarkets. Our research explores this technology's potential to address the issue.

During development, we encountered challenges with custom layer implementation, which affected our ability to obtain accurate confidence scores. Testing was also limited to a small sample of our target audience. This has caused our model to achieve an MAE of 11.7 years, which unfortunately is not within the SMART goal from our product requirements.

Furthermore, UK supermarkets have demonstrated the feasibility of using facial recognition to verify alcohol buyers' ages, achieving an MAE of 2.9 years. Despite the technical challenges faced in the paper, this product provides a strong proof of concept for supermarkets looking to improve alcohol purchasing efficiency through automated age estimation at self-service checkouts.

Lastly, this work identified several ethical considerations. Among them was bias in public datasets, which often results in performance differences across ethnicities and gender. By choosing the UTKFace dataset we were able to address this issue with the FairLearn framework. Although differences in performance across gender and ethnicity were not similar to other studies, the differences indicate potential unfair outcomes for some subgroups. We encourage others to contribute improvements in this area.

APPENDIX

A. EMPATHY MAP

The empathy in figure 15 is designed to understand the challenges that employees face and to identify the opportunities for improvement in the age identification process. This empathy map captures the thoughts, feelings, and observations of employees during the age verification process for alcohol purchases [38].

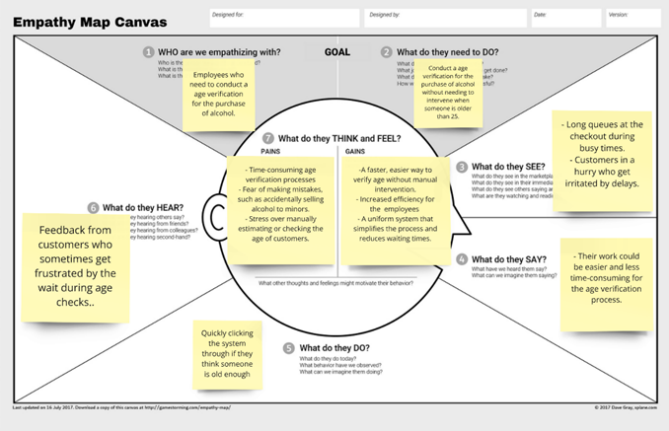

Fig. 15. Empathy Map

The following can be concluded with the empathy map: Employees experience time pressure, particularly during busy times, as the process of age identification is often time-consuming and carries the risk of making mistakes, for example, selling alcohol to minors. Employees also get feedback from frustrated customers about the wait of an age check which leads to stress for employees. The employees desire a more efficient process and a reduce in manual intervention with the help of a uniform system that simplifies the process and reduces the wait times.

B. REQUIREMENTS

The requirements are structured using the MoSCoW prioritization technique, which organizes requirements into four categories: Must have, Should have, Could have, and Won't have. This approach helps prioritize and clearly define the necessary and optional features of the project. The requirements are categorised into product requirements and model-specific requirements. All the mentioned requirements are formulated according to the SMART criteria. This section contains the full MoSCoW requirements.

1) Must-have Product Requirements:

- **Single Register Interface:** The interface must be designed to allow employees to manage a single register without interruptions.
- **Human-in-the-Loop Option:** The system must include a manual verification option.
- **User Data Handling:** The system must inform users about data handling practices, explicitly stating that photos will not be saved.

2) Should-have Product Requirements:

- **Manual Control for Groups:** The system should support manual checks when more than one face is detected
- **Countdown Timer:** The product should provide a countdown function of three seconds before capturing a photo.
- **Supermarket Chain Interface:** The interface should align with the supermarket chain's existing interface (e.g., AH-interface).

- **Multi-Register Interface:** The product should allow employees to switch between two registers at once.

3) Could-have Product Requirements:

- **Feedback System:** A feedback system (e.g., thumbs-up or thumbs-down) could be implemented to validate if the estimated age was accurate, improving system performance over time.
- **Multilingual Support:** The system could provide multilingual support, offering both Dutch and English.
- **Employee Training and Support:** Training and documentation could be provided for supermarket staff on how to use the system.

4) Won't-have Product Requirements:

- **Full Identity Verification:** The system will not perform full identity verification, only age estimation.
- **Bank or ID Integration:** The system will not integrate with banking information or identification documents.
- **Additional User Insights:** The system will not infer additional information such as the user's lifestyle (e.g., whether the user is a smoker or athlete).

5) Must-have Model Requirements:

- **User Interface:** The model must be able to receive an image and return an age prediction through a user interface.
- **Face Detection:** The model must detect faces before performing any prediction. If no face is detected, no age estimation should be done.
- **Age Estimation:** The model must estimate the customer's age achieving a Mean Absolute Error (MAE) of less than 6.0 [23].
- **Response Time:** The model must provide predictions within 10 seconds.

6) Should-have Model Requirements:

- **Confidence Score:** The model should be able to log confidence scores, and ..to decide whether a manual check is necessary. [39]
- **Multiple Faces Detection:** The model should detect multiple faces in an image and flag these instances for manual verification.

7) Could-have Model Requirements:

- **Reinforcement Learning:** The model could log feedback on whether predictions were correct, for future reinforcement learning.
- **Handling Difficult Conditions:** The model could be capable of performing over 80% confidence score under varying conditions such as different light sources, low-quality images, or turned heads. [39]

8) Won't-have Model Requirements:

- **Real-Time Prediction Adjustments:** The model will not adjust predictions in real-time based on user movements or different camera angles.
- **Deception Detection:** The model will not be able to differentiate between real faces and deceptive inputs like photos or masks.
- **Facial Recognition:** The model will be unable to do full identity verification with facial recognition.

C. DESIGN PATTERNS

While designing multiple iterations of our prototype, we also took design patterns into account, including those referenced in [26]. We believe that by implementing these design patterns, our system will be more understandable for our stakeholders. The following design patterns were implemented in our prototypes:

- **Set the right expectations**: this is important to do as we want to minimise the confusion of customers when they suddenly see the option of "automatic age estimation" with a camera popping up.
- **Determine how to show model confidence, if at all**: we are displaying advice on what the most fitting action will be for the customer based on the confidence score, instead of just showing this percentage alone.
- **Let users supervise automation**: we still want to keep the human in the loop and therefore we will give suggestions for possible actions to be taken, instead of fully automating and forcing an action to be picked.
- **Give control back to the user when automation fails**: in-case face detection is not possible there will always be a fall-back option to let the employees still perform a manual age check.
- **Be transparent about privacy and data settings**: to ensure privacy and data transparency an additional pop-up screen has been added that specifically mentions things such as "images will be immediately deleted after age estimation".
- **Explain the benefit, not the technology**: we will also only be explaining the benefit instead of the technology, as we expect that not many people will understand the technology behind it.

REFERENCES

[1] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, *Alcoholwet*, nl, Last Modified: 2024-09-26, Apr. 2024. [Online]. Available: https://wetten.overheid.nl/BWBR0002458/2024-04-01 (visited on 10/04/2024).

[2] Ministerie van Volksgezondheid, Welzijn en Sport, *Home NIX18 - NIX18*, nl-NL, webpagina, Last Modified: 2023-10-09T19:07 Publisher: Ministerie van Volksgezondheid, Welzijn en Sport, Aug. 2021. [Online]. Available: https://www.nix18.nl/ (visited on 10/24/2024).

[3] Centraal Bureau Levensmiddelenhandel, "CBL Code 'Verantwoorde alcoholverkoop in de supermarkt'," nl-NL, Centraal Bureau Levensmiddelenhandel, Tech. Rep., 2022. [Online]. Available: https://www.cbl.nl/onderwerpen/verkoop-van-alcohol/ (visited on 10/23/2024).

[4] Centraal Bureau Levensmiddelenhandel, "Verantwoorde alcoholverkoop in de supermarkt," nl-NL, Centraal Bureau Levensmiddelenhandel, Tech. Rep., 2024. [Online]. Available: https://www.cbl.nl/onderwerpen/verkoop-van-alcohol/ (visited on 10/23/2024).

[5] S. van Dorsselaer, M. de Looze, M. Boer, S. de Roos, H. Brons, and R. van den Eijnden, "HBSC 2021," nl-NL, Trimbos instituut, Tech. Rep., 2022, p. 183. [Online]. Available: https://www.trimbos.nl/aanbod/webwinkel/af2022-hbsc-2021/ (visited on 10/04/2024).

[6] A. Cheng, B. Combee, J. Mekkelholt, T. Kahmann, and R. Steenhorst, "Logboek Team 9," NL, Hogeschool van Amsterdam, Amsterdam, Logbook, Nov. 2024.

[7] Ministerie van Algemene Zaken, "Landelijk onderzoek naleving leeftijdsgrens bij de verkoop van alcohol en tabak in 2022," nl-NL, Rijksoverheid, rapport, Aug. 2022, Last Modified: 2022-10-04T09:34 Publisher: Ministerie van Algemene Zaken. [Online]. Available: https://www.rijksoverheid.nl/documenten/rapporten/2022/08/31/rapport-landelijk-onderzoek-naar-de-naleving-van-de-leeftijdsgrens-bij-alcohol-en-tabaksverkoop-in-2022 (visited on 10/23/2024).

[8] R. Halfhide and R. Beijersbergen, "Uitgescholden en bedreigd: Jonge zelfscanmedewerkers voelen zich onveilig," nl, *NOS*, Mar. 2023. [Online]. Available: https://nos.nl/artikel/2469362-uitgescholden-en-bedreigd-jonge-zelfscanmedewerkers-voelen-zich-onveilig (visited on 10/22/2024).

[9] J. Wakefield, "Supermarket cameras to guess age of alcohol buyers," en-GB, *BBC*, Feb. 2022. [Online]. Available: https://www.bbc.com/news/technology-60215258 (visited on 10/27/2024).

[10] C. Hymas, "AI face-scanning technology to be rolled out at supermarkets to check age of shoppers," en-GB, *The Telegraph*, Jan. 2024, ISSN: 0307-1235. [Online]. Available: https://www.telegraph.co.uk/news/2024/01/24/ai-face-scanning-technology-supermarkets-age-alcohol-id/ (visited on 10/27/2024).

[11] Yoti Ltd., "Yoti Facial Age Estimation," en-GB, Yoti, Whitepaper, Mar. 2023. [Online]. Available: https://www.yoti.com/wp-content/uploads/Yoti-Age-Estimation-White-Paper-March-2023.pdf (visited on 10/27/2024).

[12] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and Improving Fairness of AI Systems," *Journal of Machine Learning Research*, vol. 24, 2023. [Online]. Available: http://jmlr.org/papers/v24/23-0389.html.

[13] S. Studer, T. B. Bui, C. Drescher, *et al.*, *Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology*, arXiv:2003.05155, Feb. 2021. [Online]. Available: http://arxiv.org/abs/2003.05155 (visited on 10/24/2024).

[14] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Istanbul, Turkey: IEEE, Oct. 2020, pp. 1–5, ISBN: 978-1-72819-136-2. DOI: 10.1109/ASYU50717.2020.9259802. [Online]. Available: https://ieeexplore.ieee.org/document/9259802/ (visited on 10/04/2024).

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," en, in *Procedings of the British Machine Vision Conference 2015*, Swansea: British Machine Vision Association, 2015, pp. 41.1–41.12, ISBN: 978-1-901725-53-7. DOI: 10.5244/C.29.41. [Online]. Available: http://www.bmva.org/bmvc/2015/papers/paper041/index.html (visited on 10/24/2024).

[16] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments," en, HAL, 2008. [Online]. Available: https://inria.hal.science/inria-00321923 (visited on 10/27/2024).

[17] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep EXpectation of Apparent Age from a Single Image," en, in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile: IEEE, Dec. 2015, pp. 252–257, ISBN: 978-1-4673-9711-7. DOI: 10.1109/ICCVW.2015.41. [Online]. Available: http://ieeexplore.ieee.org/document/7406390/ (visited on 10/22/2024).

[18] Z. Zhang, Y. Song, and H. Qi, *Age Progression/Regression by Conditional Adversarial Autoencoder*, arXiv:1702.08423, Mar. 2017. DOI: 10.48550/arXiv.1702.08423. [Online]. Available: http://arxiv.org/abs/1702.08423 (visited on 10/24/2024).

[19] S. I. Serengil and A. Ozpinar, "HyperExtended LightFace: A Facial Attribute Analysis Framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, Istanbul, Turkey: IEEE, Oct. 2021, pp. 1–4, ISBN: 978-1-66542-714-2. DOI: 10.1109/ICEET53442.2021.9659697. [Online]. Available: https://ieeexplore.ieee.org/document/9659697/ (visited on 10/04/2024).

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA: IEEE, Jun. 2014, pp. 1701–1708, ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.220. [Online]. Available: https://ieeexplore.ieee.org/document/6909616 (visited on 10/04/2024).

[21] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015, ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2014.2362759. [Online]. Available: http://ieeexplore.ieee.org/document/6920084/ (visited on 10/27/2024).

[22] D. Walker, L. Bourne, and A. Shelley, "Influence, stakeholder mapping and visualization," *Construction*

Management Economics, vol. 26, pp. 645–658, Jun. 2008. DOI: 10.1080/01446190701882390.

[23] E. Moyse, "Age estimation from faces and voices: A review," English, *Psychologica Belgica*, vol. 54, no. 3, 2014, ISSN: 0033-2879. DOI: 10.5334/pb.aq.

[24] T. Rintamäki and H. Saarijärvi, "An integrative framework for managing customer value propositions," *Journal of Business Research*, vol. 134, pp. 754–764, 2021, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2021.05.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296321003593.

[25] Alex Cheng, Bart Combee, Jim Mekkelholt, Rick Steenhorst, and Tijn Kahmann, *Maai / 2024-2025-1-b1 / class-2-team-09 · GitLab*, en, Gitlab repository, Sep. 2024. [Online]. Available: https://gitlab.fdmci.hva.nl/maai/2024-2025-b1/class-2-team-09 (visited on 11/04/2024).

[26] Google. "People + ai guidebook." Accessed: Oct. 04, 2024. (2024), [Online]. Available: https://pair.withgoogle.com/guidebook.

[27] Europees Parlement en de Raad, *VERORDENING (EU) 2016/679 VAN HET EUROPEES PARLEMENT EN DE RAAD*, NL, Usr_lan: NL, Apr. 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/NL/TXT/HTML/?uri=CELEX%3A32016R0679 (visited on 10/23/2024).

[28] N. Panić, M. Marjanović, and T. Bezdan, "Addressing demographic bias in age estimation models through optimized dataset composition," *Mathematics*, vol. 12, no. 15, 2024, ISSN: 2227-7390. DOI: 10.3390/math12152358. [Online]. Available: https://www.mdpi.com/2227-7390/12/15/2358.

[29] M. C. Paganoni, "Ethical Concerns over Facial Recognition Technology," eng, *Anglistica AION*, vol. 23, no. 1, pp. 85–94, Aug. 2020, ISSN: 2035-8504. DOI: 10.19231/angl-aion.201915. [Online]. Available: https://doi.org/10.19231/angl-aion.201915 (visited on 10/04/2024).

[30] Y. Song and Z. Zhang, *Utkface: A large-scale face dataset*, 2017. [Online]. Available: https://susanqq.github.io/UTKFace/ (visited on 09/30/2024).

[31] J. Paplham and V. Franc, *A Call to Reflect on Evaluation Practices for Age Estimation: Comparative Analysis of the State-of-the-Art and a Unified Benchmark*, arXiv:2307.04570, Mar. 2024. DOI: 10.48550/arXiv.2307.04570. [Online]. Available: http://arxiv.org/abs/2307.04570 (visited on 10/23/2024).

[32] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015. arXiv: 1503.03832. [Online]. Available: http://arxiv.org/abs/1503.03832.

[33] S. Serengil and A. Özpınar, "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules," *Bilişim Teknolojileri Dergisi*, vol. 17, no. 2, pp. 95–107, Apr. 2024, ISSN: 1307-9697. DOI: 10.17671/gazibtd.1399077. [Online]. Available: http://dergipark.org.tr/en/doi/10.17671/gazibtd.1399077 (visited on 10/15/2024).

[34] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5285–5294. DOI: 10.1109/CVPR.2018.00554.

[35] M. Steininger, K. Kobs, P. Davidson, A. Krause, and A. Hotho, "Density-based weighting for imbalanced regression," *Machine Learning*, vol. 110, no. 8, pp. 2187–2211, Aug. 2021. DOI: 10.1007/s10994-021-06023-5. [Online]. Available: https://doi.org/10.1007/s10994-021-06023-5.

[36] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org (visited on 10/14/2024).

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] P. Sinansari, S. Salsabila, S. Hanoum, A. Łopatka, and W. Wlodarski, "Identify customer element through empathy map and user persona," *Procedia Computer Science*, vol. 225, pp. 4148–4156, Jan. 2023. DOI: 10.1016/j.procs.2023.10.411.

[39] M. Wood, *P values, confidence intervals, or confidence levels for hypotheses?* 2014. arXiv: 0912.3878. [Online]. Available: https://arxiv.org/abs/0912.3878.