

Deep Learning-Based Sex Classification and Chronological Age-Group Estimation from Knee Radiographs Using a Multi-Task Framework with Rank-Fusion and Tri-Loss Learning

Mehmet Kurt

Kutahya Saglik Bilimleri Universitesi

Zekeriya Katılmış

zekeriya.katilmis@ksbu.edu.tr

Kutahya Saglik Bilimleri Universitesi

Selim Kestel

kutahya City Hospital

Doğukan Öcal

kutahya City Hospital

Research Article

Keywords: Knee Radiographs, Sex Classification, Chronological Age-Group estimation, Deep Learning, Forensic Anthropology, Knee Anatomy

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11155947/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The sex classification and chronological age-group estimation from the knee radiographs are a significant area of interest in forensic anthropology and clinical diagnostics. As the knee exhibits clear sexual dimorphism and predictable age-related morphological changes, knee radiographs provide a useful basis for demographic profiling. While most existing research focuses on pediatric maturation, relatively limited research has addressed automated, fine-grained frameworks for adult demographic assessment. This study proposes an automated, deep learning-based multi-task framework designed to extract demographic information from knee radiographs. The research is structured into two investigative tasks: Task 1 addresses sex classification through a Rank-Fusion Ensemble strategy, which combines the ranked predictions of multiple deep neural networks to identify sex-related morphological patterns. Task 2 focuses on chronological age-group estimation using a ConvNeXt-Tiny model. To capture the ordinal progression of chronological age groups, Task 2 utilizes a proposed Tri-Loss strategy that combines CORAL, Cross-Entropy, and Ordinal Regression loss functions. The results demonstrate promising performance for demographic estimation from knee radiographs. For sex classification, the Rank-Fusion Ensemble achieved an accuracy of 96.65% with an ROC-AUC of 0.9840. For chronological age-group estimation, the model reached a Top-1 accuracy of 71.29%. A detailed evaluation showed an ordinal MAE of 0.32 age-group intervals, suggesting that age-group predictions are largely aligned with adjacent categories, reflecting the ordinal and overlapping nature of age-related morphological variation in the knee. By utilizing knee radiographs as the primary imaging modality, the multi-task approach offers a quantitative alternative to traditional qualitative radiological observations in forensic and medical contexts.

1. Introduction

The accurate determination of an individual's biological profile, specifically biological sex and chronological age, is a fundamental pillar of forensic anthropology, legal medicine, and clinical diagnostics. Biological sex and chronological age-group estimation provide critical evidence in the identification of unknown remains and the assessment of age-related characteristics in living individuals. Among various skeletal sites, knee anatomy is considered a high-value anatomical backbone for demographic profiling. As a complex articulation involving the distal femur, proximal tibia, and the patella, the knee exhibits significant sexual dimorphism and predictable morphological changes throughout the human lifespan—ranging from epiphyseal fusion in adolescence to progressive degenerative alterations in late adulthood [1].

Traditional methods for demographic assessment primarily rely on manual scoring systems or standardized radiological atlases applied to knee radiographs, which are often labor-intensive and susceptible to intra- and inter-observer variability. These qualitative approaches frequently face challenges in maintaining consistency across different populations due to secular trends in growth and diverse ethnic backgrounds [2]. Furthermore, the subtle nature of skeletal aging within the knee anatomy, especially in the adult population, makes it difficult for human observers to differentiate between

adjacent age decades with high precision. Consequently, there is an urgent need for more objective, automated, and non-invasive assessment tools that can extract high-dimensional features from knee radiographs without being influenced by subjective interpretation [3].

In recent years, the rapid advancement of deep learning (DL) and convolutional neural networks (CNNs) has advanced the analysis of knee radiographs. Unlike traditional computer vision techniques, deep learning models can automatically learn hierarchical feature representations directly from raw radiographs, capturing intricate textural and structural features within the knee anatomy that may be imperceptible to the human eye. Modern architectures, such as EfficientNets and Vision Transformers, have demonstrated success in complex medical classification tasks, providing a potential alternative to manual skeletal assessments [4]. However, most current research in knee-based biometrics is restricted to adolescent cohorts for determining the legal threshold of majority, leaving a significant gap in the comprehensive classification of the full adult age spectrum using knee radiographs [5, 6].

To address these limitations, this research proposes an integrated multi-task framework for individual demographic estimation using knee radiographs as the primary imaging modality. While current literature is heavily focused on pediatric cohorts for determining the legal threshold of majority, there is a gap in the comprehensive classification of the full adult age spectrum (20–70 years). Traditional studies often treat chronological age-group estimation as a continuous regression problem; however, the clinical and forensic assessment of adults often requires categorical grouping to better align with physiological life stages. To address this, our study leverages modern ensemble strategies and specialized loss functions to enhance the predictive reliability of knee-based assessments across a broad adult age range. By utilizing high-resolution radiographic data, we aim to provide a scalable and automated solution that targets the unique features of knee anatomy for both sex and age-group identification across a broad adult age range (20–70 years).

In the present study, the demographic estimation problem is structured into two distinct investigative tasks centered on the information within knee radiographs. Task 1 focuses on binary sex classification, utilizing a proposed Rank-Fusion Ensemble strategy to identify sex-related morphological patterns in knee anatomy. Task 2 addresses chronological age-group estimation across five distinct decades using a specialized ConvNeXt-Tiny architecture. To handle the ordinal nature of aging within knee structures, Task 2 incorporates a novel Tri-Loss strategy consisting of Consistent Ordinal Regression (CORAL), Cross-Entropy, and Ordinal Regression losses. Both tasks were evaluated through a rigorous analytical protocol, including ROC analysis, confusion matrices, and calibration assessments, to ensure the forensic and clinical reliability of demographic predictions derived from knee radiographs. The main contributions of this study are summarized as follows:

- **Integrated Multi-Task Framework:** A comprehensive two-task deep learning framework is proposed for demographic estimation from knee radiographs, unifying biological sex classification and chronological age-group estimation.

- **Rank-Fusion for Sex Classification:** A Rank-Fusion Ensemble strategy is introduced by integrating EfficientNet-V2-S, Swin Transformer, and DenseNet-121, ensuring reliable identification of sex-related morphological patterns in knee anatomy.
- **Specialized Age-Group Estimation:** A ConvNeXt-Tiny-based architecture is developed for adult age-group estimation, a demographic cohort often overlooked in current radiological research. It utilizes a novel Tri-Loss strategy (CORAL, Cross-Entropy, and Ordinal Regression) to capture the sequential nature of biological aging.
- **Ordinal Consistency Analysis:** The study emphasizes the ordinal structure of age groups, demonstrating that predictive errors are primarily limited to adjacent categories, which is essential for ensuring forensic and clinical reliability.
- **Comprehensive Evaluation:** The proposed framework provides an extensive validation using independent test data, including accuracy, ROC-AUC, Top-K metrics, and ordinal error analysis, establishing its potential as an objective tool for demographic profiling.

2. Literature Review

Knee imaging has emerged as a critical focal point in forensic anthropology and clinical assessment for the determination of biological profiles. Traditionally, this process relied on manual scoring of epiphyseal fusion or the application of standardized radiological atlases; however, the recent integration of advanced imaging modalities and artificial intelligence has significantly enhanced the objectivity and precision of these assessments. The evolution from qualitative morphological observations to quantitative deep learning frameworks has allowed researchers to explore complex knee joint structures, such as the distal femur, proximal tibia, and patella, as informative morphological features for individual identification. The relevant literature studies addressing these methodologies are summarized below.

Dedouit et al. [7] developed an original magnetic resonance imaging (MRI) staging system to assess the epiphyseal fusion of the knee's growth plates in a population aged 10 to 30 years. Their study demonstrated that MRI is a reliable, non-invasive method for age assessment, showing that females generally achieve skeletal maturation earlier than males. The researchers emphasized that this radiation-free approach provides contemporary information on epiphyseal union, effectively overcoming the limitations of traditional skeletal collections.

Pyle and Hoerr [8] examined the reliability of age estimation using their established atlas in a modern Scottish population through knee radiographs of 442 individuals. Their analysis revealed a strong correlation between chronological and estimated age, particularly for females between 9 and 15 years, where the method achieved high accuracy. The study concluded that while the atlas remains a valuable tool, slight modifications are necessary to maintain precision in contemporary populations due to secular trends in growth.

Wan et al. [9] reviewed radiographs of 479 Chinese subjects aged 10 to 20 years to identify age based on epiphyseal union and sex based on patellar measurements. While the relationship between the three

knee epiphyses and chronological age was found to be strong, the patella demonstrated limited value as a sex indicator with a maximum accuracy of 73%. Their findings highlighted ethnic variations, noting that epiphyseal union occurred earlier in the Chinese population compared to previously published Western datasets.

Welson and El Basset [10] conducted a study on an Egyptian population aged 10 to 20 years using knee roentgenographic assessments for age and sex estimation. They found that the total Score of Knee Joint (SKJ) provided the highest correlation with age, whereas the femur bicondylar distance was a significant indicator for sexual dimorphism with an accuracy of 79.4%. In contrast, patellar parameters did not show statistically significant sexual differences in this specific adolescent age group, suggesting a need for older cohorts in sex estimation.

Dallora et al. [11] proposed a fully automated deep learning approach using MRI scans of the knee to assess the age of youth and young adults between 14 and 21 years. Utilizing a pre-trained GoogLeNet architecture, the study achieved a Mean Absolute Error (MAE) of 0.793 years for males and 0.988 years for females. Their research demonstrated that automated convolutional neural network (CNN) models could effectively classify individuals around the critical legal threshold of 18 years with high accuracy.

Auf der Mauer et al. [12] developed a three-step automated framework involving image pre-processing, bone segmentation, and age estimation based on 3D knee MRIs of male subjects. Their model, which combined CNNs with tree-based machine learning algorithms, reached a mean absolute error of 0.67 years in age regression and 90.9% accuracy in classification. The study underscored the potential of user-independent deep learning approaches to provide reliable forensic assessments without the need for manual intervention.

Daghighi et al. [13] investigated the fusion stages of the tibia and femur epiphyses using MRI in an Iranian population aged 15 to 40 years. Their results confirmed that MRI is a suitable modality for distinguishing different stages of maturation and that multiple epiphyses can be evaluated simultaneously in a single anatomical location. The study advocated for the use of knee MRI as a safe and practical alternative to CT and radiography, especially for legal purposes where ionizing radiation is discouraged.

Demircioğlu et al. [14] developed a deep neural network to estimate chronological age from 3,816 knee radiographs in a pediatric German population. The network was trained to predict age across independent and external validation cohorts, achieving a mean absolute error of approximately 0.86 years. This research highlighted the scalability of standard deep learning architectures in processing large retrospective datasets to separate age classes with high area under the curve (AUC) values.

Oura et al. [15] conducted a proof-of-concept study utilizing the Terry Anatomical Collection to estimate sex from radiographs of reconstructed cadaver knee joints. Their MhNet-based model reached an overall testing accuracy of 90.3%, correctly classifying all females in the dataset. This research provided early

evidence that artificial intelligence-based classification of the knee joint can serve as a robust tool for sex estimation in forensic osteology, even with small-to-medium scale datasets.

Wan et al. [16] compared the feasibility of MRI and radiograph-based age estimation in a Chinese Han nationality population of 1,018 volunteers. The study found that while both modalities showed a reasonable correlation with age, MRI was generally superior in regression models for males. Their results provided a foundation for subsequent machine learning analyses, emphasizing that MRI technology offers a more objective and radiation-free framework for skeletal maturation assessment.

Despite the significant contributions of these studies, current research is predominantly focused on the maturation phase of the knee joint, specifically targeting the transition from adolescence to early adulthood (ages 10–21) to determine the legal threshold of majority. Crucially, there is a distinct lack of research focusing on fine-grained, categorical chronological age-group estimation across the broader adult lifespan (20–70 years); the literature remains heavily saturated with regression-based models for pediatric growth assessment. Furthermore, existing literature often treats age and sex estimation as isolated tasks, frequently relying on manual scoring systems or single-architecture CNNs. There remains a significant need for an integrated framework that addresses demographic classification across a broader adult lifespan, from early adulthood to older adulthood (ages 20–70). This study addresses these existing limitations by proposing a multi-task approach that utilizes a modern Rank-Fusion Ensemble for sex classification and a specialized Tri-Loss-optimized ConvNeXt architecture for fine-grained chronological age-group estimation in a balanced adult dataset.

3. Methodology

This study proposes a deep learning-based framework for demographic classification, focusing on chronological age-group estimation and sex classification from knee radiographs. The proposed framework consists of two complementary tasks designed to extract demographic information from knee radiographs. The first task focuses on sex classification using multiple deep learning architectures integrated through a Rank-Fusion Ensemble strategy, whereas the second task addresses chronological age-group estimation using a ConvNeXt-Tiny architecture optimized with a Tri-Loss strategy to account for the ordinal nature of age progression. The overall methodology and experimental workflow of the proposed multi-task framework, including data preparation, task-specific model development, optimization strategies, and performance evaluation, are illustrated in Fig. 1.

3.1. Dataset and Characteristics

The dataset used in this study was retrospectively collected from the Kütahya Health Sciences University Evliya Çelebi Training and Research Hospital. It comprises a total of 1,042 digital knee radiographs. All images were assigned at the subject level, and no subject contributed images to more than one subset. The dataset was specifically curated to represent a balanced demographic distribution for both sex and chronological age-group classification tasks.

The demographic composition of the dataset is nearly perfectly balanced to ensure the deep learning models are not biased toward a specific demographic. Regarding biological sex, the dataset includes 519 male (49.81%) and 523 female (50.19%) subjects. The age of the participants ranges from 22 to 69 years, with a mean age of 45.0 ± 14.8 years and a median age of 45 years.

For the age-group estimation task (Task 2), the subjects were categorized into five chronological decades: 20–30, 30–40, 40–50, 50–60, and 60–70. As shown in Table 1, each age category contains approximately 20% of the total data, providing an anatomical spectrum that covers the physiological changes in the knee joint from early adulthood to the elderly stage. This diversity may support the models in learning various skeletal features, such as bone density variations, joint space narrowing, and the development of osteophytes across different life stages.

To evaluate the generalization performance of the proposed models, the dataset was split into training, validation, and independent test sets using a 70-10-20% stratification strategy. This ensures that the demographic ratios are preserved across all subsets. The detailed statistical and demographic distribution of the dataset is summarized in Table 1 and Table 2. Representative knee radiographs across different chronological age groups and sex categories are illustrated in Fig. 2.

Table 1
General demographic characteristics and statistical distribution of the dataset.

Total Subjects	Total Images	Gender % (Male / Female)	Age Range (Min–Max)	Mean Age ($\pm$ SD)	Median Age
1,042	1,042	49.8 / 50.2	22–69	45.0 ± 14.8	45

Table 2
Detailed distribution for Task 1 (Sex) and Task 2 (Age-Group).

Category	Group	Total	Train	Validation	Test	Percentage (%)
Sex	Male	519	363	52	104	49.81
	Female	523	366	52	105	50.19
Age Group	20–30	210	147	21	42	20.15
	30–40	209	146	21	42	20.06
	40–50	210	147	21	42	20.15
	50–60	206	144	20	42	19.77
	60–70	207	145	21	41	19.87
Total		1,042	729	104	209	100

3.2. Image Pre-processing and Augmentation

To enhance the image quality of the knee radiographs and ensure optimal feature extraction, a multi-stage pre-processing pipeline was implemented. Each radiograph was first converted to the RGB color space and subjected to Contrast Limited Adaptive Histogram Equalization (CLAHE) [17] to standardize the contrast across varying exposure levels, which is crucial for highlighting trabecular structures and joint boundaries. All images were then resized to a uniform resolution of 320×320 pixels using a strategy that preserves the original anatomical aspect ratio through padding. To improve the robustness of the models against intra-class variability, a comprehensive data augmentation suite was employed via the Albumentations library [18], including horizontal flipping, random rotations, and brightness-contrast adjustments. Finally, pixel values were normalized using the ImageNet mean and standard deviation to ensure compatibility with pre-trained weights and to facilitate faster convergence.

3.3. Deep Learning Architectures

The methodology of this study utilizes a diverse set of state-of-the-art architectures to capture both local textural patterns and global structural relationships within the knee anatomy. EfficientNet-V2-S [19] was selected for its optimized parameter efficiency and improved scaling, which allows for rapid and accurate identification of subtle sex-related morphological patterns. To capture long-range spatial dependencies, the Swin Transformer (Swin-T) [20] was integrated, utilizing its hierarchical shifted-window attention mechanism to analyze multi-scale anatomical features. Additionally, DenseNet-121 [21] was employed to leverage its dense connectivity patterns, which facilitate feature reuse and mitigate the vanishing gradient problem through continuous feature-map concatenation. Together, these models provide a comprehensive analytical framework, ranging from convolutional feature maps to hierarchical transformer-based representations of the knee joint.

3.4. Training Protocol and Optimization

All models were developed using the PyTorch framework [22] and initialized with ImageNet-1K pre-trained weights to leverage transfer learning capabilities. A differential learning rate strategy was implemented, where the backbone parameters were optimized with a lower rate while the classification head was assigned a higher rate to accelerate task-specific adaptation. The training process utilized the AdamW optimizer [23] and incorporated a Label Smoothing loss of 0.05 [24] to prevent overconfident predictions and improve model calibration. To ensure stable convergence, a Cosine Annealing scheduler with an initial warmup phase was utilized. Furthermore, Automatic Mixed Precision (AMP) was integrated into the training loop to enhance computational efficiency and reduce memory overhead without compromising the numerical precision of the gradients.

3.5. Rank-Based Fusion Ensemble

To enhance predictive performance and reduce the variance associated with individual architectures, a Rank-Based Fusion ensemble strategy was employed [25]. Unlike standard ensemble methods that rely on the simple averaging of softmax probabilities, this approach aims at reducing sensitivity to differences in probability calibration among the base models by focusing on the relative ordering of class predictions. For each sample, the predicted class probabilities generated by EfficientNet-V2-S,

Swin-T, and DenseNet-121 were ranked in descending order, with the highest probability receiving the highest normalized rank. The normalized rank was calculated as:

$$r_{im} = \frac{R_{im} - 1}{N - 1}$$

where R_{im} is the rank of sample i for model m , and N is the number of samples. The final fusion score was obtained by averaging the normalized ranks of the three models:

$$F_i = \frac{1}{M} \sum_{m=1}^M r_{im}$$

The final class prediction was then determined from the fusion score using the predefined decision threshold. By prioritizing ordinal ranking, this strategy provides a consensus of architectural perspectives while reducing sensitivity to differences in probability calibration among the base models [26].

3.6. ConvNeXt-Tiny and Tri-Loss Strategy for Chronological Age-group Estimation

For the chronological age-group estimation (Task 2), a specialized architectural and optimization framework was developed to address the ordinal nature of human aging. Unlike biological sex, age groups are sequential; therefore, misclassifying a subject into a neighboring age group is a less severe error than misclassifying them into a distant one. To capture the fine-grained osteological changes associated with aging, the ConvNeXt-Tiny architecture [27] was proposed. This model modernizes the standard convolutional framework by integrating design principles from vision transformers, such as large kernel sizes and inverted bottlenecks, providing a hierarchical representation of joint morphology.

To further focus the model on relevant anatomical features, a specialized Region-of-Interest (ROI) preprocessing step was implemented to extract the central 85% of the radiograph and isolate the tibiofemoral joint space from peripheral artifacts. The central 85% crop was selected empirically during preliminary experiments to reduce peripheral artifacts while retaining the knee joint region. This is followed by a localized CLAHE enhancement to emphasize age-related features such as subchondral sclerosis and osteophyte formation.

The classification head was modified using a Consistent Ordinal Regression (CORAL) framework [28], which replaces the standard softmax output with a series of shared-weight binary classifiers. This ensures that the predicted probabilities satisfy the ordinal constraint, reflecting the cumulative nature of aging. The model was optimized using a novel Tri-Loss function, which combines CORAL Loss (for rank consistency, weight: 0.5), Cross-Entropy Loss (for categorical discriminative power, weight: 0.4), and Ordinal Regression Loss (to minimize the absolute error between predicted and true age indices, weight: 0.1).

Training was conducted using a differential learning rate strategy, with the backbone set at 1×10^{-4} and the CORAL head at 5×10^{-4} . This Tri-Loss approach, combined with a 60-epoch Cosine Decay schedule and a 5-epoch warmup phase, allows the model to learn a continuous representation of the aging process, improving the Top-2 accuracy and reducing the Ordinal Mean Absolute Error (MAE) compared to standard classification baselines. For comparative evaluation, EfficientNet-V2-S and the Rank-Fusion Ensemble were also evaluated on the age-group classification task using the same training, validation, and test partitions.

4. Results And Discussion

This section presents and discusses the performance of the proposed deep learning framework for sex classification and chronological age-group estimation from knee radiographs. The results are evaluated using quantitative performance metrics and visual analyses to assess the effectiveness and generalization capability of the developed models. Performance was evaluated using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and Matthews Correlation Coefficient (MCC). The 95% confidence intervals for accuracy were calculated using the Wilson score method. Calibration was assessed using the Expected Calibration Error (ECE) with 10 equal-width confidence bins. For the five-class age-group task, class-wise ROC-AUC values were calculated using a one-vs-rest strategy, and macro-AUC was obtained by averaging the class-wise AUCs.

4.1. Task 1: Sex Classification Performance

The performance of the deep learning architectures and the proposed ensemble strategy for sex classification from knee radiographs was evaluated using a comprehensive suite of metrics, including accuracy with 95% confidence intervals (CI), precision, sensitivity, specificity, F1-score, ROC-AUC, and the Matthews Correlation Coefficient (MCC). The Rank-Fusion Ensemble model achieved the highest overall performance with an accuracy of 96.65% (95% CI: 93.78–98.56), outperforming all individual architectures. While individual models such as Swin-T and DenseNet-121 demonstrated high accuracy (both 96.17%), the Rank-Fusion strategy improved the MCC to 0.9331, indicating a favorable level of classification consistency. Notably, although EfficientNet-V2-S achieved a perfect sensitivity of 1.0000, it exhibited a lower specificity (0.8269) compared to the ensemble model (0.9712), which provided a more balanced and reliable classification capability across both sexes. The superior ROC-AUC of 0.9840 achieved by the Rank-Fusion Ensemble further confirms its robustness in discriminating between male and female knee morphologies. Detailed performance metrics for all models are presented in Table 3.

Table 3
Comparative performance metrics for sex classification.

Model	Acc. (%) [95% CI]	Prec. (%)	Sens. (%)	Spec. (%)	F1 (%)	ROC-AUC	MCC
EfficientNet-V2-S	91.39 [87.08–94.74]	85.37	100.0	82.69	92.11	0.9795	0.84
Swin-T	96.17 [93.30–98.56]	94.50	98.10	94.23	96.26	0.9783	0.92
DenseNet-121	96.17 [93.30–98.56]	95.33	97.14	95.19	96.23	0.9833	0.92
Rank-Fusion Ensemble	96.65 [93.78–98.56]	97.12	96.19	97.12	96.65	0.9840	0.93

4.2. Visual Analysis of Sex Classification Performance

The discriminative power of the evaluated architectures was further validated through Receiver Operating Characteristic (ROC) analysis. All models exhibited curves that clustered toward the top-left coordinate, indicating exceptional classification sensitivity and specificity. The Rank-Fusion Ensemble achieved a peak Area Under the Curve (AUC) of 0.9840, surpassing the individual performances of DenseNet-121 (0.9833), EfficientNet-V2-S (0.9795), and Swin-T (0.9783). This high AUC score demonstrates that the ensemble approach provides reliable classification performance across varying morphological characteristics in individual radiographs. The comparative ROC curves for the individual models and the proposed ensemble are illustrated in Fig. 3.

To investigate the error distribution of the best-performing model, a confusion matrix was generated for the Rank-Fusion Ensemble. The analysis reveals a highly balanced performance across both classes, with only 7 total misclassifications out of 209 test samples. Specifically, the model correctly identified 101 out of 104 male subjects (97.1%) and 101 out of 105 female subjects (96.2%). The minimal number of false positives and false negatives confirms that the proposed method effectively captures sex-related morphological patterns in knee anatomy with high precision and minimal sex-related bias. The confusion matrix for the Rank-Fusion Ensemble is shown in Fig. 4.

4.3. Task 2: Chronological Age-Group Estimation Performance

The estimation of chronological age groups from knee radiographs represents a more challenging task than sex classification due to the subtle and continuous nature of skeletal aging across five distinct decades. Performance metrics including Macro Precision, Recall, F1-score, Cohen’s Kappa, and Matthews Correlation Coefficient (MCC) were utilized to evaluate the models. The ConvNeXt-Tiny model, optimized with the Tri-Loss strategy, demonstrated improved performance, achieving the highest accuracy of 71.29%. This result notably outperforms the Rank-Fusion Ensemble (59.81%) and the

baseline EfficientNet-V2-S (42.58%). Furthermore, the Cohen’s Kappa of 0.6412 and MCC of 0.6439 achieved by ConvNeXt-Tiny indicate substantial agreement and a favorable level of classification consistency, respectively. These results suggest that the integration of ordinal regression (CORAL) and depthwise convolutional structures may help the model better capture progressive morphological patterns associated with aging. The comprehensive performance metrics for Task 2 are presented in Table 4.

Table 4
Performance metrics for chronological age-group estimation.

Model	Acc. (%)	Macro Prec.	Macro Recall	Macro F1	Weighted F1	Cohen’s Kappa	MCC
EfficientNet-V2-S	42.58	34.19	42.86	31.28	31.17	0.28	0.33
Swin-T	50.72	0.40	50.92	40.08	39.96	0.38	0.42
DenseNet-121	53.11	52.95	53.30	46.91	46.84	0.41	0.44
Rank-Fusion Ensemble	59.81	58.73	59.91	58.19	58.13	0.50	0.50
ConvNeXt-Tiny	71.29	71.04	71.38	70.58	70.51	0.64	0.64

4.4. Class-Wise Performance and One-vs-Rest (OvR) ROC analysis of ConvNeXt-Tiny

A detailed class-wise analysis of the proposed ConvNeXt-Tiny model was conducted to investigate its classification accuracy across different decades. The model exhibited exceptional performance in the youngest (20–30 years, F1-score: 83.33%) and oldest (60–70 years, F1-score: 85.06%) age groups. Specifically, the recall for the 20–30 age group reached 95.24%, indicating that the model is highly effective at identifying the radiographic features of early adulthood. In contrast, performance metrics showed a relative decrease in the middle-age categories (30–40 and 40–50 years), where radiographic features are often more subtle and transitional, leading to a higher degree of morphological overlap. However, the model regained high precision in the 50–60 group (81.25%), potentially reflecting age-related morphological patterns such as significant joint space narrowing or early osteophytic changes. These findings confirm that while mid-adult stages present inherent classification challenges, the proposed model maintains an overall weighted F1-score of 70.51%.

The discriminative capability of the ConvNeXt-Tiny model for each specific age category was further evaluated using One-vs-Rest (OvR) ROC analysis. The model achieved exceptional AUC values for the polar age groups, specifically 98.43% for the 20–30 cohort and 97.58% for the 60–70 cohort, indicating near-perfect separation of early-adult and elderly knee morphologies. While the middle-age groups (40–50) showed a relatively lower AUC of 82.02% due to the transitional nature of skeletal features in these decades, the overall Macro AUC of 92.00% indicates strong overall discriminative performance across the entire age spectrum. These results suggest that the model performs particularly well at identifying

physiological extremes while maintaining high global discriminative power. The detailed class-wise metrics, including precision, recall, and AUC, are summarized in Table 5, and the corresponding One-vs-Rest ROC curves are shown in Fig. 5.

Table 5
Detailed class-wise and overall performance for age-group estimation.

Age Group	Precision (%)	Recall (%)	F1-Score (%)	AUC	Support
20–30	74.07	95.24	83.33	98.43	42
30–40	61.54	57.14	59.26	89.16	42
40–50	57.89	52.38	55.00	82.02	42
50–60	81.25	61.90	70.27	92.79	42
60–70	80.43	90.24	85.06	97.58	41
Accuracy			71.29		209
Macro Average	71.04	71.38	70.58	92.00	209
Weighted Average	70.99	71.29	70.51	91.97	209

4.5. Confusion Matrix Analysis

The confusion matrix for the proposed ConvNeXt-Tiny model demonstrates high classification precision, particularly at the boundaries of the age spectrum. The model achieved high classification performance for the youngest (20–30: 40/42) and oldest (60–70: 37/41) cohorts. In the intermediate groups (30–40, 40–50, and 50–60), misclassifications were predominantly distributed among adjacent age categories, with almost no extreme errors (e.g., predicting a 20-year-old as 60). This concentration of errors around the diagonal highlights the model's ability to maintain ordinal consistency, reflecting the progressive and overlapping nature of skeletal maturation and joint degeneration in middle adulthood. The detailed distribution of true versus predicted classes is illustrated in Fig. 6.

4.6. Ordinal Age Error Analysis

To evaluate the potential utility in forensic and clinical demographic assessment of the models, an ordinal error analysis was performed to measure the distance between the predicted and true age categories. The ConvNeXt-Tiny model outperformed all other architectures, achieving the lowest Mean Absolute Error (MAE) of 0.32. A critical finding is that for ConvNeXt-Tiny, 96.65% of all predictions (Correct + ± 1 Group) fell within the correct or adjacent age categories. Furthermore, the proposed model completely eliminated extreme classification errors (± 3 and ± 4 groups), indicating limited occurrence of large ordinal errors in the test set. In contrast, baseline models like EfficientNet-V2-S exhibited a non-negligible rate of extreme errors. The detailed distribution of ordinal prediction errors is presented in Table 6.

Table 6
Ordinal error distribution and Mean Absolute Error (MAE).

Model	MAE	Correct (%)	± 1 Group (%)	± 2 Groups (%)	± 3 Groups (%)	± 4 Groups (%)
EfficientNet-V2-S	0.85	42.58	35.89	16.75	3.83	0.96
Swin-T	0.62	50.72	37.80	10.53	0.48	0.48
DenseNet-121	0.58	53.11	36.36	10.05	0.48	0.00
Rank-Fusion Ensemble	0.48	59.81	33.01	6.22	0.96	0.00
ConvNeXt-Tiny (Proposed)	0.32	71.29	25.36	3.35	0.00	0.00

4.7. Calibration and Top-K Accuracy Analysis

To assess the reliability of the model's confidence in its own predictions, a calibration and reliability analysis was performed. The ConvNeXt-Tiny model achieved an Expected Calibration Error (ECE) of 0.0630, indicating a high degree of alignment between the predicted probabilities and the actual classification accuracy. In the highest-confidence bin, where the model's average confidence reached 96.3%, the observed accuracy was 94.6%, demonstrating that the system provides well-calibrated probability estimates with limited overconfidence. The reliability diagram illustrating this calibration is shown in Fig. 7.

The Top-K accuracy analysis was further utilized to evaluate the model's potential decision-support utility. While the exact-match (Top-1) accuracy was 71.29%, the Top-2 accuracy substantially increased to 89.47%, and the Top-3 accuracy reached 97.13%. Furthermore, the Top-4 accuracy was recorded at 98.56%, with a cumulative Top-5 accuracy of 100.00%. These results are particularly relevant for chronological age-group estimation, as they indicate that in nearly 90% of cases, the true age group is identified within the model's first two ranked predictions. This high performance in ranked outputs suggests that the proposed framework effectively captures the morphological proximity of adjacent age decades.

5. Conclusion

This study examined the application of modern deep learning architectures and optimization strategies for demographic profiling using knee radiographs as the primary imaging modality. By partitioning the research into two investigative tasks, biological sex classification and chronological age-group estimation, a framework for the automated analysis of knee anatomy was evaluated for forensic and medical contexts.

For sex classification (Task 1), the Rank-Fusion Ensemble strategy, which incorporates EfficientNet-V2-S, Swin Transformer, and DenseNet-121, yielded an accuracy of 96.65% and an ROC-AUC of 0.9840. These

results indicate that the ensemble-based ranking approach can assist in identifying sex-related morphological patterns within the knee anatomy as visualized through knee radiographs with consistent classification performance.

Regarding the chronological age-group estimation (Task 2), a specialized ConvNeXt-Tiny architecture optimized with a Tri-Loss strategy was utilized to address the complexities of skeletal aging within knee joint structures. The model achieved a Top-1 accuracy of 71.29% and a Top-2 accuracy of 89.47%. The analysis recorded a Mean Absolute Error (MAE) of 0.32, and no extreme classification errors were observed within the test set, suggesting that age-group predictions are generally aligned with adjacent categories. The recorded Expected Calibration Error (ECE) of 0.0630 provides additional information regarding the reliability of probability estimates in knee-based chronological age-group estimation.

In conclusion, the integration of ordinal regression and multi-model fusion provides a structured approach to skeletal biometrics. These findings suggest that the automated analysis of knee anatomy via knee radiographs can serve as a supportive and objective tool for forensic anthropology and clinical demographic assessment. Future research may focus on more diverse populations and multicenter datasets and investigate the impact of degenerative joint diseases on the performance of models trained on knee radiographs to further refine the generalizability of these prediction systems.

Declarations

Acknowledgements

Author information

Mehmet Kurt: Conceptualization, supervision, validation, investigation, writing-review & editing. **Zekeriya Katılmış:** Conceptualization, methodology, formal analysis, software, writing-original draft, writing-review & editing. **Selim Kestel:** Data curation, resources. **Doğukan Öcal:** Data curation, visualization..

Authors and Affiliations

Mehmet Kurt

Kütahya Health Sciences University, Department of Medical Informatics, Kütahya Türkiye,

Zekeriya Katılmış

Kütahya Health Sciences University, Department of Computer Engineering, Kütahya Türkiye

Selim Kestel

Department of Orthopedics and Traumatology

Doğukan Öcal

Corresponding author

Correspondence to Zekeriya Katılmış.

Funding:

The authors received no financial support for the research, authorship, and/or publication of this article.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Data availability:

The data that support the findings of this study are available on request from the corresponding author, upon reasonable request and with the provision of a data sharing agreement.

Ethics declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Human Ethics and Consent to Participate:

This retrospective study was approved by the Ethics Committee of Kütahya Health Sciences University (Approval Number: 2026/04-40). All eligible pelvic radiographs obtained from patients evaluated at Kütahya Health Sciences University Evliya Çelebi Training and Research Hospital between 2016 and 2024 were retrospectively analyzed for sex classification and chronological age-group estimation using deep learning models. The study was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. The requirement for informed consent was waived due to the retrospective nature of the study. Clinical trial number: not applicable.

References

1. Spradley MK, Jantz RL (2011) Sex estimation in forensic anthropology: Skull versus postcranial elements. *J Forensic Sci* 56(2):289–296. <https://doi.org/10.1111/j.1556-4029.2010.01635.x>
2. Schmeling A, Dettmeyer R, Rudolf E, Vieth V, Geserick G (2016) Forensic age estimation: Methods, certainty, and the law. *Deutsches Ärzteblatt international* 113(4):44–50. <https://doi.org/10.3238/arztebl.2016.0044>

3. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML (2019) Deep learning in medical imaging and radiation therapy. *Med Phys* 46(1):e1–e36.
<https://doi.org/10.1002/mp.13264>
4. Li S, Liu B, Li S, Zhu X, Yan Y, Zhang D (2022) A deep learning-based computer-aided diagnosis method of X-ray images for bone age assessment. *Complex Intell Syst* 8(3):1929–1939.
<https://doi.org/10.1007/s40747-021-00376-z>
5. Gilsanz V, Ratib O (2005) *Hand bone age: A digital atlas of skeletal maturity*. Springer
6. Ritz-Timme S, Cattaneo C, Collins MJ, Waite ER, Schütz HW, Kaatsch HJ, Borrman HIM (2000) Age estimation: The state of the art in relation to the specific demands of forensic practise. *Int J Legal Med* 113(3):129–136. <https://doi.org/10.1007/s004140050283>
7. Dedouit F, Auriol J, Rousseau H, Rouge D, Crubezy E, Telmon N (2012) Age assessment by magnetic resonance imaging of the knee: A preliminary study. *Forensic Sci Int* 217(1–3):232–e1.
<https://doi.org/10.1016/j.forsciint.2011.11.013>
8. Hackman L, Black S (2013) Age estimation from radiographic images of the knee. *J Forensic Sci* 58(3):732–737. <https://doi.org/10.1111/1556-4029.12077>
9. Aly SM, Shrestha B, Hong DJ, Omran A, Wang W (2016) Identification of age and sex based on knee radiography. *Forensic Sci Int* 267:231–e1. <https://doi.org/10.1016/j.forsciint.2016.08.001>
10. Welson NN, Abd E, Basset AS (2019) Age and sex estimation by knee roentgenographic assessment: An Egyptian population study. *J Forensic Radiol Imaging* 18:4–10.
<https://doi.org/10.1016/j.jofri.2019.07.002>
11. Dallora AL, Berglund JS, Brogren M, Kvist O, Ruiz SD, Dübbel A, Anderberg P (2019) Age assessment of youth and young adults using magnetic resonance imaging of the knee: A deep learning approach. *JMIR Med Inf* 7(4):e16291. <https://doi.org/10.2196/16291>
12. Mauer MAD, Well EJ, Herrmann J, Groth M, Morlock MM, Maas R, Säring D (2021) Automated age estimation of young individuals based on 3D knee MRI using deep learning. *Int J Legal Med* 135(2):649–663. <https://doi.org/10.1007/s00414-020-02465-z>
13. Daghighi MH, Pourisa M, Javanpour-Heravi H, Ghojzadeh M, Mirza-Aghazadeh-Attari M, Daghighi S, Arasteh O, Zarrintan A (2021) Application of knee MRI in forensic age estimation: A retrospective cohort. *Radiography* 27(1):108–114. <https://doi.org/10.1016/j.radi.2020.06.019>
14. Demircioğlu A, Quinsten AS, Forsting M, Umutlu L, Nassenstein K (2022) Pediatric age estimation from radiographs of the knee using deep learning. *Eur Radiol* 32:4813–4822.
<https://doi.org/10.1007/s00330-022-08582-0>
15. Oura P, Junno J-A, Hunt D, Lehenkari P, Tuukkanen J, Maijanen H (2023) Deep learning in sex estimation from knee radiographs – A proof-of-concept study utilizing the Terry Anatomical Collection. *Leg Med* 61:102211. <https://doi.org/10.1016/j.legalmed.2023.102211>
16. Wan L, Hao H, Wang Y, Wang M, Song F, Shi Y, Xia W, Zhang Z (2024) The feasibility of age estimation based on MRI and radiographs of the knee joint in the Chinese population of Han

- nationality. *Australian J Forensic Sci* 56(3):242–252.
<https://doi.org/10.1080/00450618.2022.2143562>
17. Pizer SM (1977) Adaptive histogram equalization and its variations. *Comput Graphics Image Process* 6(9):184–195
 18. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA (2020) Albumentations: Fast and flexible image augmentations. *Information* 11(2):125.
<https://doi.org/10.3390/info11020125>
 19. Tan M, Le Q (2021) Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning* (pp. 10096–10106). PMLR
 20. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9992–10002). <https://doi.org/10.1109/ICCV48922.2021.00986>
 21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2261–2269). <https://doi.org/10.1109/CVPR.2017.243>
 22. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desfasiaux T, Kopf A, Yang E, DeVito Z, Raison M, Teuwen A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*, 32
 23. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. *arXiv preprint*.
<https://doi.org/10.48550/arXiv.1711.05101>. arXiv:1711.05101
 24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826)
 25. Kumar R, Vassilvitskii S (2010) Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 571–580).
<https://doi.org/10.1145/1772690.1772749>
 26. Dietterich TG (2000) Ensemble methods in machine learning. In *Multiple Classifier Systems* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1
 27. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S (2022) A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11966–11976).
<https://doi.org/10.1109/CVPR52688.2022.01167>
 28. Cao W, Mirjalili V, Raschka S (2020) Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognit Lett* 140:325–331.
<https://doi.org/10.1016/j.patrec.2020.11.008>

Figures

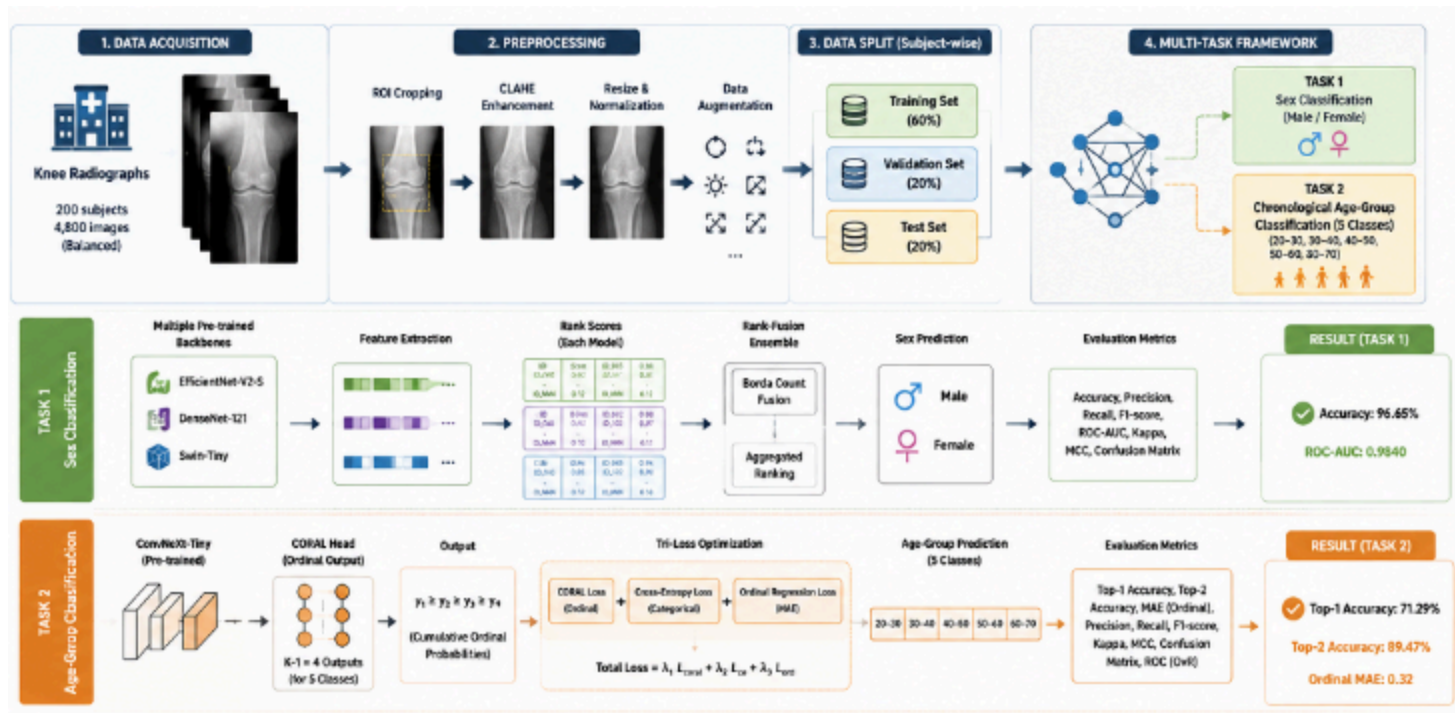


Figure 1

Overall workflow for chronological age-group estimation and sex classification from knee radiographs.



Figure 2

Representative knee radiographs across chronological age groups and sex categories.

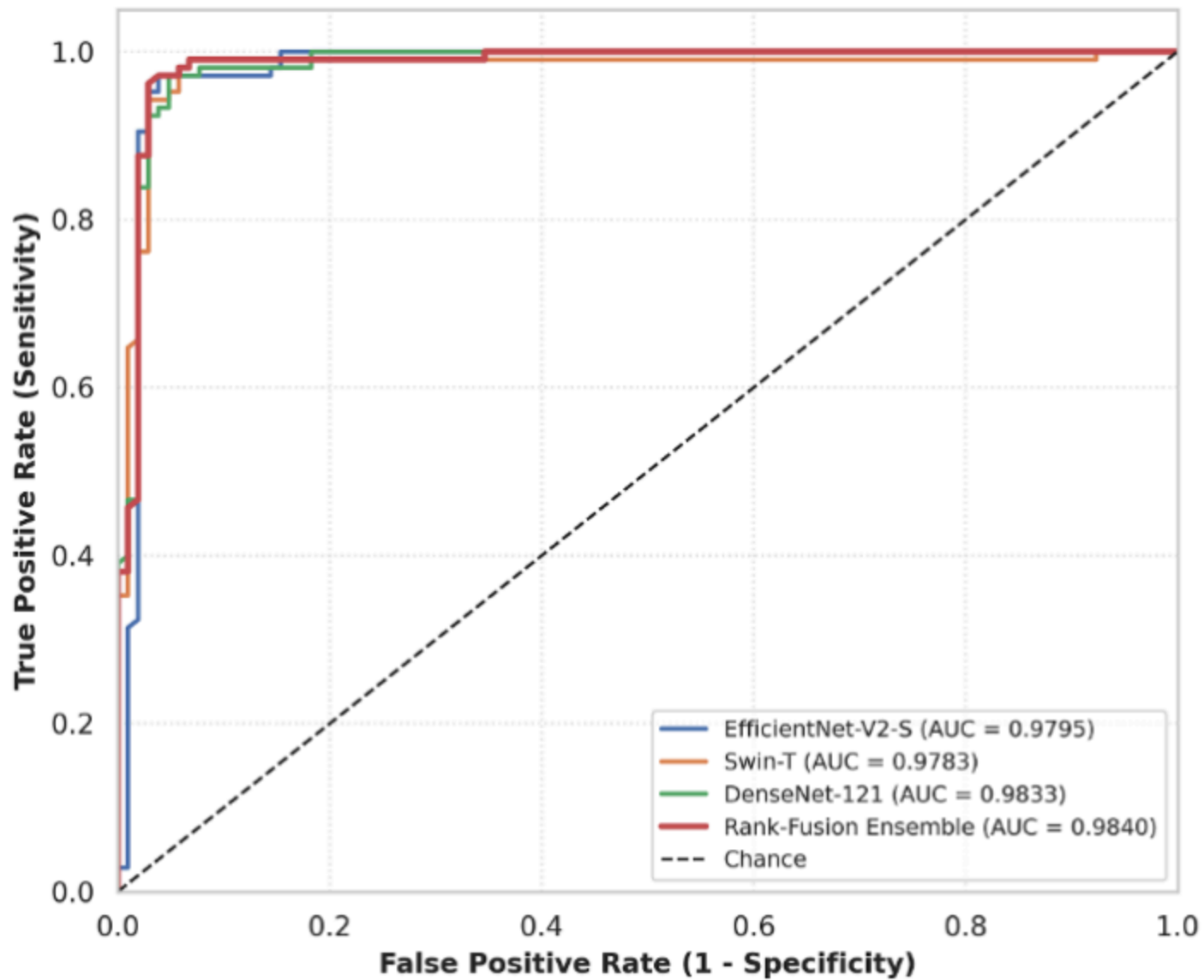


Figure 3

ROC curves for sex classification.

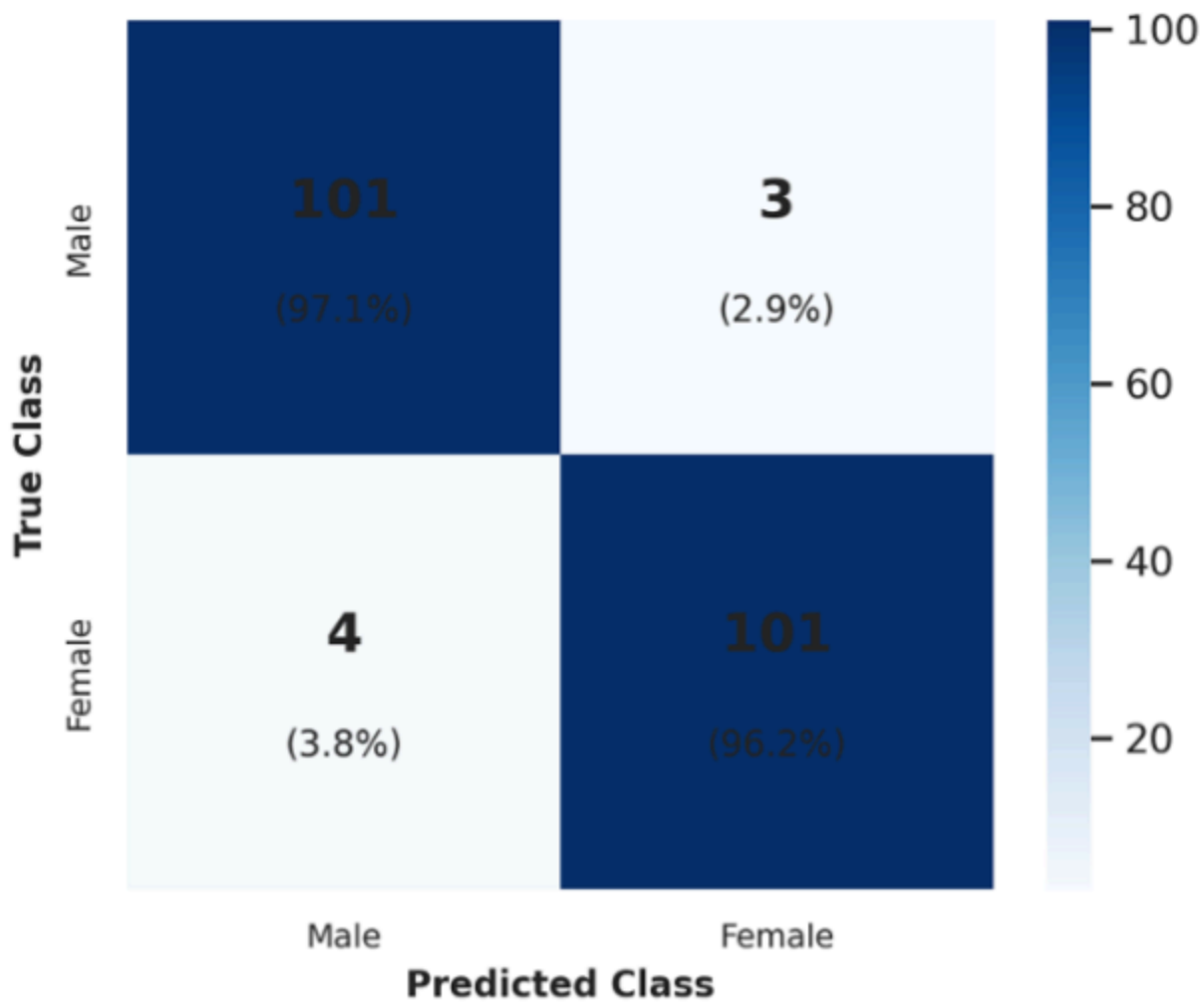


Figure 4

Confusion matrix for Rank-Fusion Ensemble.

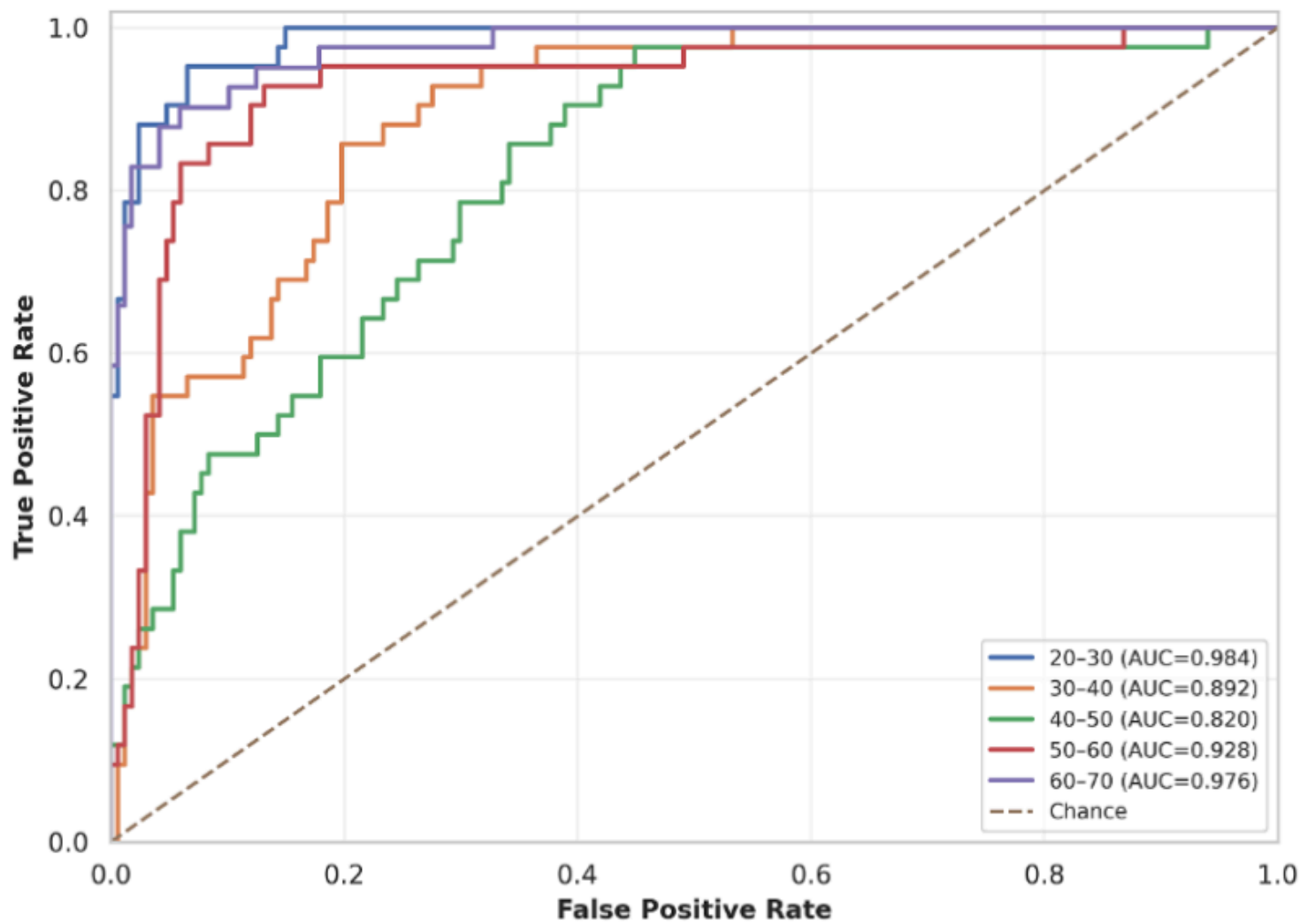


Figure 5

One-vs-Rest ROC curves for age-group estimation.

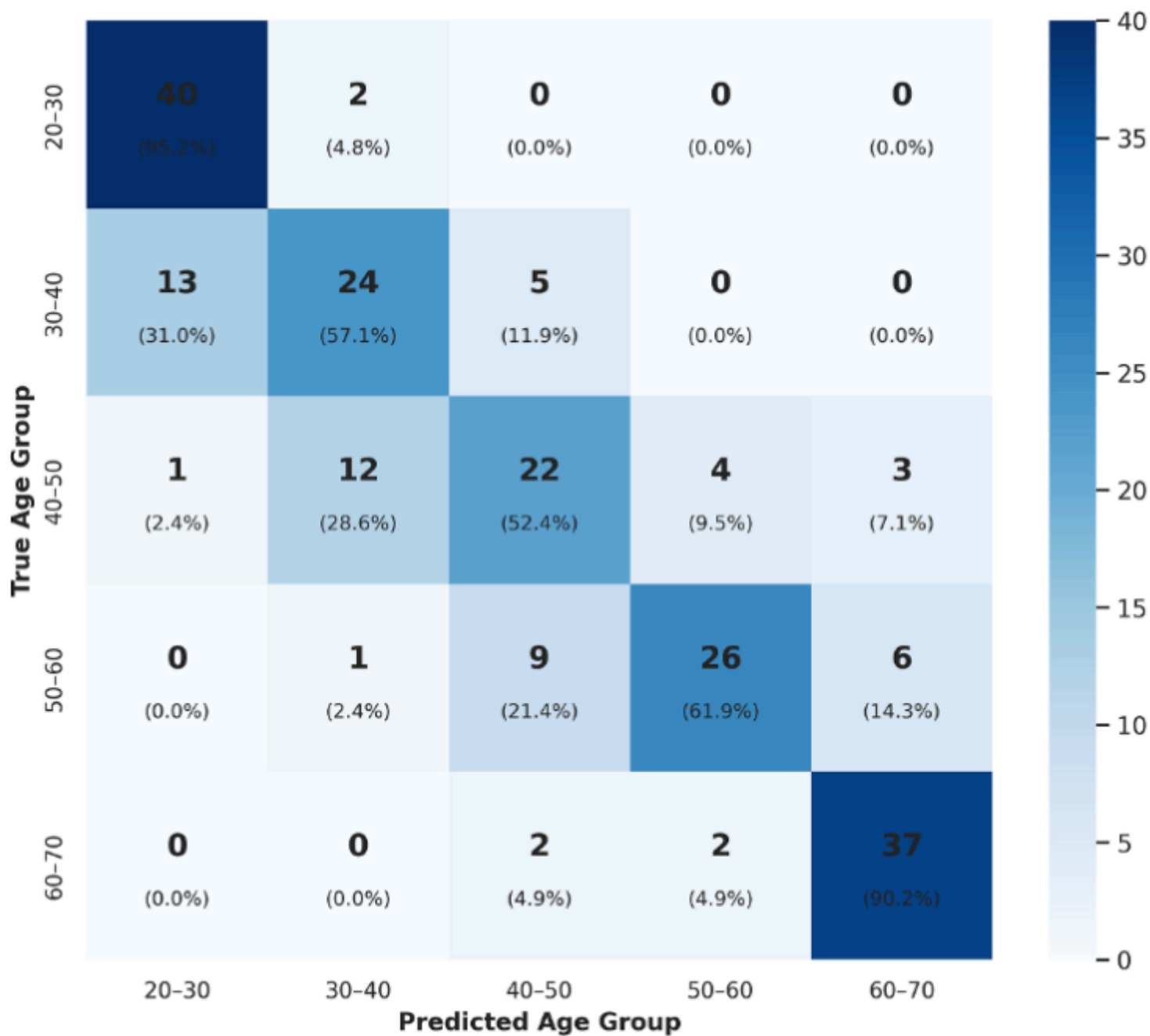


Figure 6

Confusion matrix for age-group estimation.

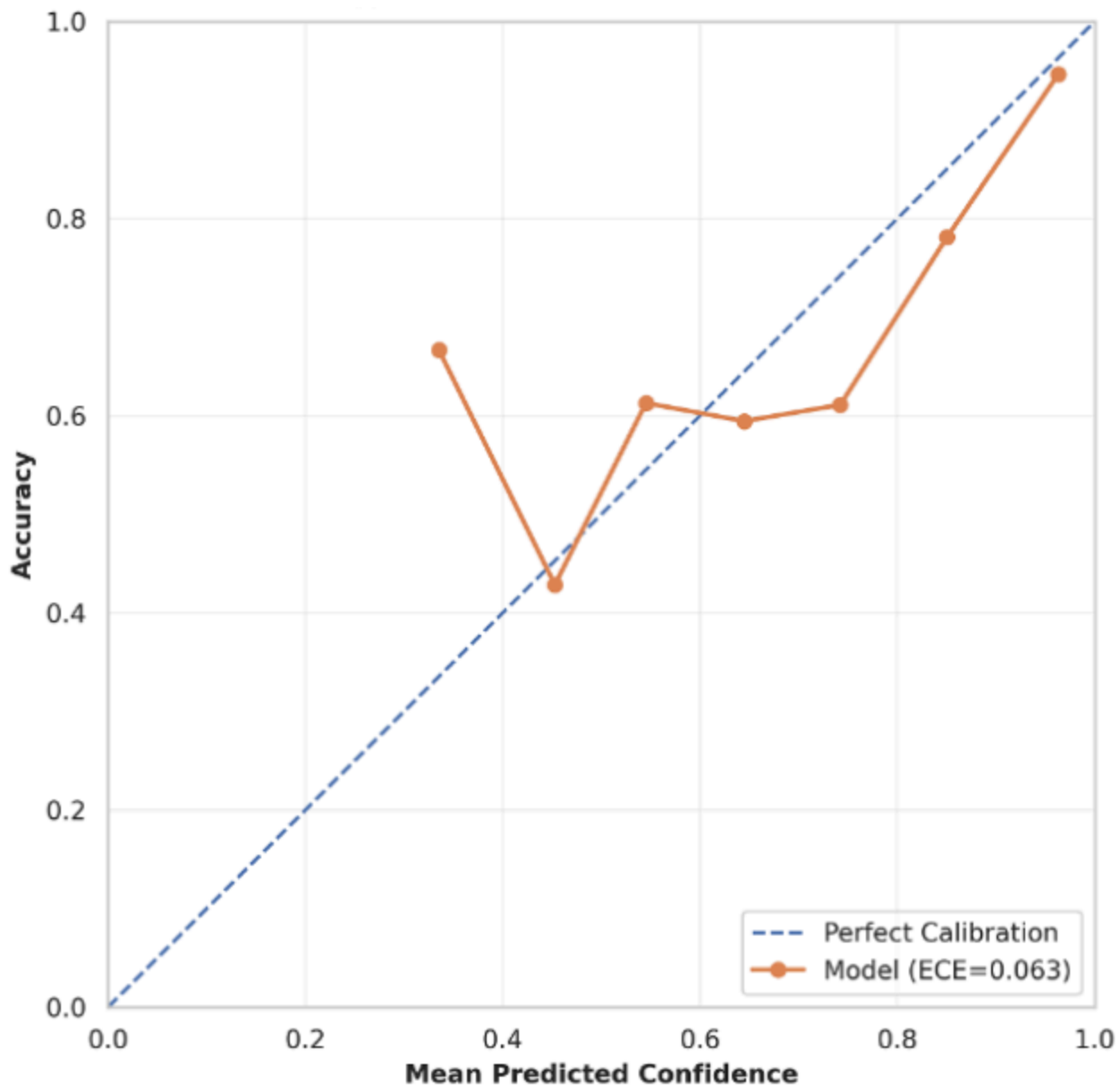


Figure 7

Reliability diagram and ECE for Task 2.