

MD-ViT: Multidomain Vision Transformer Fusion for Fair Demographic Attribute Recognition

Rezky Arisanti Putri^{1*}, Ricky Eka Putra², Yuni Yamasari³

^{1,2,3} Department of Informatics, Universitas Negeri Surabaya, Surabaya, Indonesia

Corresponding author: rickyeka@unesa.ac.id

ARTICLE INFO

Article history:

Submitted 24-11-2025

Accepted 5-12-2025

Available online 6-12-2025

Keywords:

Vision Transformer, demographic classification, fairness, feature fusion, XGBoost

DOI:

<https://doi.org/10.26740/jieet.v9n2.p64-79>

ABSTRACT

Demographic attribute recognition, particularly race and gender classification from facial images, plays a critical role in applications ranging from precision healthcare to digital identity systems. However, existing deep learning approaches often suffer from algorithmic bias and limited robustness, especially when trained on imbalanced or non-representative data. To address these challenges, this study proposes MD-ViT, a novel framework that leverages multidomain Vision Transformer (ViT) fusion to enhance both accuracy and fairness in demographic classification. Specifically, we integrate embeddings from two task-specific pretrained ViTs: ViT-VGGFace (fine-tuned on VGGFace2 for structural identity features) and ViT-Face Age (trained on UTKFace and IMDB-WIKI for age-related morphological cues), and then classify using XGBoost to model complex feature interactions while mitigating overfitting. Evaluated on the balanced DemogPairs dataset (10,800 images across six intersectional subgroups), our approach achieves 89.07% accuracy and 89.06% F1-score, outperforming single-domain baselines (ViT-VGGFace: 88.61%; ViT-Age: 78.94%). Crucially, fairness analysis reveals minimal performance disparity across subgroups (F1-score range: 87.38%–91.03%; $\sigma = 1.33$), indicating effective mitigation of intersectional bias. These results demonstrate that cross-task feature fusion can yield representations that are not only more discriminative but also more equitable. We conclude that MD-ViT offers a principled, modular, and ethically grounded pathway toward fairer soft biometric systems, particularly in high-stakes domains such as digital health and inclusive access control.



This work is licensed under the Creative Commons Attribution
Non-Commercial-Share Alike 4.0 International License.

INTRODUCTION

The human face is one of the most information-rich biometric modalities in identity identification and verification systems (Jatain & Jailia, 2023). Beyond its central role in individual recognition, facial images also encode key demographic attributes such as gender, age, and race that are highly relevant across diverse technological applications, including public security, digital forensics, precision healthcare, and socio-demographic analysis (Iloanusi et al., 2022). In clinical

settings, for instance, demographic inference from facial appearance can support early detection of medical conditions with epidemiologically distinct distributions across ethnic or gender groups (Bonner et al., 2023; Ha et al., 2021). Consequently, both the accuracy and fairness of demographic classification systems have become critical considerations in the design of responsible artificial intelligence (AI)-based solutions.

Nevertheless, developing such systems presents two significant challenges: (i) Technical challenges, stemming from intrinsic variations in facial appearance due to genetic, environmental, and contextual factors, including pose, illumination, and expression; and (ii) Ethical challenges, arising from potential algorithmic bias, particularly when training data fail to reflect the true diversity of the target population (Scheuerman et al., 2020; Sehrawat & Ali, 2023). Class imbalance in datasets often leads to uneven model performance across minority groups. (Karkkainen & Joo, 2021; Nixon et al., 2025). Recent studies emphasise that effective bias mitigation must begin at the data curation stage: fair and inclusive representation in training data is a foundational requirement that cannot be fully compensated for by architectural adjustments or post-hoc correction strategies alone. (Kotwal & Marcel, 2024; Robinson et al., 2020).

On the architectural front, Convolutional Neural Networks (CNNs), such as ResNet-50, have long served as the backbone of facial analysis systems, including demographic classification. (Singh & Chauhan, 2023). However, CNNs' reliance on local convolutional operations inherently limits their ability to capture long-range spatial dependencies among facial components. In response, Vision Transformers (ViTs) have recently demonstrated competitive advantages through their self-attention mechanism, enabling global and holistic image processing. (Ranftl et al., 2021). Empirical evidence shows that ViTs not only achieve higher accuracy on various facial analysis tasks but are also more robust to real-world challenges such as partial occlusion (e.g., masks or glasses) and pose variation. (Al-Otaiby & El-Alfy, 2023; Gao et al., 2022).

More importantly, ViTs pretrained on task-specific facial datasets such as identity recognition (VGGFace), age estimation, or facial expression analysis yield highly informative and transferable feature representations. (Bulat et al., 2022). However, the majority of existing studies still rely on features from a single task domain, thereby underutilising the rich, multidimensional information naturally embedded in human faces.

Building upon these insights, this study proposes an innovative cross-task feature fusion approach: we integrate embeddings from two ViT models pretrained on distinct domains, ViT-VGGFace (structural/identity features) and ViT-Age (age-related morphological cues), to construct a more comprehensive and discriminative facial representation. This strategy draws on the principle of multidomain learning, which has been proven effective in related tasks such as deepfake detection (Ding et al., 2024). The fused features are then classified using XGBoost, selected for its strong capacity to model nonlinear feature interactions and its tendency to produce robust generalisation through explicit regularisation and staged optimisation (Wiens et al., 2025). By comparing performance across configurations, single-domain features (ViT-VGGFace or ViT-Age) versus fused features (ViT-VGGFace + ViT-Age), we rigorously assess the consistency of cross-domain fusion benefits and evaluate how classifier choice influences system stability and overall effectiveness. In summary, the key contributions of this work are: (i) A novel cross-task

ViT-based feature fusion framework that leverages transferable knowledge from complementary facial analysis tasks to enhance demographic representation; and (ii) A comprehensive empirical evaluation quantifying the added value of multidomain fusion over conventional single-domain approaches.

By integrating rich representation learning, robust architecture design, and a commitment to data inclusivity, this research advances the development of soft biometric systems that are not only accurate but also fair, transparent, and ready for responsible real-world deployment, particularly in high-stakes domains such as digital health and identity-based access control.





METHOD



This section delineates the experimental framework employed in this study, encompassing dataset specification, model architectures, feature extraction and fusion strategies, classifier configuration, and evaluation protocol. The proposed MD-ViT pipeline integrates multidomain pretrained Vision Transformers with XGBoost-based classification to achieve robust and fair demographic attribute recognition, balancing technical performance and ethical fairness in a systematic and reproducible manner.

1. Dataset

The DemogPairs dataset is a specialised benchmark designed to measure and analyse demographic bias in deep learning-based face recognition systems. It comprises 10,800 facial images, evenly distributed across six demographic subgroups: Asian Males, Asian Females, Black Males, Black Females, White Males, and White Females. (Hupont & Fernández, 2019). From these images, approximately 58.3 million identity verification pairs were constructed, each labelled as either genuine (both images belong to the same individual) or impostor (images from different individuals). This structure enables fine-grained performance evaluation across demographic subgroups, thereby facilitating rigorous fairness analysis and robustness assessment against intergroup bias in large-scale face verification scenarios. Table 1 illustrates sample images from each class in the DemogPairs dataset.

Table 1. Sample Images from the DemogPairs Dataset.

Sample Images	Race_Gender Label
	Asian_Females
	Asian_Males
	Black_Females
	Black_Males

Sample Images	Race_Gender Label
	White_Females
	White_Males

2. Vision Transformer (ViT)

Vision Transformer (ViT) is a deep learning architecture based on the Transformer framework, introduced by (Dosovitskiy et al., 2020) in the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." Unlike Convolutional Neural Networks (CNNs), which rely on local convolutional operations, ViT processes images by dividing them into fixed-size patches and treating each patch as a token akin to word tokens in natural language processing (NLP). Through the self-attention mechanism, ViT models global spatial relationships across image regions, enabling it to capture not only local features but also the holistic context of the entire object.

In the context of demographic attribute classification, such as race and ethnicity, this approach offers several key advantages. First, ViT can holistically model interactions among facial features (e.g., the relationship between eye shape, nose structure, and skin tone distribution), rather than relying solely on isolated local regions. Second, the flexibility of self-attention allows the model to dynamically assign higher weights to the most discriminative features for a given demographic group, thereby improving robustness against variations in expression, pose, and illumination. Third, since ViT does not depend on rigid local patterns, it is more adaptable to phenotypic diversity across populations, potentially mitigating the race-other-effect, a phenomenon wherein models exhibit reduced accuracy when recognising faces from ethnic minority groups. (Pereira et al. 2024).

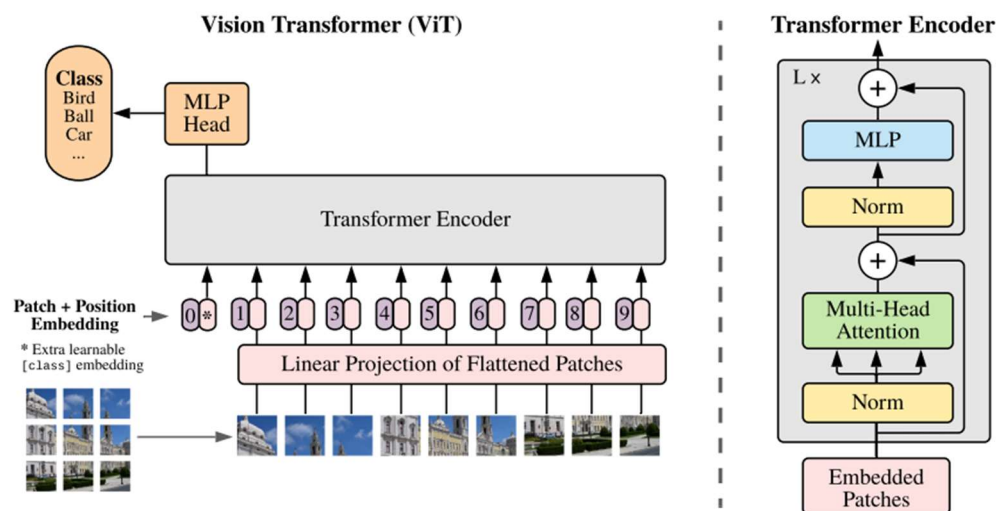


Figure 1. Vision Transformer Architecture (Dosovitskiy et al., 2020)

Architecturally, as illustrated in Figure 1, Vision Transformer (ViT) begins with image patching, where the input image is divided into N fixed-size patches; each patch is flattened and projected into an embedding space via a linear layer. To preserve spatial positional information, positional encoding is added to each patch embedding before it enters the Transformer Encoder. The core component of the encoder is Multi-Head Self-Attention (MHSA), which enables each patch to interact with all other patches via an attention score computed as defined in (1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where Q (query), K (key), and V (value) are linear projections of the input embeddings, and d_k denotes the dimensionality of the key vectors, introduced to ensure numerical stability. To enhance representational capacity, the self-attention mechanism is executed in parallel across h attention heads; the resulting outputs are then concatenated and linearly projected using the output weight matrix W^O , as specified in (2).

$$MHSA(X) = concat(head_1, \dots, head_h)W^O \quad (2)$$

Following the Multi-Head Self-Attention (MHSA) stage, each token is independently processed by a Feed-Forward Network (FFN), a module composed of two linear layers separated by the non-linear activation function Gaussian Error Linear Unit (GELU). The FFN is defined in (3).

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \quad (3)$$

The GELU function can be analytically approximated in (4).

$$GELU(x) \approx 0.5x\left(1 + tanh\left[\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right]\right) \quad (4)$$

FFN enriches token representations by mapping MHSA outputs into a more expressive feature space after global contextual information has been extracted. To ensure training stability in deeply stacked architectures, each sub-module (MHSA and FFN) is equipped with a residual connection and layer normalisation. The residual connection preserves information from the input by adding it directly to the sub-module's output, while layer normalisation stabilises the activation distribution per sample, as formalised in (5).

$$LayerNorm(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta \quad (5)$$

Where μ and σ^2 denote the mean and variance of the input vector components, respectively, and γ and β are learnable scaling and shifting parameters. After passing through the entire stack of encoder layers, the [CLS] token representation, which aggregates information from all image patches via self-attention, is extracted as the final feature vector. Although in many implementations this vector is fed directly into a classification head, in this study, the ViT is employed exclusively as a feature extractor: the [CLS] token embedding is taken as the latent representation and subsequently passed to an external classifier (XGBoost). This design enhances modularity, interpretability, and fine-grained control over the decision-making process.

3. Pretrained ViT-VGGFace

VIT-VGGFace is a Vision Transformer (ViT) model fine-tuned on the VGGFace2 dataset. (Greco et al., 2020), a large-scale face recognition benchmark comprising 3.31 million images from 9,131 unique identities, exhibiting extreme variations in pose, illumination, age, ethnicity, expression, and accessories, thereby reflecting realistic in-the-wild conditions (skutaada, 2024). In contrast to CNN-based architectures, VIT-VGGFace leverages self-attention to capture global spatial correlations among facial features, such as the relationships between jawline shape, eye position, and nasal structure, yielding a 768-dimensional embedding that is semantically rich, highly discriminative, and invariant to external perturbations, such as head rotation or lighting changes. Thanks to end-to-end fine-tuning initialised from ImageNet-21k pretrained weights, the model becomes sensitive to persistent identity cues, such as interpupillary distance and the zygomatic bone contour, which remain stable despite variations in expression or ageing. The resulting embedding enables zero-shot face verification via cosine similarity, making it highly efficient for integration into biometric systems, albeit at a higher computational cost. Overall, VIT-VGGFace provides a more comprehensive, robust, and generalizable facial representation, making it an ideal candidate for high-stakes applications such as authentication, security, large-scale face clustering, and deep learning-based demographic analysis.

4. Pretrained ViT-Face Age

ViT Facial Age Image Detection is a Vision Transformer (ViT) model fine-tuned for age estimation from facial images, leveraging a combined dataset of UTKFace (~23.7k images) and IMDB-WIKI (~983k images), spanning ages 0–100+. (dima806, 2023). Rather than performing age prediction via regression, the model adopts a classification-based approach, dividing age into 23 non-uniform bins from '01' (1–2 years) to '90+', structured to reflect the non-linear pace of facial morphological change (denser bins in early childhood, sparser in later adulthood). Training is conducted end-to-end, initialised with ImageNet-21k-pretrained weights. (dima806, 2023), enabling the model to holistically capture age-related cues through self-attention, including spatial proportions (e.g., eye-to-chin ratio, relative nose width), skin texture, and changes in facial volume, going beyond local features such as wrinkles alone.

Despite significant class imbalance (younger age groups being heavily overrepresented), the model demonstrates robust performance, thanks to standardised preprocessing including ImageNet-style normalisation, horizontal flipping, and brightness/contrast augmentation, as well as the noise-resilient age binning strategy, which mitigates label ambiguity inherent in coarse age annotations. Notably, the model achieves competitive accuracy without explicitly utilising gender or race information, successfully distinguishing developmental stages from childhood traits (full cheeks, large eyes) to ageing signs (wrinkles, loss of skin elasticity). This makes it well-suited for real-world applications such as automated demographic analysis, age-based access control, or digital service personalisation, where a balance among age granularity, prediction reliability, and resilience to data bias is essential.

5. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an enhanced version of the Gradient Boosting Machine (GBM), offering improvements in computational efficiency, convergence speed, and robustness against overfitting through explicit regularisation on both tree structure and leaf

weights. Built upon the principle of additive modelling, XGBoost constructs an ensemble of decision trees sequentially, where each new tree is trained to minimise the residuals (errors) of the cumulative predictions from previous iterations, typically by optimising a loss function such as log-loss for multiclass classification tasks (Wiens et al., 2025). Regularisation is implemented via hyperparameters such as λ (the L2 penalty on leaf weights) and γ (the minimum loss reduction required to split a node), which effectively constrain model complexity and improve generalisation. In addition to its high predictive performance, XGBoost provides feature importance metrics, enabling interpretability by quantifying each feature's relative contribution to class discrimination.

Mathematically, XGBoost initiates the learning process with an initial prediction equal to the mean of the training labels, as defined in (6).

$$f_0(x) = \frac{1}{n} \sum_{i=1}^n y_i \quad (6)$$

The residual vector is then computed as defined in (7).

$$\hat{Y} = y - f_0(X) \quad (7)$$

This residual serves as the target for the first weak learner (decision tree). In each boosting iteration, a decision tree is grown by selecting optimal splits based on two key criteria: similarity score and gain. The similarity score quantifies the homogeneity of residuals within a node and is formulated in (8).

$$Similarity = \frac{(\sum \hat{y})^2}{\sum [previous f_1(x_i) \times (1 - previous f_1(x_i)) + \lambda]} \quad (8)$$

Where $\lambda \geq 0$ is the L2 regularisation parameter that penalises nodes with few observations, preventing them from dominating tree growth. A higher similarity score indicates greater residual coherence within the node. The gain measures the loss reduction achieved by a candidate split and is calculated as the difference between the combined similarity of the child nodes and that of the parent node, as defined in (9).

$$Gain = (Left Similarity + Right Similarity) - Root Similarity \quad (9)$$

Only splits with maximal gain are retained; nodes with gain below a specified threshold are pruned to control model complexity. Once the tree structure is finalised, the output value (weight) of each leaf is computed in (10).

$$Output Value = \frac{\sum \hat{y}_i}{\sum [F_{i-1}(x_i) \times (1 - F_{i-1}(x_i))] + \lambda} \quad (10)$$

Where I_j denotes the set of samples in leaf j , and w_j represents the additive update to the cumulative prediction. The final prediction after T iterations is obtained using the logistic (sigmoid) function (11).

$$F_n(x) = \frac{1}{1 + e^{-\left(\frac{h_0(x)}{1 - h_0(x)} + \sum_{i=1}^n [\eta \times h_i(x)]\right)}} \quad (11)$$

Where η is the learning rate (shrinkage parameter) that controls the contribution of each tree, and $h_t(x)$ is the prediction from the t -th tree. This formulation maps the raw output to a

probability value in $[0,1]$, suitable for binary classification. XGBoost's superiority stems from its integration of explicit regularisation ($\lambda, \gamma, \text{min_child_weight}$), second-order gradient optimisation, and gain-based pruning. Together, these mechanisms enhance convergence stability, mitigate overfitting, and enable high scalability, making XGBoost a preferred choice for large-scale data-driven applications (Wiens et al., 2025).

6. Model Architecture

Figure 2 illustrates the end-to-end framework for facial-based demographic classification (race and gender), integrating cross-domain feature extraction using Vision Transformers (ViT) and the XGBoost classification algorithm. The pipeline consists of seven main stages, as depicted in Figure 2.

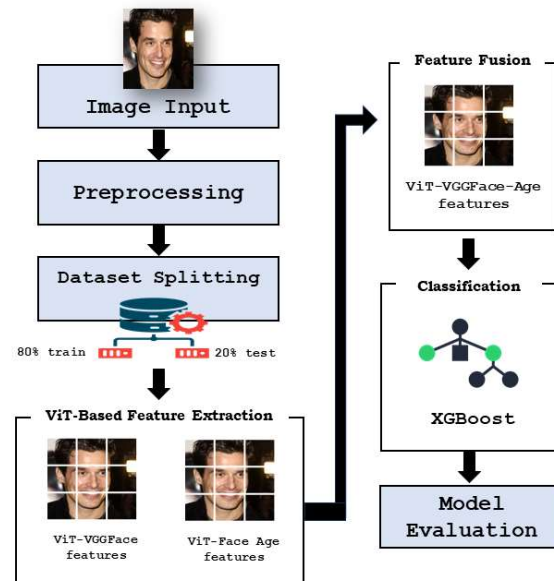


Figure 2. Model Architecture

Stage 1: Image Input

The process begins with the input of raw facial images sourced from the DemogPairs dataset, a curated, manually annotated collection labelled for race and gender attributes.

Stage 2: Preprocessing

All facial images, originally varying in size and spatial resolution, are resized to a standardised dimension of 224×224 pixels. This resolution aligns with the architectural specifications of the Vision Transformer (ViT), which is explicitly designed to process inputs of this size. Adopting this standard enables direct utilisation of pretrained weights without requiring structural modifications or retraining of the initial embedding layers, as validated in prior studies. (Dosovitskiy et al., 2020).

Stage 3: Dataset Splitting

The dataset is randomly partitioned into two subsets: 80% for training and 20% for testing. The split is performed stratified by class labels (race and gender) to preserve the original class

distribution in both subsets. The training set is used to fine-tune two ViT architectures in parallel: ViT-Face (for structural facial features) and ViT-Age (for age-related texture and morphological cues), both initialised from ImageNet-pretrained weights and subsequently adapted to their respective downstream tasks. The test set, held out entirely from the training process, is used only once to evaluate the final model's performance, using metrics such as Accuracy, Precision, Recall, and F1-Score.

Stage 4: ViT-Based Feature Extraction

At this stage, two distinct Vision Transformer models are employed in parallel to extract domain-specific features. First, ViT-VGGFace extracts global structural features, including facial bone structure, symmetry, and spatial proportions, by leveraging the self-attention mechanism to model long-range dependencies across image patches. Second, ViT-Face Age focuses on local texture features, such as wrinkles, skin pores, and pigmentation patterns. To reduce colour-induced bias and emphasise textural patterns, this model uses grayscale-converted inputs.

Stage 5: Feature Fusion

Each ViT outputs a fixed-dimensional embedding vector (768-D). These are concatenated to form a 1536-dimensional fused representation that integrates complementary information from both the structural and textural domains. This fusion is not a naive aggregation but the core innovation of the proposed framework, enabling the model to become more robust to real-world variations in pose and illumination. To rigorously assess the contribution of each component, a systematic ablation study is conducted across three configurations: (i) Full model using fused features from both ViTs, (ii) ViT-VGGFace, and (iii) ViT-Face Age.

Stage 6: Classification

The features extracted from the two ViT models are used as inputs to the XGBoost classifier. XGBoost was selected for its strong capacity to model nonlinear feature interactions and robustness against overfitting, enabled by explicit regularisation mechanisms including L2 regularisation of leaf weights, a minimum gain threshold for splitting, and shrinkage via the learning rate. To ensure optimal configuration, hyperparameter tuning was performed using a 5-fold cross-validation grid search on the training set. The search ranges and selected optimal values are presented in Table 2.

Table 2. Optimised XGBoost Hyperparameters

Parameter	Tested Value
tree_method	['approx', 'hist']
max_depth	[3, 6, 8]
gamma	[0.0, 0.1, 0.3]
min_child_weight	[1, 3, 5]
n_jobs	[1]
random_state	[42]

The optimal hyperparameter combination minimised validation log-loss while maintaining a stable F1-score across demographic subgroups, indicating that regularisation successfully

mitigated overfitting without compromising fairness. After tuning, the model was retrained on the whole training set and evaluated *once* on the completely held-out test set.

Stage 7: Model Evaluation

Model performance is evaluated using four standard confusion matrix–based metrics: Accuracy, Precision, Recall, and F1-Score. The confusion matrix compares model predictions against ground-truth labels across four fundamental components. (Pardede & Kleb, 2024): (i) True Positive (TP): correctly predicted positive samples, (ii) True Negative (TN): correctly predicted negative samples, (iii) False Positive (FP): negative samples misclassified as positive, and (iv) False Negative (FN): positive samples misclassified as negative. The metrics are computed as follows in (12)-(15).

Accuracy measures the overall proportion of correct predictions as computed in (12).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Precision assesses the reliability of optimistic predictions as computed in (13).

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

Recall (or *sensitivity*) quantifies the model's ability to identify all actual positives as computed in (14).

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

F1-Score, the harmonic mean of Precision and Recall, provides a balanced measure, especially critical for multiclass or imbalanced settings, as computed in (15).

$$F1 - Score = \frac{2 \times (Precision \cdot Recall)}{Precision + Recall} \quad (15)$$

These metrics are complementary: Accuracy provides a macro-level performance overview, while Precision, Recall, and F1-Score enable fine-grained analysis of potential biases or performance gaps across demographic subgroups, which is essential for ensuring fairness and reliability in demographic attribute recognition systems.

RESULTS

This section presents the experimental results of the proposed MD-ViT framework, structured to facilitate clear interpretation through tabulated performance metrics and subgroup analyses. The findings are organised into two main parts: (1) overall model evaluation, comparing the effectiveness of single-domain versus fused multidomain ViT features when paired with XGBoost; and (2) fairness analysis across demographic subgroups, assessing Equity in performance across intersectional categories of race and gender. Each result is followed by a targeted discussion that interprets the empirical outcomes in light of the research objectives, namely, improving accuracy and fairness in demographic attribute recognition through cross-task feature fusion.

1. Model Evaluation ViT and XGBoost

This subsection presents a comparative performance analysis of three feature extraction strategies: ViT-Face Age, ViT-VGGFace, and their fused counterpart, ViT-VGGFace-Age, all paired with XGBoost as the downstream classifier. The evaluation focuses on overall classification

efficacy across the six demographic classes in the DemogPairs dataset, using standard metrics Accuracy, Precision, Recall, and F1-Score to quantify the added value of cross-domain feature fusion in demographic attribute recognition.

Table 3. Model Evaluation ViT and XGBoost

Features	Accuracy	Precision	Recall	F1-Score
ViT-Face Age	0.7894	0.7895	0.7894	0.789
ViT-VGGFace	0.8861	0.8863	0.8861	0.8859
ViT-VGGFace-Age	0.8907	0.8909	0.8907	0.8906

Table 3 summarises the performance of demographic classification across three Vision Transformer-based feature extraction schemes, all paired with XGBoost as the classifier. The configurations evaluated are: (1) features from ViT-Age (a model trained for age estimation), (2) features from ViT-Face (a model pretrained on the VGGFace identity dataset), and (3) fused features from both models. Evaluation metrics, including accuracy, precision, recall, and F1-score, are computed on the held-out test set to ensure an unbiased assessment.

A ViT-VGGFace-based model achieves 88.61% accuracy, reaffirming the dominant role of structural facial features, such as geometric proportions and the spatial configurations of facial components, in demographic classification. Notably, the fused model (ViT-VGGFace-Age) yields the highest performance across all metrics: 89.07% accuracy, 89.09% precision, 89.07% recall, and 89.06% F1-score. This consistent improvement of approximately 0.4–0.5 percentage points over the ViT-VGGFace baseline demonstrates that age-related information, though insufficient on its own (ViT-Age: 78.94% accuracy), provides meaningful predictive gain when integrated with structural features, provided the downstream classifier (here: XGBoost) can adaptively model feature dependencies and suppress redundancy.

2. Evaluation of XGBoost Hyperparameter Tuning

To ensure optimal model performance and robustness, a systematic hyperparameter tuning of XGBoost was conducted via 5-fold cross-validation on the training set. This subsection presents the configuration search space, the selected optimal parameters, and their impact on both overall accuracy and fairness-aware metrics, demonstrating how explicit regularisation and architectural constraints mitigate overfitting while preserving equitable performance across demographic subgroups.

Table 4. Fairness Analysis by Demographic Subgroups

Params	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Acc Mean	F1 Score Mean	Precision Mean	Recall Mean	Train Time Mean
gamma = 0.0										
max_depth = 3										
min_child_weight = 3										
tree_method = hist										
gamma = 0.0										
max_depth = 3										
min_child_weight = 3	0.9068	0.8999	0.8912	0.8964	0.8935	0.8976	0.8974	0.898	0.8976	3981.618
tree_method = approx										

Params	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Acc Mean	F1 Score Mean	Precision Mean	Recall Mean	Train Time Mean
gamma = 0.1 max_depth = 3 min_child_weight = 5 tree_method = approx	0.9022	0.8981	0.8918	0.8964	0.8964	0.897	0.8969	0.8973	0.897	2353.176
gamma = 0.3 max_depth = 3 min_child_weight = 1 tree_method = hist	0.9051	0.8976	0.8883	0.8947	0.8958	0.8963	0.8962	0.8967	0.8963	237.6889
...
gamma = 0.3 max_depth = 8 min_child_weight = 3 tree_method = approx	0.8912	0.8814	0.8738	0.875	0.89	0.8823	0.882	0.8829	0.8823	2649.573
gamma = 0.3 max_depth = 8 min_child_weight = 1 tree_method = hist	0.8924	0.8843	0.8773	0.8738	0.8837	0.8823	0.8819	0.8826	0.8823	365.9889
gamma = 0.3 max_depth = 8 min_child_weight = 3 tree_method = hist	0.8947	0.8767	0.8767	0.8762	0.8831	0.8815	0.8812	0.8819	0.8815	334.3973
gamma = 0.3 max_depth = 8 min_child_weight = 1 tree_method = approx	0.8883	0.8767	0.8704	0.8756	0.886	0.8794	0.8791	0.8799	0.8794	2678.01

As shown in Table 4, the configuration with $gamma = 0.0$, $max_depth = 3$, $min_child_weight = 3$, and $tree_method = hist$ achieves the best overall performance with accuracy up to 89.81%, F1-score 89.81% and training time: 239.16 seconds, the fastest among all configurations. This configuration not only delivers absolute performance but also high cross-fold stability, with a standard deviation in accuracy of only 0.46 points, significantly lower than alternatives. This indicates greater robustness to data partitioning, crucial when the dataset includes minority subgroups prone to sampling fluctuations.

A comparison between $tree_method = hist$ and $approx$ reveals a significant trade-off: (i) $hist$ (exact histogram-based method) is consistently faster (mean < 400 seconds) and more stable, and (ii) $approx$ (quantile sketch-based approximation) incurs substantially longer training times (up to ~4000 seconds) without meaningful performance gains and often reduces stability. Moreover, increasing max_depth from 3 to 8 yields no predictive advantage; instead, it degrades performance (F1-score drops by up to 1.6 points) and increases variance indicative of overfitting. Similarly, raising $gamma$ or min_child_weight beyond certain thresholds overly restricts tree growth, diminishing the model's capacity to capture subtle, discriminative patterns across demographic subgroups.

These findings support the final selection of $gamma = 0.0$, $max_depth = 3$, $min_child_weight = 3$, and $tree_method = hist$ as the optimal configuration, successfully balancing the bias–variance trade-off, minimising overfitting risk, and ensuring training efficiency

without compromising fairness or accuracy. This configuration was subsequently used to train the final model on the whole training set before evaluation on the held-out test set.

3. Fairness Analysis by Demographic Subgroups

To assess the Equity and robustness of the proposed MD-ViT framework, performance is evaluated not only in aggregate but also disaggregated across six intersectional demographic subgroups as defined in the DemogPairs dataset. This fine-grained analysis enables systematic detection of potential disparities. It provides empirical evidence on the model's adherence to group fairness, a critical requirement for ethical deployment in real-world applications.

Table 5. Fairness Analysis by Demographic Subgroups

Class	Accuracy	Precision	Recall	F1-Score
Black_Males	0.9694	0.9176	0.8972	0.9073
White_Females	0.9662	0.8889	0.9111	0.8999
Asian_Males	0.9579	0.8726	0.8750	0.8738
White_Males	0.9694	0.8910	0.9306	0.9103
Black_Females	0.9593	0.8931	0.8583	0.8754
Asian_Females	0.9593	0.8820	0.8722	0.8771

As shown in Table 5, the model achieves consistently high accuracy across all subgroups, ranging from 95.79% (Asian Males) to 96.94% (Black Males and White Males), with a maximum disparity of only 1.15 percentage points. This indicates the absence of systematic bias toward any particular subgroup in terms of overall accuracy. Moreover, metrics more sensitive to class imbalance, namely precision, recall, and F1-score, also exhibit relatively balanced distributions. The F1-score ranges from 87.38% (Asian Males) to 91.03% (White Males), with a standard deviation of ± 1.33 points. The most significant deviations occur in recall for Black Females (85.83%) and precision for White Females (88.89%); however, these differences remain within acceptable limits for low-to-moderate risk demographic applications.

Notably, no consistent disparity pattern emerges along race or gender axes alone. For instance, the Black subgroup demonstrates high performance for both males (F1-score: 90.73%) and females (F1-score: 87.54%). Similarly, the Asian subgroup shows stable performance, 87.38% (F1-score for males) and 87.71% (F1-score for females), slightly lower than others, possibly reflecting intrinsic phenotypic variation or relatively fewer training samples for this group.

This performance consistency across intersectional subgroups serves as a strong indicator that the system satisfies group fairness within reasonable bounds, thereby addressing critical ethical considerations in the deployment of demographic inference technologies, particularly in sensitive domains such as healthcare and public services.

DISCUSSION

The fusion of features from two distinct task domains, face identification (ViT-VGGFace) and age estimation (ViT-Face Age), has been shown to improve model performance consistently. When paired with XGBoost, this fusion strategy increased accuracy from 88.61% (using ViT-VGGFace features alone) to 89.07%, accompanied by comparable improvements in precision, recall, and F1-score. These results indicate that age-related cues such as wrinkles, skin texture, and

age-induced facial shape changes, though insufficient on their own (accuracy: 78.94%), provide meaningful predictive gain when combined with structural facial features.

This improvement reflects a clear cross-domain synergy: structural features capture stable interpopulation differences, while age-related features encode dynamic phenotypic variations, some of which may further discriminate between groups. XGBoost, leveraging its built-in feature weighting and gain-based splitting mechanisms, adaptively prioritises informative features while suppressing redundant or noisy ones. Consequently, these findings support the hypothesis that multidimensional facial representations integrating identity and ageing cues are more comprehensive and robust for demographic attribute recognition, provided the fusion strategy is coupled with a classifier capable of flexibly modelling complex feature interactions.

CONCLUSION

Based on the experimental results using XGBoost as the sole classifier, it can be concluded that fusing features from two distinct task domains, face identification (ViT-VGGFace) and age estimation (ViT-Face Age), consistently enhances the performance of race and gender classification systems compared to single-domain feature usage. Specifically, the fused ViT-VGGFace-Age configuration achieves the highest performance, with 89.07% accuracy and 89.06% F1-score, outperforming ViT-VGGFace alone (88.61%) and ViT-Age alone (78.94%). These results directly address the research question, confirming that multidimensional facial representations integrating both structural and temporal aspects are indeed more comprehensive and practical for demographic attribute recognition, particularly when paired with a model capable of adaptively modelling feature interactions, such as XGBoost, via gain-based splitting and explicit regularisation.

Moreover, subgroup-level evaluation demonstrates that the system achieves a relatively high degree of fairness, with only a ~3.6-point gap in F1-score across demographic subgroups indicating adequate mitigation of intersectional bias.

In light of these findings, we recommend that developers of soft biometric systems, especially in healthcare and public-access applications, adopt cross-task feature fusion strategies based on Vision Transformers, coupled with ensemble tree-based classifiers such as XGBoost, as this combination has proven effective for simultaneously boosting accuracy and ensuring equitable performance across groups.

Furthermore, we urge policymakers and technology ethics practitioners to promote evaluation protocols that go beyond aggregate metrics: routine, fine-grained fairness audits across demographic subgroups should be institutionalised, and training datasets must be curated to include representative variations in pose, lighting, and background. Only then can demographic inference systems transition from laboratory accuracy to real-world reliability, ensuring they are not only technically sound but also inclusive, trustworthy, and socially responsible when deployed in diverse societal contexts.

REFERENCES

Al-Otaiby, N., & El-Alfy, E. S. M. (2023). Effects of Face Image Degradation on Recognition with Vision Transformers: Review and Case Study. *2023 3rd International Conference on*

- Computing and Information Technology, ICCIT 2023.*
<https://doi.org/10.1109/ICCIT58132.2023.10273970>
- Bonner, S. N., Thumma, J. R., Valbuena, V. S. M., Stewart, J. W., Combs, M., Lyu, D., Chang, A., Lin, J., & Wakeam, E. (2023). The intersection of race and ethnicity, gender, and primary diagnosis on lung transplantation outcomes. *Journal of Heart and Lung Transplantation*, 42(7). <https://doi.org/10.1016/j.healun.2023.02.1496>
- Bulat, A., Cheng, S., Yang, J., Garbett, A., Sanchez, E., & Tzimiropoulos, G. (2022). Pre-training Strategies and Datasets for Facial Representation Learning. *Lecture Notes in Computer Science*, 13673 LNCS. https://doi.org/10.1007/978-3-031-19778-9_7
- dima806. (2023). *dima806/face_age_image_detection · Hugging Face.*
https://huggingface.co/dima806/face_age_image_detection
- Ding, Y., Bu, F., Zhai, H., Hou, Z., & Wang, Y. (2024). Multi-feature fusion-based face forgery detection with local and global characteristics. *PLoS ONE*, 19(10 October). <https://doi.org/10.1371/journal.pone.0311720>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.*
<https://doi.org/10.48550/arxiv.2010.11929>
- Gao, W., Li, L., & Zhao, H. (2022). Facial Expression Recognition Method Based on SpResNet-ViT. *Proceedings - 2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science, ACCTCS 2022.* <https://doi.org/10.1109/ACCTCS53867.2022.00046>
- Greco, A., Percannella, G., Vento, M., & Vigilante, V. (2020). Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, 31(7), 67. <https://doi.org/10.1007/s00138-020-01123-z>
- Ha, F., John, A., & Zumwalt, M. (2021). Gender/sex, race/ethnicity, similarities/differences among SARS-CoV, MERS-CoV, and COVID-19 patients. *The Southwest Respiratory and Critical Care Chronicles*, 9(37). <https://doi.org/10.12746/swrccc.v9i37.795>
- Hupont, I., & Fernández, C. (2019). DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG, 2019)*, 1–7. <https://doi.org/10.1109/FG.2019.8756625>
- Iloanusi, O., Flynn, P. J., & Tinsley, P. (2022). Similarities in African Ethnic Faces from the Biometric Recognition Viewpoint. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2022.* <https://doi.org/10.1109/WACVW54805.2022.00048>
- Jatain, R., & Jailia, D. M. (2023). Automatic Human Face Detection and Recognition Based On Facial Features Using a Deep Learning Approach. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(2). <https://doi.org/10.17762/ijritcc.v11i2s.6146>
- Karkkainen, K., & Joo, J. (2021). FairFace: A face attribute dataset with balanced race, gender, and age for bias measurement and mitigation. *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021.* <https://doi.org/10.1109/WACV48630.2021.00159>
- Kotwal, K., & Marcel, S. (2024). Demographic Fairness Transformer for Bias Mitigation in Face Recognition. *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. <https://doi.org/10.1109/IJCB62174.2024.10744457>

- Nixon, S., Ruiiu, P., Cadoni, M., Lagorio, A., & Tistarelli, M. (2025). Assessing bias and computational efficiency in vision transformers using early exits. *Eurasip Journal on Image and Video Processing*, 2025(1). <https://doi.org/10.1186/s13640-024-00658-9>
- Pardede, J., & Kleb, S. S. (2024). Face Race Classification using ResNet-152 and DenseNet-121. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, 12(3), 798. <https://doi.org/10.26760/elkomika.v12i3.798>
- Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision Transformers for Dense Prediction. *Proceedings of the IEEE International Conference on Computer Vision*. <https://doi.org/10.1109/ICCV48922.2021.01196>
- Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y., & Timoner, S. (2020). Face recognition: Too biased, or not too biased? *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2020-June. <https://doi.org/10.1109/CVPRW50498.2020.00008>
- Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How We Have Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1). <https://doi.org/10.1145/3392866>
- Sehrawat, J. S., & Ali, M. (2023). Morpho-facial variations in physical features of two tribal populations of Kargil (Ladakh, India): A bio-anthropological investigation. *Anthropological Review*, 86(3). <https://doi.org/10.18778/1898-6773.86.3.01>
- Singh, S., & Chauhan, A. S. (2023). Attendance Compilation by Facial Recognition Methods of Image Processing: A Review. *International Journal for Research in Applied Science and Engineering Technology*, 11(5). <https://doi.org/10.22214/ijraset.2023.51708>
- skutaada. (2024). *skutaada/VIT-VGGFace at main*. <https://huggingface.co/skutaada/VIT-VGGFace/tree/main>
- Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., & Burton, J. (2025). A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clinical and Translational Science*, 18(3). <https://doi.org/10.1111/cts.70172>
- Xue, M., Duan, X., & Liu, W. (2019). Eliminating the other-race effect for multi-ethnic facial expression recognition. *Mathematical Foundations of Computing*, 2, 43–53. <https://doi.org/10.3934/mfc.2019004>