Estimating Age and Sex from Dental Panoramic Radiographs Using Neural Networks and Vision–Language Models

Salem Shamsul Alam [1] , Nabila Rashid [1], Tasfia Azrin Faiza [1], Saif Ahmed [1] , Rifat Ahmed Hassan [1] , James Dudley [2] and Taseef Hasan Farook [2,3,*]

1 Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh; saif.ahmed02@northsouth.edu (S.A.)
2 Adelaide Dental School, The University of Adelaide, Adelaide, SA 5000, Australia; james.dudley@adelaide.edu.au
3 Research and Innovations, Dental Loop Pty Ltd., Adelaide, SA 5000, Australia
* Correspondence: taseef.farook@adelaide.edu.au

Abstract: Purpose: The purpose of this study was to compare multiple deep learning models for estimating age and sex using dental panoramic radiographs and identify the most successful deep learning models for the specified tasks. **Methods:** The dataset of 437 panoramic radiographs was divided into training, validation, and testing sets. Random oversampling was used to balance the class distributions in the training data and address the class imbalance in sex and age. The models studied were neural network models (CNN, VGG16, VGG19, ResNet50, ResNet101, ResNet152, MobileNet, DenseNet121, DenseNet169) and vision–language models (Vision Transformer and Moondream2). Binary classification models were built for sex classification, while regression models were developed for age estimations. Sex classification was evaluated using precision, recall, F1 score, accuracy, area under the curve (AUC), and a confusion matrix. For age regression, performance was evaluated using mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), R^2, and mean absolute percentage error (MAPE). **Results:** In sex classification, neural networks achieved accuracies of 85% and an AUC of 0.85, while Moondream2 had much lower accuracy (49%) and AUC (0.48). DenseNet169 performed better than other models for age regression, with an R^2 of 0.57 and an MAE of 7.07. Among sex classes, the CNN model achieved the highest precision, recall, and F1 score for both males and females. Vision Transformers that specialised in identifying objects from images demonstrated weaker performance in dental panoramic radiographs, with an inference time of 4.5 s per image. **Conclusions:** The CNN and DenseNet169 were the most effective models for classifying sex and age regression, performing better than other models for estimating age and sex from dental panoramic radiographs.

Keywords: age regression; orthopantomograms; convolutional neural networks; vision–language model; random oversampling

1. Introduction

Age and sex estimation from medical imaging plays a critical role in forensic science, legal proceedings, and clinical dentistry. Methods for age estimation typically involve evaluating the stages of tooth development and eruption, assessing root formation, and examining secondary dentin deposition or pulp chamber reduction, especially in older individuals [1,2]. Sex determination often relies on analysing mandibular features, as males generally have larger, more angular mandibles, while females exhibit smaller, more

rounded jaws with more prominent gonial angles [3,4]. Traditionally, these processes have been subjective and heavily influenced by clinician or investigator expertise [5].

The advent of automated dental age and sex estimation has significantly reduced subjectivity, leveraging advanced technologies such as deep learning. Convolutional neural networks (CNNs) and statistical models have achieved substantial progress in this domain [1,2]. CNNs have demonstrated effectiveness in classifying dental features from radiographs, and significant advancements in detecting and segmenting dental structures [6,7]. More recently, vision–language models (VLMs), such as Vision Transformer and Moondream2, have shown promise in image captioning and segmentation [8]. However, their application to medical imaging classification tasks, such as predicting age and sex from dental radiographs, remains largely unexplored.

In addition to VLMs, other modern deep learning techniques have shown success in medical imaging. For example, Visual Geometric Group (VGG) models have been applied to detect pathologies in chest radiographs [9] and retinal imaging [10], while Residual Networks (ResNets) have been commonly used for brain tumour identification from magnetic resonance imaging scans [11]. Dense Convolutional Networks (DenseNets) have proven effective for multiclass classification in histological images [12]. Notably, ResNet and DenseNet models have been applied successfully to estimate age from dental radiographs [13–15].

While three-dimensional imaging methods, such as cone beam computed tomography (CBCT), enable volumetric analysis of teeth for age estimation, their use has been limited by higher radiation exposure and limited availability in clinical settings [15]. Consequently, two-dimensional panoramic radiographs remain the standard for full mouth evaluations.

Building on previous work where our research group developed a hybrid neural network model for estimating age and sex, the present study extends the foundational research, which generated accuracies between 67% and 85%, by evaluating various deep learning approaches for this purpose [16]. This area remains under-researched despite its clinical relevance. The objective was to identify the model that delivers the highest prediction accuracy while maintaining transparency in its decision-making process, thereby enhancing trust among practitioners. For this study, the following models were assessed: VGG16, VGG19, ResNet50, ResNet101, ResNet152, MobileNet, DenseNet121, DenseNet169, Vision Transformer, and Moondream2.

2. Material and Methods

2.1. Dataset Description

This study was approved by the University of Adelaide Human Research and Ethics Committee (HREC-2023-073) and the Institutional Review Board of North South University (2023/OR-NSU/IRB/0503). A Large Language Model (ChatGPT; OpenAI, San Francisco, CA, USA) was used to proofread and improve writing consistency during the revisions and editing of the manuscript draft.

The dataset consisted of 706 deidentified panoramic radiographic images taken within a single year at a specialist dental imaging centre. These images included dentulous and partially edentulous patients who sought general dental consultations or were scheduled for procedures such as multiple restorations, extractions, or third molar extractions. A dentist with six years of experience in clinical image analysis (THF) reviewed the dataset to exclude completely edentulous cases, patients with implants, those with odontogenic or non-odontogenic growths, images indicating spreading or chronic infections, and cases with radiologically visible trauma. Subsequently, the data science team (SSA, NR, TAF, SA, RAH) screened the remaining images for quality and suitability for machine learning applications, ultimately determining 437 images to be appropriate for use. The images

were classified into sex (male and female) and age based on the information retrieved from the system (Figure 1A,B).

(**A**)

(**B**)
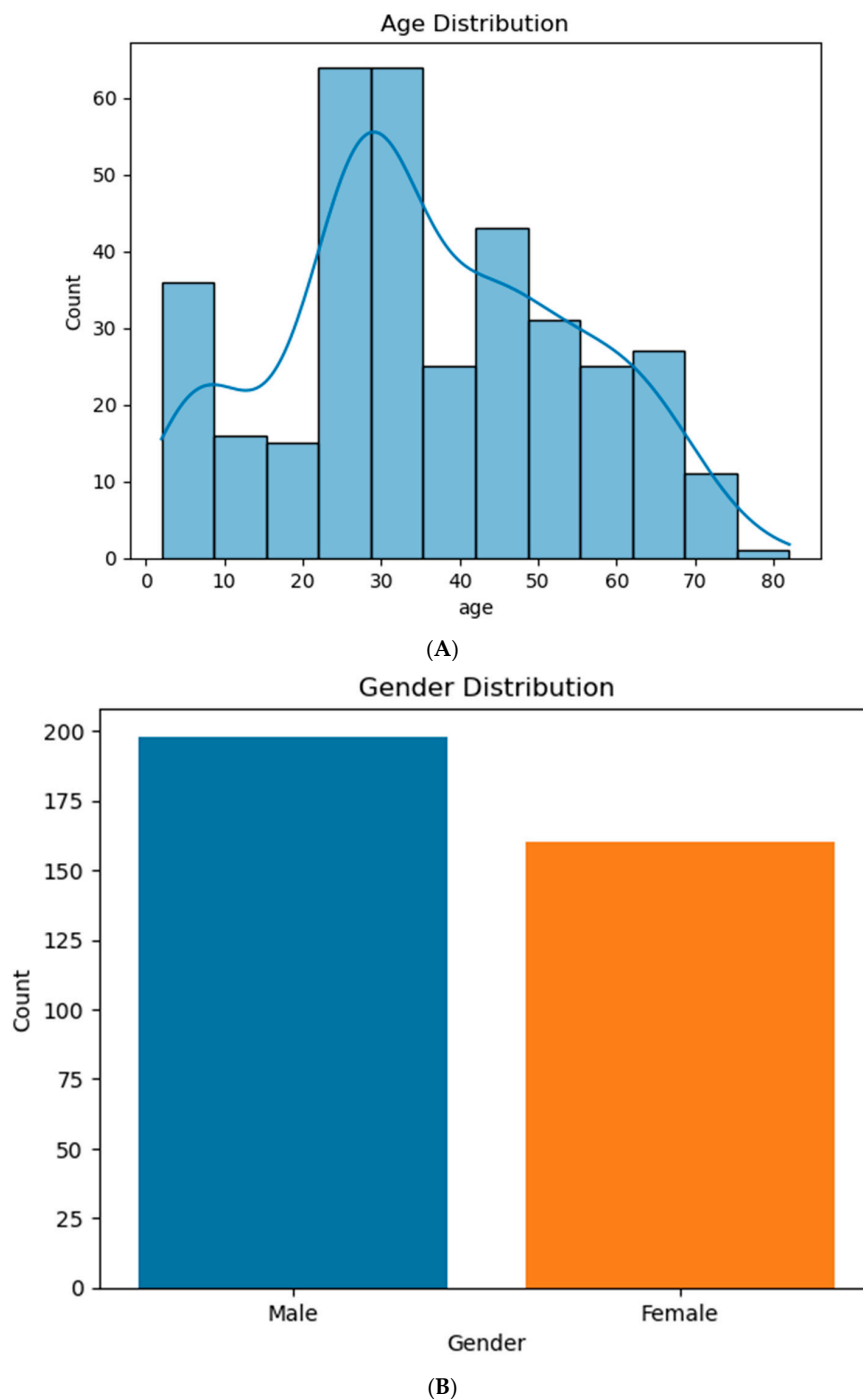
Figure 1. (**A**) Age distribution of the dataset; (**B**) sex distribution of the dataset.

The dataset was divided into training, validation, and testing sets. A total of 358 images were prepared for model training, of which 20% were reserved for model validation. The remaining 79 were used to test the performance of the models that were trained.

2.2. Relevant Literature

CNNs have shown promise in medical radiology due to their ability to identify minute variations in anatomic structures. For instance, Simonyan and Zisserman introduced VGG [17], a CNN architecture well suited for dental radiograph analysis. The effectiveness

of CNN architecture in image classification encouraged its inclusion in our study to assess its performance on sex classification tasks. He et al. developed ResNet, which overcame the vanishing gradient problem in deep networks, yielding improved results for age regression [18]. The success of ResNet in handling deep networks guided our choice of architectures for age regression in the current study. Howard et al. [19] contributed to MobileNet, which emphasised computational efficiency and adaptability for resource-constrained tasks such as age estimation systems capable of running on mobile devices in remote clinics. Inspired by its lightweight nature, we incorporated MobileNet into our study for sex classification, especially considering its efficiency in real-time applications.

Li et al. [20] applied deep learning for forensic age estimation using pelvic radiographs and demonstrated that CNNs can effectively identify age-specific skeletal features. This approach to age estimation from radiographs informed the adaptation of our models to handle dental radiographs for age prediction. Advanced architectures such as DenseNet employ a feature reuse mechanism that further refines age regression tasks [21]. Given DenseNet's success in handling complex image data, we adopted it to improve age regression performance in our study.

Oktay [22] and Kuo et al. [23] used CNNs in the detection and classification of teeth from radiographs, while Farhadian et al. [24] used CNNs to identify the pulp-to-tooth ratio from radiographs. Sironi et al. [25] discussed the role of Bayesian networks in assessing pulp chamber volume as a marker of age estimation from radiography [26]. While Bayesian networks were not directly used in our study, this approach helped inform our understanding of the potential for integrating advanced feature extraction techniques for better age and sex prediction.

Milošević et al. applied CNNs to estimate chronological age, indirectly using sex-related features automatically [27]. The Bayesian CNN applied in the study by De Back et al. [28] can be further extended with modifications to sex classification by incorporating those features of sex-related dental characteristics. The idea of modifying CNNs for sex-specific analysis inspired our implementation of sex classification as a separate task, optimising the model's ability to distinguish sex in dental images. The triplet network approach has lately been suggested to discriminate between age periods, which might help determine sex, which can be an area for further research [29]. Moreover, Tuzoff et al. have already successfully detected and numbered the dentition using CNNs, delivering the basic improvements toward sex-specific analysis relevant to dental radiographs [30].

Other emerging technologies also include VLMs. A hybrid VLM–transformer model was recently presented for dental age and sex classification [31]. This proof of concept demonstrated the adaptability of other VLMs, such as Moondream2, in enhancing the accuracy of the prediction. Chu et al. applied octuplet Siamese networks for osteoporosis analysis. Hence, it is possible to easily extend such architectures to sex prediction in dental imaging [32]. The concept of Siamese networks in medical imaging highlighted potential methods for improving the robustness of our models in distinguishing sex from dental radiographs.

Mualla et al. [33] evidenced the effectiveness of machine learning methods applied to radiographs for dental age estimation, which informed the strategies embraced in the current study for sex classification. Vila-Blanco et al. [34] used a deep neural network in chronological age estimation from panoramic radiographs, further informing the strategies appropriate to evaluate the interaction of age with sex in forensic analysis. We drew on their work to refine our approach to age prediction, ensuring that sex considerations were incorporated effectively into the model's structure. The use of deep transfer learning on panoramic radiographs for age estimation also guided the current study's strategy [35]. The transfer learning strategy outlined by Atas influenced our approach to leveraging

pre-trained models, helping to enhance the generalization of our models across diverse dental radiographs [35].

2.3. Data Preprocessing

Our study used random oversampling to balance the class imbalance in the dataset. This strategy entails randomly repeating samples from minority classes. By adjusting for class imbalance using augmented synthetic data, the model may effectively learn to predict sex and age without bias toward majority classes.

The radiographs were resized to grayscale images of 224×224 pixels and then converted to the RGB colour format for compatibility with deep learning models. The pixel intensity values were normalised to the [0, 1] range, which improved training efficiency and model performance.

2.4. Deep Learning Model Used

The current study examined the following models: Convolutional Neural Networks (CNNs), VGGNet [17], ResNet [18], MobileNet [19], DenseNet [21], and Vision Transformer (ViT) [36]. Due to their distinct designs and capabilities, each model has demonstrated great potential in medical imaging applications.

This study was conducted using a personal computer with the following specifications: Processor: Intel Core i5 (9th generation), RAM: 16 GB, Graphics Card: NVIDIA GTX 1650 Super (4 GB), Operating System: Windows 10.

The tests were conducted using Python on Jupyter Notebook, facilitated by Anaconda Navigator (Anaconda, Inc., Austin, TX, USA). It is there that the required packages were installed and set up: TensorFlow, Keras, NumPy, Matplotlib, sci-kit-learn, and many others necessary for this work. The use of these packages enabled the seamless execution of workflows and ensured their reproducibility. In this paper, the computing resources used are modest compared with high-performance computing systems, and careful optimisation of model training, batch size, and augmentation ensured that the experiments could be performed without significant computation bottlenecks.

2.4.1. Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are the standard architecture for image classification tasks [37]. CNNs extract hierarchical characteristics from images, beginning with low-level features like edges and progressing to more complicated patterns like textures and object components.

In this study, we used custom-built CNN architecture, starting with a $224 \times 224 \times 3$ input layer, corresponding to the reshaped grayscale radiographic images converted to RGB for the model input. The model started with a convolutional layer that uses 323×3 filters and a ReLU activation function, followed by max-pooling with a 2×2 pool size to minimise spatial dimensions. The second convolutional block employed 64 filters of comparable size with ReLU activation, followed by max-pooling. These layers concentrated on recognising fundamental features such as edges and simple textures. The model was then presented with two fully linked (dense) layers, the first with 128 units and the second with 64 units, both using ReLU activations. The final output layer was a fully connected layer with a single neuron for age prediction (regression) or two neurons for sex prediction (classification), depending on the task, and employed a linear activation function for age and sigmoid activation for sex classification. Figure 2 demonstrates the architecture of the neural network.
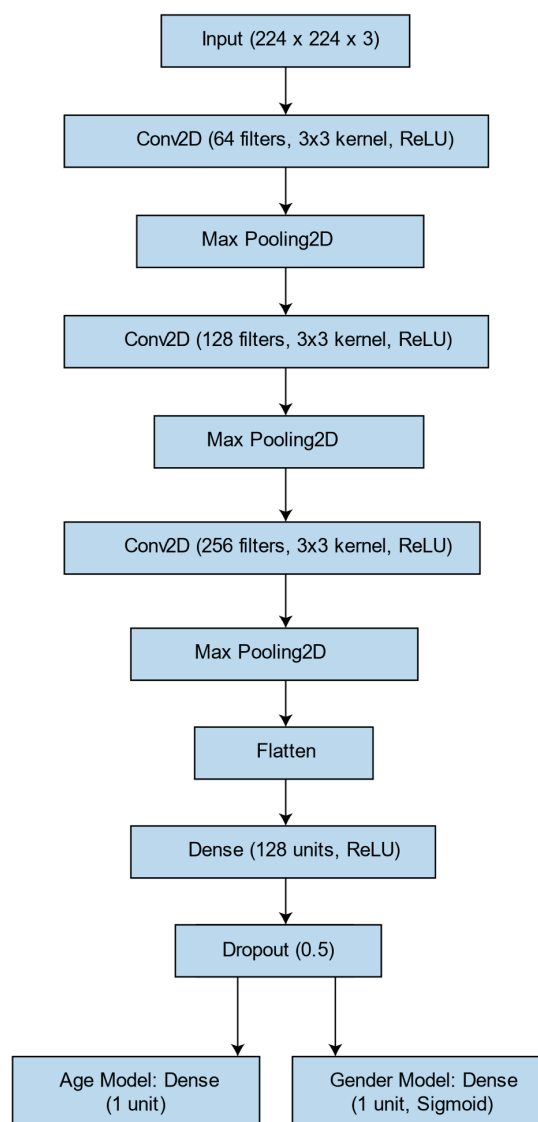
Figure 2. The constructed convolutional neural network.

2.4.2. VGGNet (Visual Geometry Group Network)

VGGNet is a deep convolutional network noted for its simplicity and consistent architecture. VGGNet has 19 layers, comprising 16 convolutional layers and three fully linked layers. VGGNet's main characteristic is its deep structure, in which convolutional layers use 3×3 filters with a stride of one, and the network grows in depth with pooling layers after every few convolutional layers [17].

For the current study, we used a pre-trained VGGNet model that was fine-tuned using the panoramic radiograph dataset. The convolutional layers extracted hierarchical characteristics, whereas deeper layers were formed to recognise more abstract elements such as bony landmarks. The final linked layers at the end of the network served as classifiers or regressors. The output layer for age prediction employed linear activation, whereas the output for sex classification employed a sigmoid activation function. Figure 3 demonstrates the architecture of the VGG Network.
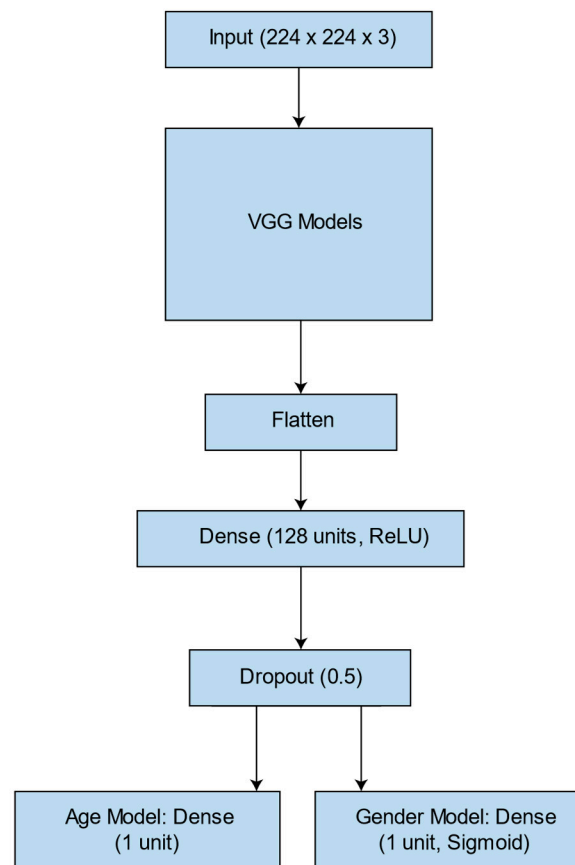
Figure 3. Visual Geometry Group Network.

2.4.3. ResNet (Residual Networks)

ResNet proposes the concept of "residual learning", which utilises identity shortcuts (skip connections) to avoid specific levels, which allows the model to learn more intricate details without degrading its learning outcomes [18].

The current study used different ResNet versions, including ResNet50, ResNet101, and ResNet152. These models have varying depths: ResNet50 has 50 layers, ResNet101 has 101, and ResNet152 has 152. All three networks were based on residual blocks, which are made up of several 3×3 convolutions with skip connections that add the input to the output of the block.

All ResNet variants employed in the current study used fully connected layers for the final output, with a linear activation function for age regression and a sigmoid activation for sex categorisation. Figure 4 demonstrates the architecture of the Residual Networks applied in the current study.
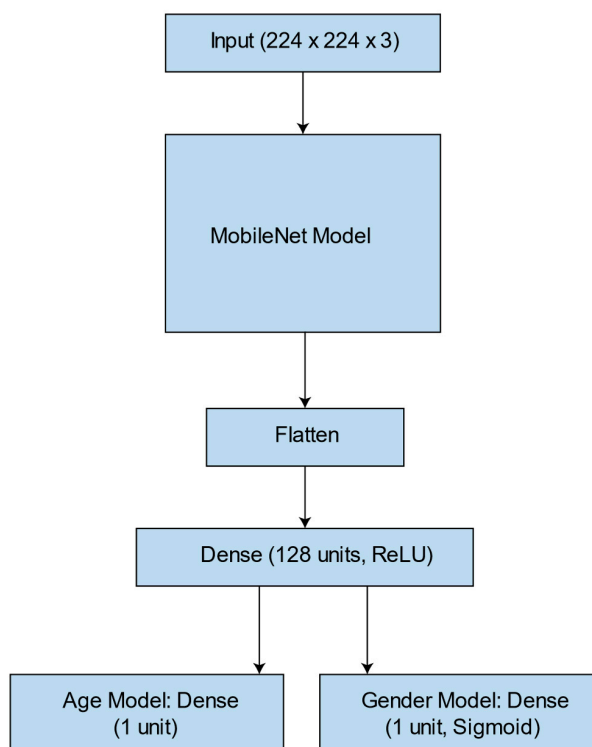
2.4.4. MobileNet

Howard et al. introduced MobileNet, which was designed to be computationally efficient, making it perfect for resource-constrained applications such as predicting age and sex on mobile devices. MobileNet uses depth-wise separable convolutions with fewer parameters and a lower computational cost than traditional convolutions. This enables MobileNet to perform competitively even on smaller, more resource-constrained devices [19].

We conducted our investigation using MobileNetV2, a lightweight version of the original model designed to improve performance on mobile devices. The model consists of an initial convolution layer, depth-wise separable convolutions with batch normalisation, and ReLU6 activations. These layers aid in extracting essential information from the raw

radiographic images, with the final layers acting as classifiers or regressors for age and sex prediction. Figure 5 shows the architecture of MobileNetV2.
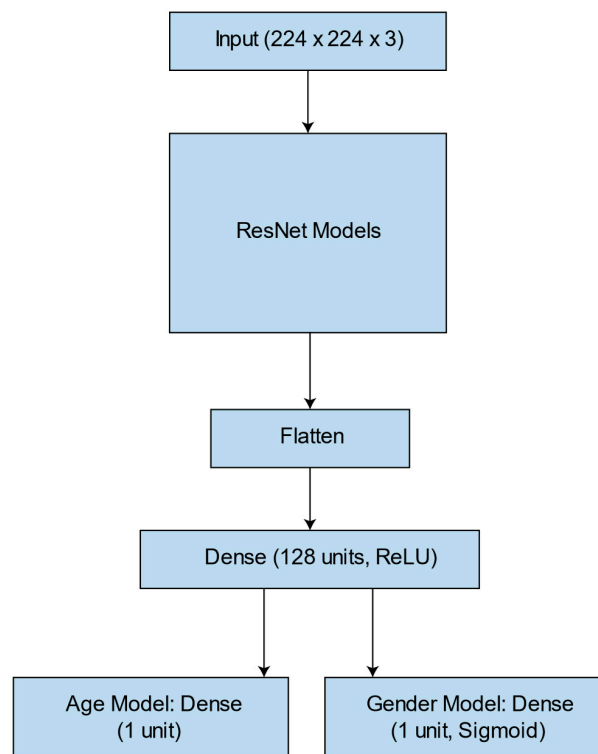
Figure 4. Residual Network architecture.

Figure 5. The MobileNetV2 architecture.

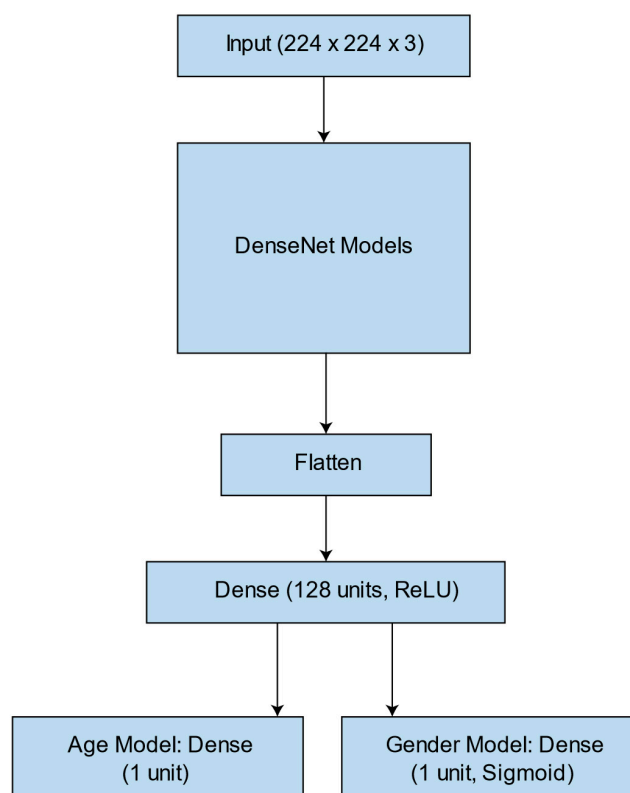2.4.5. DenseNet (Densely Connected Convolutional Networks)

DenseNet is a deep learning architecture that improves the classic convolutional network by adding dense connections across layers [21]. Unlike traditional CNNs, which

only receive input from the previous layer, DenseNet connects each layer to every other layer in a feed-forward manner. This design enhances feature propagation and reuse, resulting in more efficient learning and improved generalization.

The DenseNet models used here organised the layers into dense blocks, each receiving input from all preceding levels. DenseNet's essential feature is its growth rate, which determines how many features each layer can generate [21]. This extensive connection enables the model to learn more diversified and discriminative characteristics while requiring fewer parameters than traditional networks. DenseNet models are noted for their efficiency and ability to capture fine-grained image features.

For the current study, we employed DenseNet121 and DenseNet169, with the suffix numbers (121 and 169) annotating the number of layers present in the architecture. Figure 6 demonstrates the architecture of the DenseNet model.


Figure 6. The densely connected convolutional network (DenseNet) architecture.

2.4.6. Vision Transformer (ViT)

Vision Transformer (ViT) is a VLM that breaks an image into fixed patches of predetermined sizes, linearly embeds them, and uses self-attention mechanisms to acquire global context [36]. ViT's capacity to capture long-range dependencies in images offers it an advantage in tasks that involve understanding complicated interactions between various portions of the image.

In this study, we customized the ViT model for age and sex prediction tasks. The dental radiographic images were first divided into non-overlapping patches, then linearly embedded into a sequence, in a manner similar to the approach used in natural language processing. The self-attention mechanism in ViT allows the model to concentrate on essential regions of the image, which is critical for deciphering small details like bone structure and facial features in dental radiographs. The output layer employs a linear activation function for age regression and a sigmoid for sex categorisation. Figure 7 demonstrates the ViT architecture and workflow applied in the current study.
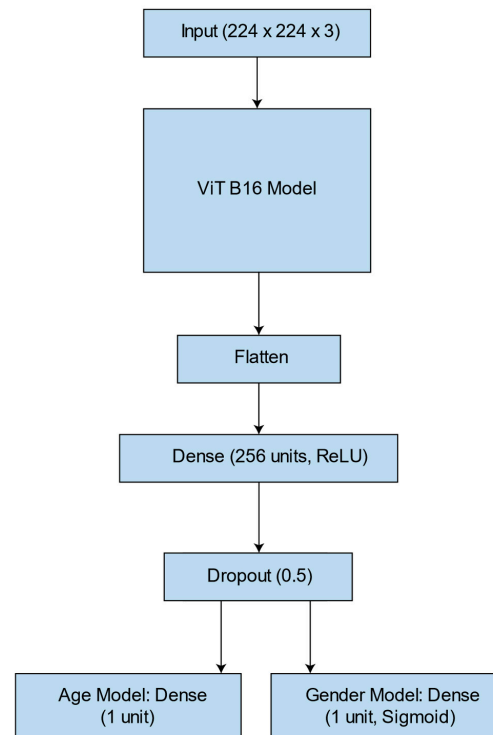
Figure 7. The Vision Transformer architecture used in the current study.

2.4.7. Moondream2

In this study, we selected Moondream2, a more miniature vision–language model (VLM), because of its smaller size and processing requirements compared with larger models such as Llama and PaliGemma, which include billions of parameters [38]. The smaller size of Moondream2 enables effective testing without sacrificing model accuracy, allowing us to run it within the limits of our available resources. A general outline of Moondream2 architecture is illustrated in Figure 8.
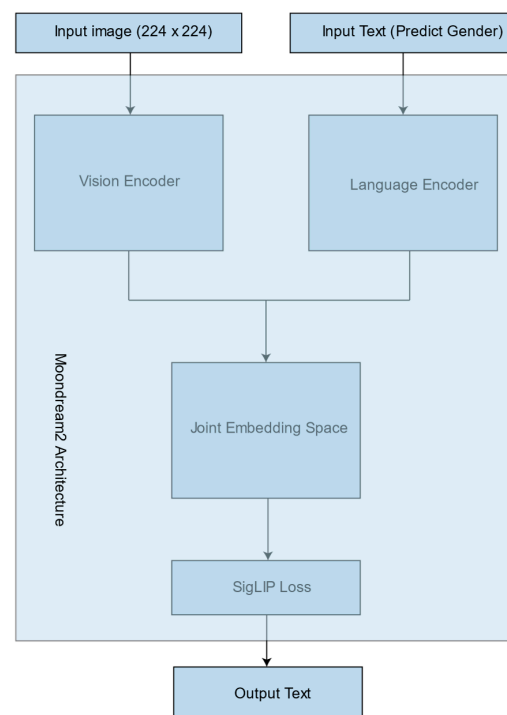
Figure 8. The Moondream architecture.

2.5. The Overall Workflow

The overall workflow for evaluating the included models is illustrated in Figure 9.
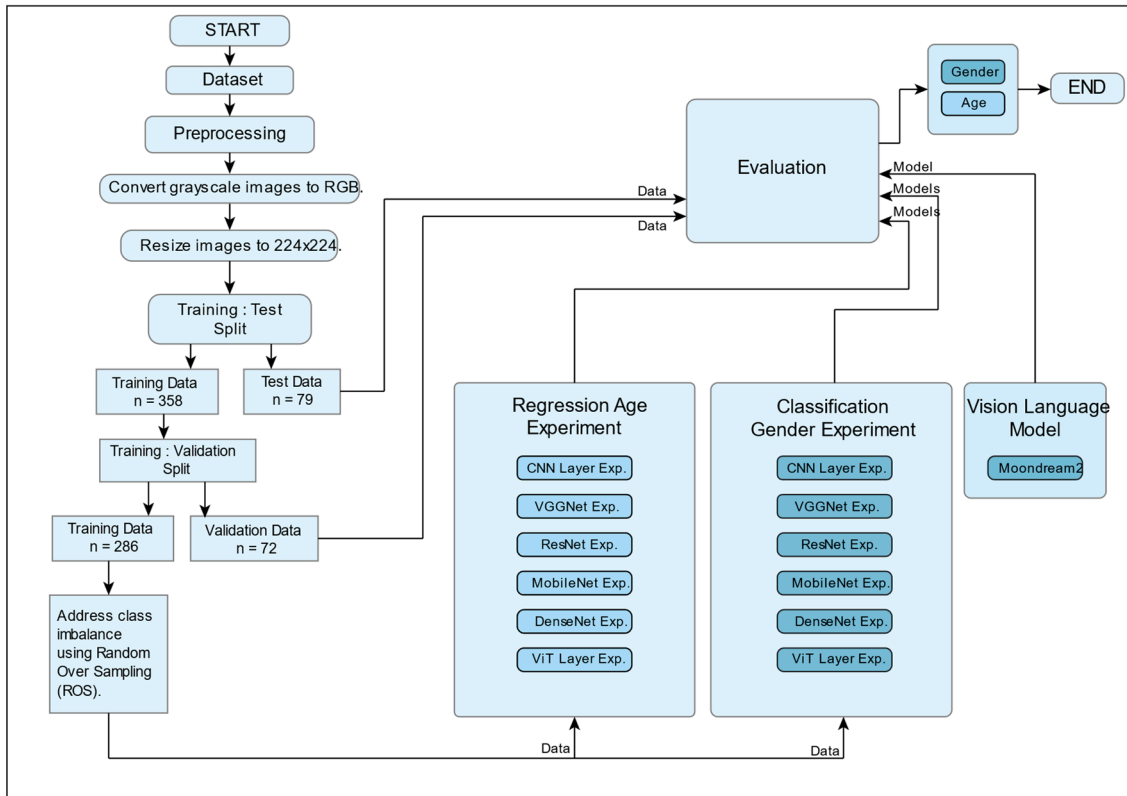
Figure 9. Deep learning workflow for age and sex estimation.

2.6. Evaluation Metrics

The following evaluation metrics were used [39,40].

2.6.1. Classification Evaluation Metrics

To evaluate the performance of the machine learning models in this study, various evaluation metrics were used to provide a complete picture of their effectiveness. These metrics included accuracy, precision, recall, F1 score, and ROC-AUC (Receiver Operating Characteristic—Area Under the Curve) score.

Accuracy measured the proportion of correctly classified instances out of the total number of instances. It is calculated as

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

Accuracy alone can be misleading, particularly when dealing with imbalanced datasets, as it does not adequately reflect the model's performance in distinguishing between classes with unequal sample sizes. To address this, precision was used, which calculates the proportion of correctly predicted positive cases out of all predicted positives. The formula for precision is

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The proportion of correctly identified positive observations within a specific class (sensitivity or recall) was evaluated next. This metric measures the effectiveness of each model in identifying all relevant instances within the class. The formula for recall is

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 score is the harmonic mean of precision and recall, offering a balanced metric that considers both factors. It is particularly useful for imbalanced datasets where one class is underrepresented, ensuring that performance is evaluated more comprehensively. The F1 score is calculated as

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC) was evaluated to compare the True Positive Rate (recall) with the False Positive Rate (1—Specificity) across various threshold settings. This metric assesses the model's ability to differentiate between classes at different thresholds, with a higher AUC indicating stronger overall performance in classification. The ROC-AUC score is computed as follows:

$$\text{ROC-AUC Score} = \int_0^1 \text{ROC Curve dFalse Positive Rate}$$

Here, dFalse Positive Rate is a differential element representing a small change in the False Positive Rate. In calculus, the differential dFalse Positive Rate is used to represent an infinitesimally small increment in the False Positive Rate. This integration process sums up the area under the ROC curve by accounting for these tiny changes in False Positive Rate, ultimately providing a measure of the model's performance in distinguishing between classes across various threshold values.

2.6.2. Regression Evaluation Metrics

In regression tasks, several evaluation metrics are commonly used to assess the performance of a model, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R^2 (coefficient of determination) [41].

The mean squared error (MSE) measures the average squared differences between the predicted and actual values. It is calculated by taking the average of the squared differences between the predicted ($\hat{y}_i$) and actual values (y_i) for each data point, as shown in the following formula:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Here, n is the number of samples in the dataset. MSE penalizes larger errors more significantly due to the squaring of differences, and a lower MSE indicates better model performance, with fewer large errors.

Root mean squared error (RMSE) is the square root of the MSE, providing an error metric in the same units as the target variable, making it more interpretable. The formula for RMSE is

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Like MSE, RMSE penalizes larger errors more heavily, but its square root transformation provides a value that is easier to interpret in the context of the original data.

Mean absolute error (MAE) measures the average of the absolute differences between the predicted and actual values. The formula for MAE is

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Unlike MSE and RMSE, MAE does not square the errors, making it less sensitive to large outliers. It provides a direct, interpretable measure of the average magnitude of the errors without exaggerating the impact of large discrepancies.

Mean absolute percentage error (MAPE) is used to measure the accuracy of predictions in terms of percentage error. It is calculated by averaging the absolute percentage errors between the predicted and actual values. The formula for MAPE is

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

In this formula, y_i represents the actual value, and $\hat{y}_i$ represents the predicted value. MAPE expresses the error as a percentage of the actual value, making it easier to interpret and compare across different datasets or scales. However, MAPE has a limitation when the actual values are close to zero, as it can lead to extremely large percentage errors.

The coefficient of determination $\left(R^2\right)$ indicates how well the model explains the variance in the target variable. It compares the model's performance to that of a simple baseline model that predicts the mean of the target variable for all instances. The formula for R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2}$$

In this formula, $\overline{y}_i$ is the mean of the actual values, and n is the total number of samples. R^2 measures the proportion of the variance in the dependent variable that is explained by the model. A higher R^2 value (closer to 1) indicates better model performance, whereas values closer to 0 indicate poor model fit. A negative R^2 suggests that the model performs worse than simply predicting the mean value of the target variable.

These regression metrics together provide a comprehensive view of how well the model is performing, with MSE, RMSE, and MAE indicating the magnitude of errors, MAPE showing the error as a percentage, and R^2 describing the model's explanatory power.

3. Results

In the current study, we performed separate experiments for classification and regression tasks. For sex classification, we applied deep learning models, namely CNN, VGG16, VGG19, ResNet (50, 101, 152), MobileNet, DenseNet (121, 169), Vision Transformer (ViT), and a vision–language model (Moondream2). We used the same models for age prediction, evaluating both tasks with various metrics.

3.1. Sex Classification (Male)

The CNN model achieved the highest precision, recall, and F1 scores (0.85, 0.85, and 0.85), indicating consistent performance in sex classification (Table 1). In contrast, the Moondream2 model had the lowest precision (0.51), suggesting that it made more false positive predictions but still performed well during recall (0.88). VGG16 and DenseNet121 also performed well, with balanced precision and recall, (0.83, 0.83) and (0.77, 0.9), respectively, showing effective sex classification. ResNet50 showed relatively lower performance, particularly in recall, indicating a higher rate of false negatives.

Table 1. Sex classification results for males.

Model	Precision	Recall	F1 Score
CNN	0.85	0.85	0.85
VGG16	0.83	0.83	0.83
VGG19	0.74	0.9	0.81
ResNet50	0.75	0.37	0.49
ResNet101	0.77	0.49	0.6
ResNet152	0.71	0.88	0.78
MobileNet	0.7	0.93	0.8
DenseNet121	0.77	0.9	0.83
DenseNet169	0.75	0.88	0.81
Vision Transformer	0.69	0.83	0.76
Moondream2	0.51	0.88	0.64

3.2. Sex Classification (Female)

The CNN model demonstrated a balanced performance, with a precision of 0.84, recall of 0.84, and F1 score of 0.84, indicating its ability to correctly classify female images with minimal errors (Table 2). The CNN performed better than models such as ResNet50 (precision: 0.56, F1 score: 0.68) and ResNet152 (precision: 0.82, F1 score: 0.70), which had higher precision but lower recall, reflecting some missed true positives. MobileNet and DenseNet121, with F1 scores of 0.70 and 0.78, respectively, also outperformed the CNN in some aspects. However, Moondream2 demonstrated poor results (precision: 0.38, F1 score: 0.13) again, indicating that the vision–language model struggled with classifying sex across panoramic radiographs.

Table 2. Sex Classification results for females.

Model	Precision	Recall	F1 Score
CNN	0.84	0.84	0.84
VGG16	0.82	0.82	0.82
VGG19	0.86	0.66	0.75
ResNet50	0.56	0.87	0.68
ResNet101	0.6	0.84	0.7
ResNet152	0.82	0.61	0.70
MobileNet	0.88	0.58	0.7
DenseNet121	0.87	0.71	0.78
DenseNet169	0.84	0.68	0.75
Vision Transformer	0.77	0.61	0.68
Moondream2	0.38	0.08	0.13

3.3. Overall Sex Classification Model Performance

Table 3 provides an overview of the sex classification performance across the various models. The custom CNN model stood out, with the highest accuracy (0.85) and AUC (0.85). DenseNet121 and VGG16 also performed well, achieving accuracies of 0.81 and 0.82, respectively, with AUC values of 0.84. ResNet152 demonstrated balanced performance, with 0.75 accuracy and 0.82 AUC across the sub-500 image dataset. ResNet50 and Moondream2, however, fell behind, with lower accuracy (0.61 and 0.49, respectively) and AUC scores (0.75 and 0.48), suggesting that they struggled with effective binary sex classification.

Table 3. Overall sex classification results.

Model	Accuracy	AUC
CNN	0.85	0.85
VGG16	0.82	0.84
VGG19	0.78	0.82
ResNet50	0.61	0.75
ResNet101	0.66	0.74
ResNet152	0.75	0.82
MobileNet	0.76	0.83
DenseNet121	0.81	0.84
DenseNet169	0.78	0.85
Vision Transformer	0.72	0.77
Moondream2	0.49	0.48

3.4. Overall Age Regression Model Performance

Table 4 presents the overall performances of different models in age regression, using various evaluation metrics. DenseNet169 and MobileNet provided the best results, with the lowest mean squared error (MSE) values of 85.83 and 95.46, respectively, and the highest R^2 values of 0.57 and 0.52, indicating strong predictive power. In terms of mean absolute error (MAE), DenseNet169 achieved the lowest value (7.07), closely followed by MobileNet (7.78), suggesting better accuracy in predicting age. The RMSE values for DenseNet169 and MobileNet were (9.26) and (9.77), respectively, highlighting their minimal average prediction errors. Vision Transformer, with an MSE of 159.98 and R^2 of 0.20, performed poorly compared to the other models. ResNet50 performed the worst, with the highest MSE (188.79) and MAE (11.59), and an R^2 of just 0.06, indicating poor predictive ability. VGG16, VGG19, and the CNN showed moderate performance, with MSE values ranging from 142 to 148, MAE values between 9.3 and 9.96, and R^2 values between 0.26 and 0.29, demonstrating average prediction accuracy.

Table 4. Overall age regression results.

| Model	MSE	MAE	MAPE	RMSE	R^2
CNN	142.88	9.3	32.37%	11.95	0.29
VGG16	148.19	9.96	36.52%	12.17	0.26
VGG19	142.13	9.57	33.83%	11.92	0.29
ResNet50	188.79	11.59	44.76%	13.74	0.06
ResNet101	170.68	10.62	38.16%	13.06	0.15
ResNet152	166.36	10.34	39%	12.9	0.17
MobileNet	95.46	7.78	26.07%	9.77	0.52
DenseNet121	97.2	7.95	27.31%	9.86	0.52
DenseNet169	85.83	7.07	22.98%	9.26	0.57
Vision Transformer	159.98	10.24	37.38%	12.65	0.2

3.5. Receiver Operating Characteristics of the Models

The AUC represents the ability of the model to discriminate between positive and negative classes, with higher values indicating superior performance [42].

The CNN demonstrated the highest AUC of 0.85, indicating excellent discriminatory power in distinguishing sex classes. VGG16 and DenseNet169 followed closely, achieving an AUC of 0.84. VGG19 documented an AUC of 0.82.

ResNet152, MobileNet, and DenseNet121 achieved AUCs of 0.82, 0.83, and 0.84 respectively. On the other hand, ResNet50 and ResNet101 achieved lower AUC values, of 0.75 and 0.74, respectively.

ViT recorded an AUC of 0.77, while Moondream2 recorded an AUC of 0.48, suggesting the selected VLMs were ineffective at discriminating sex classes. Figure 10A–K illustrate the ROC graphs for the included models. Please note that we have evaluated 10 deep learning models for age regression and 11 models for sex classification.
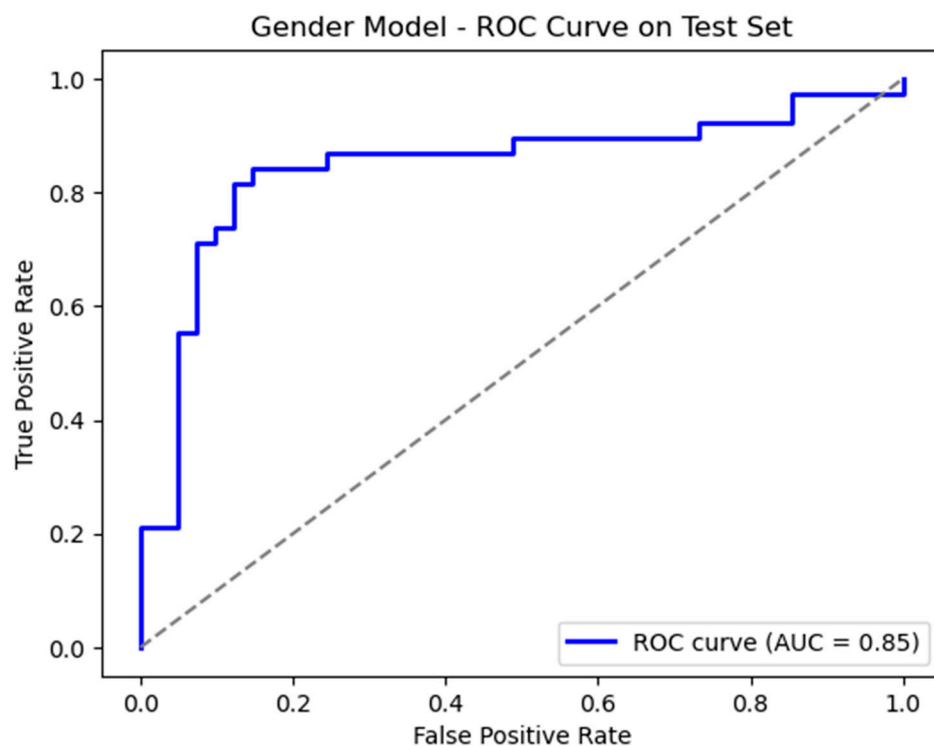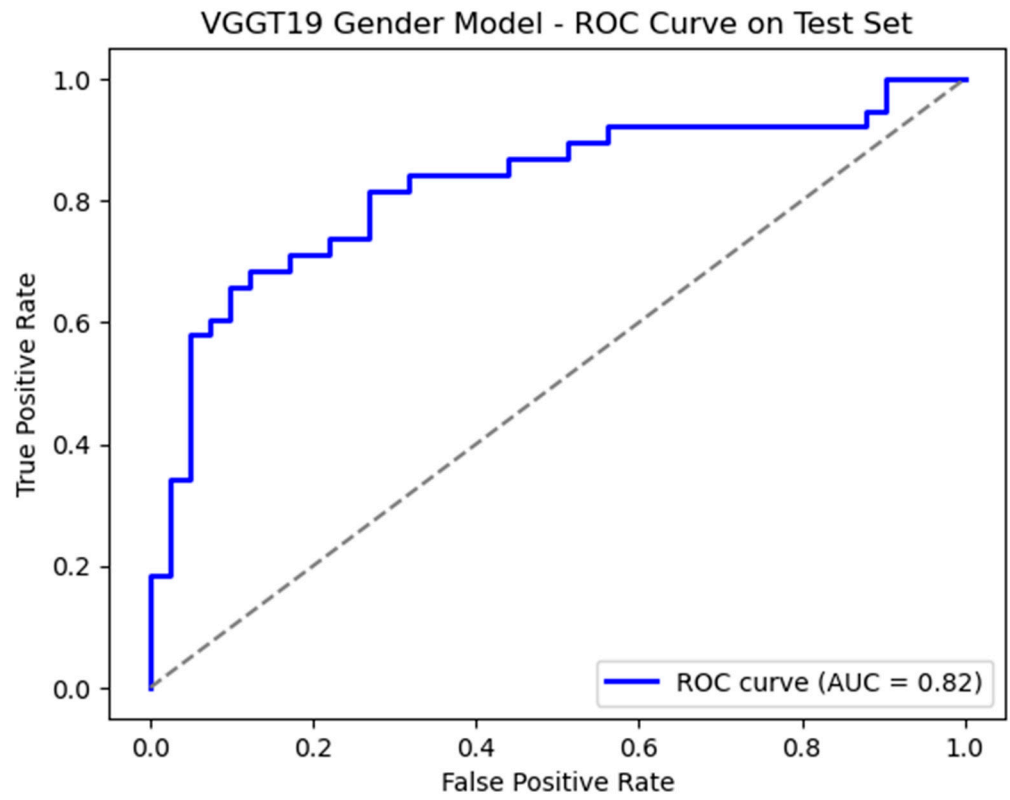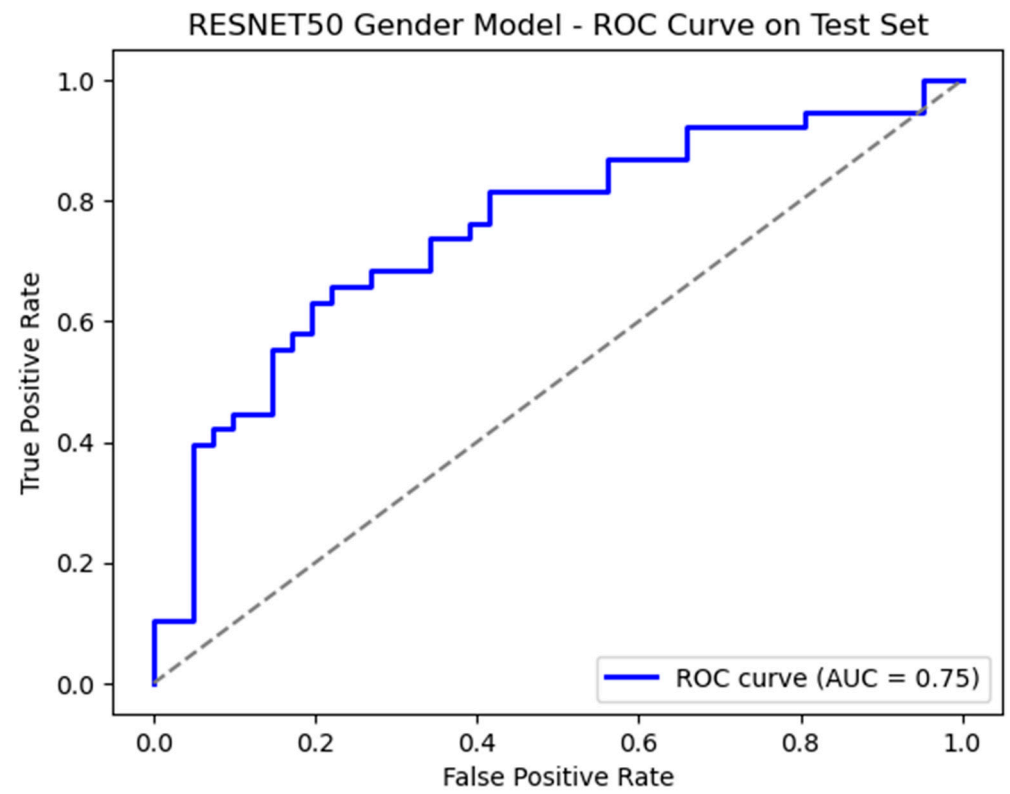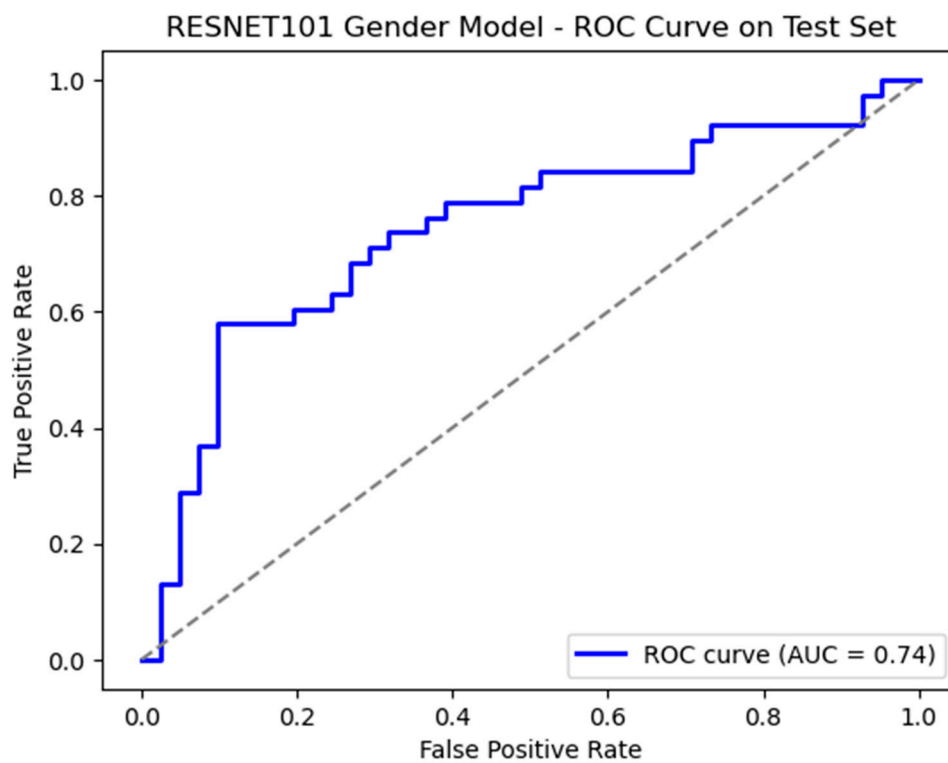

(**A**)
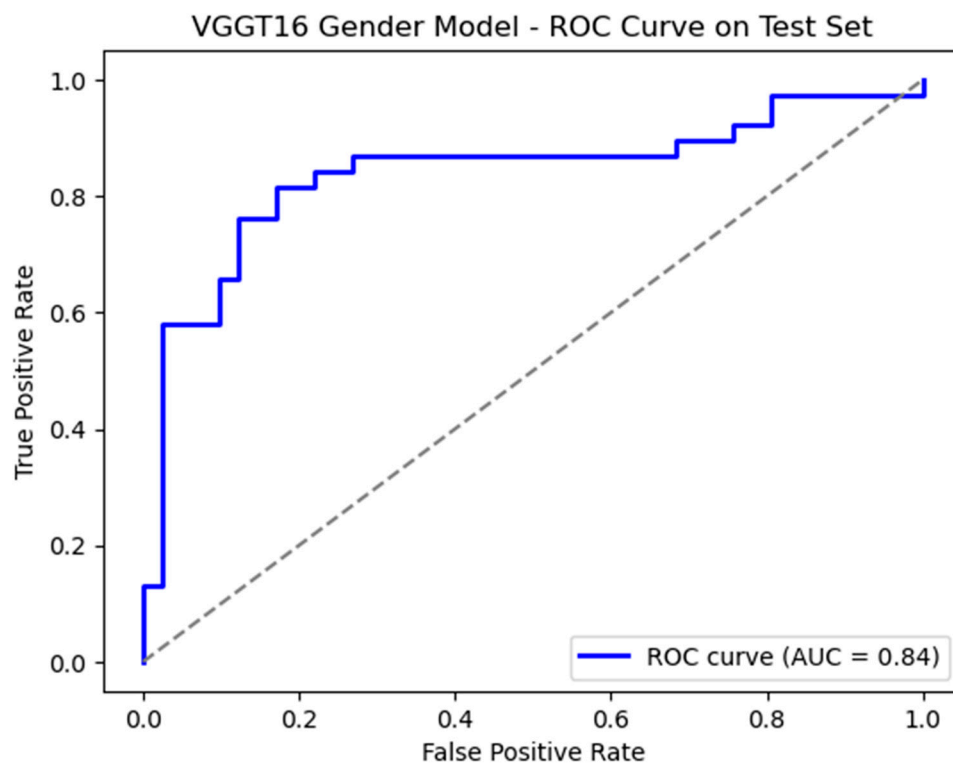
(**B**)

Figure 10. *Cont.*

(**C**)

(**D**)

Figure 10. *Cont.*

(**E**)
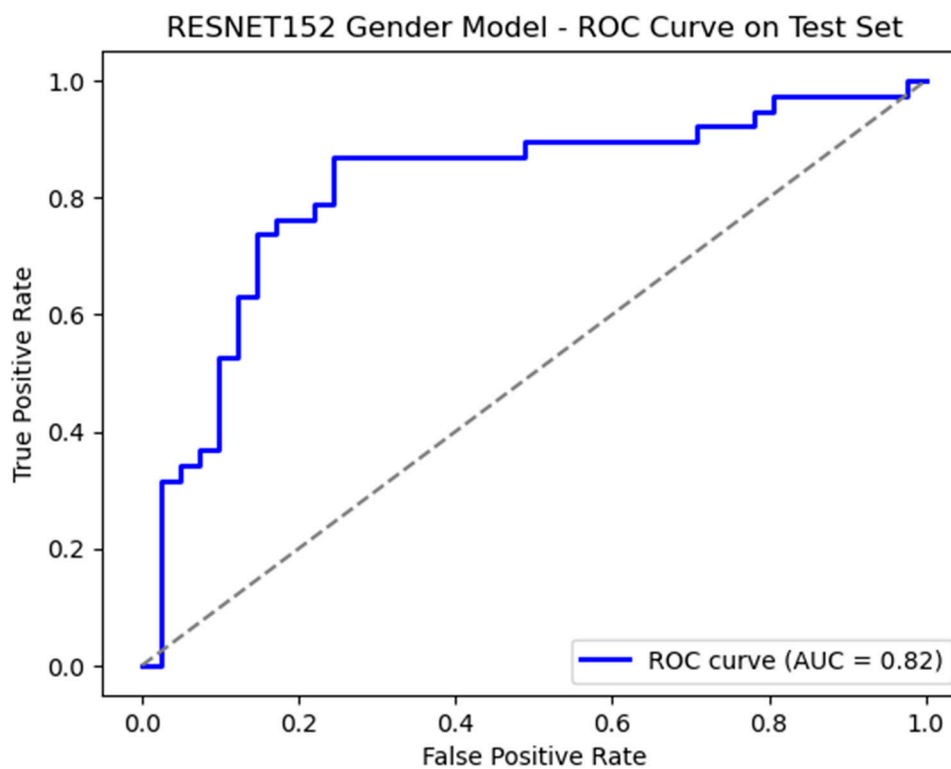
(**F**)

Figure 10. *Cont.*
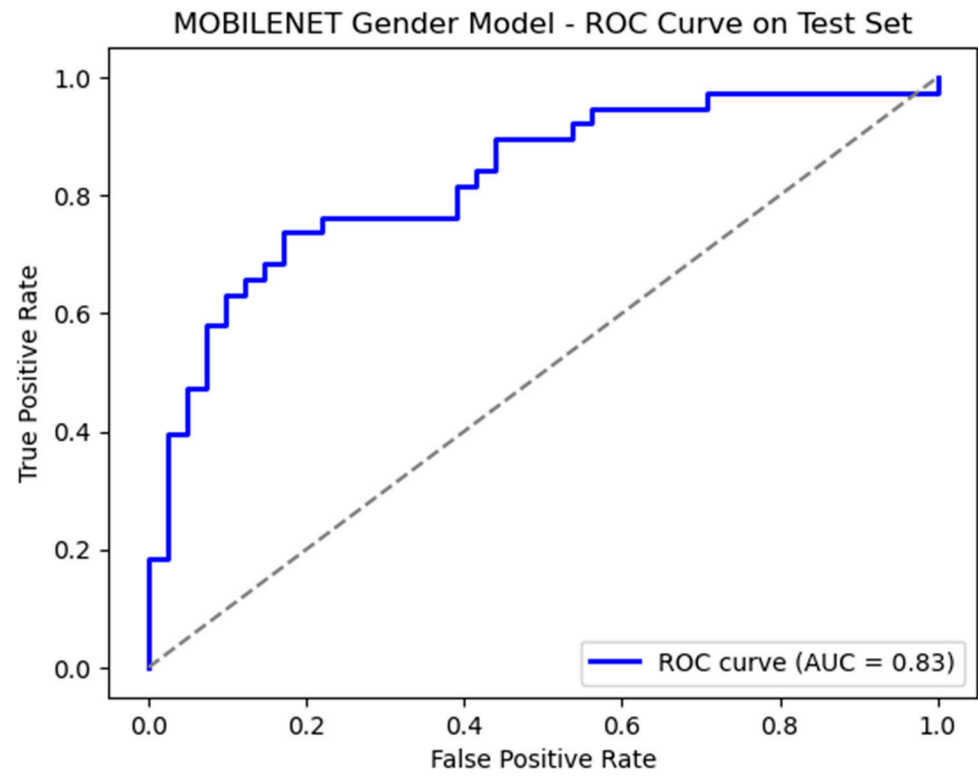
(**G**)
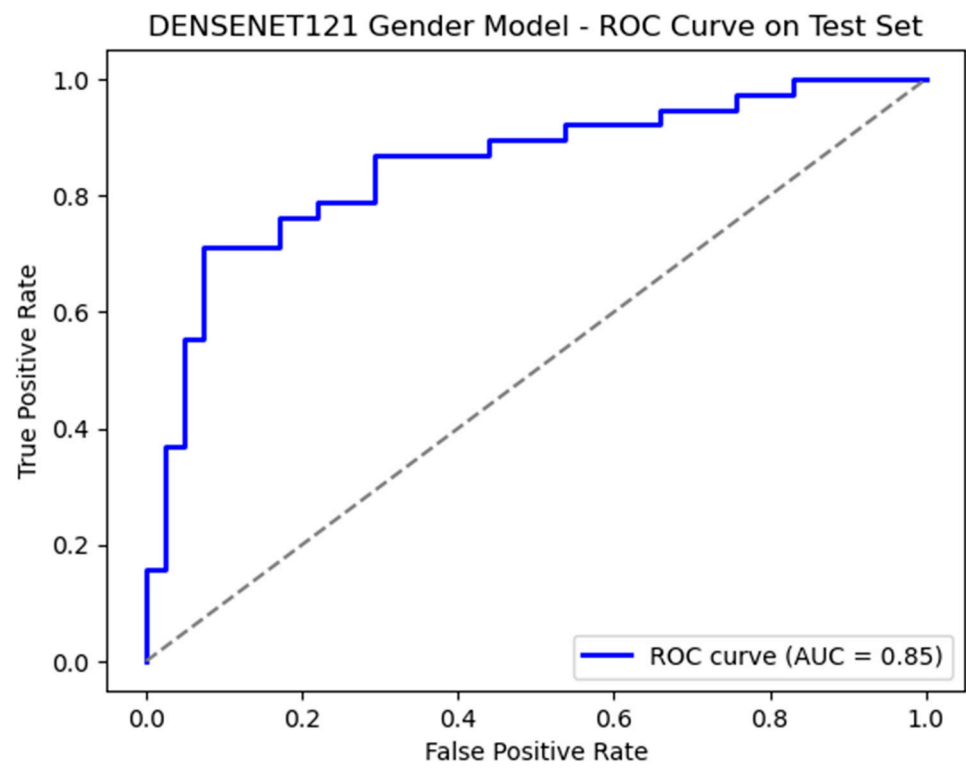
(**H**)

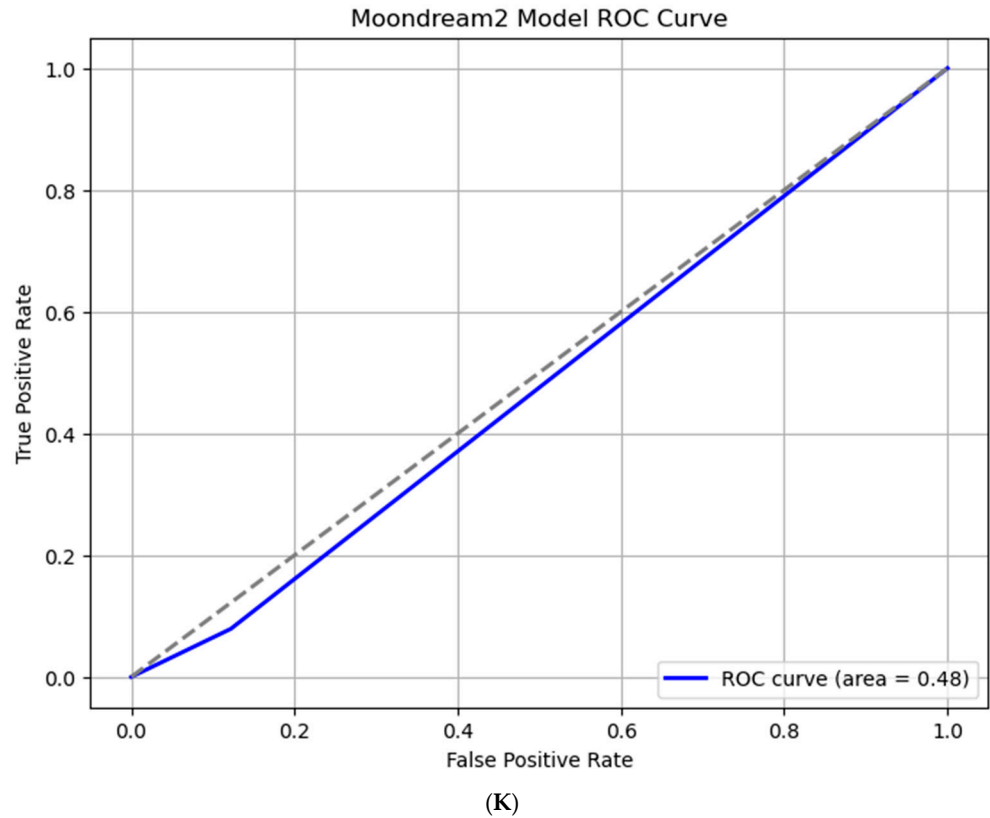**Figure 10.** *Cont.*

(**I**)

(**J**)

Figure 10. *Cont.*

(**K**)
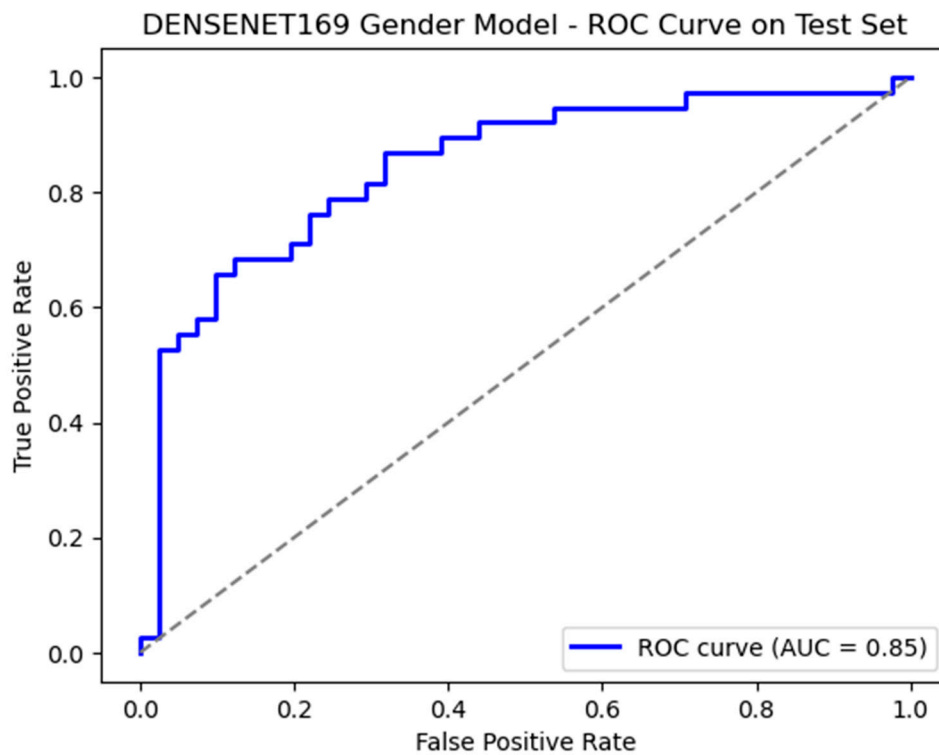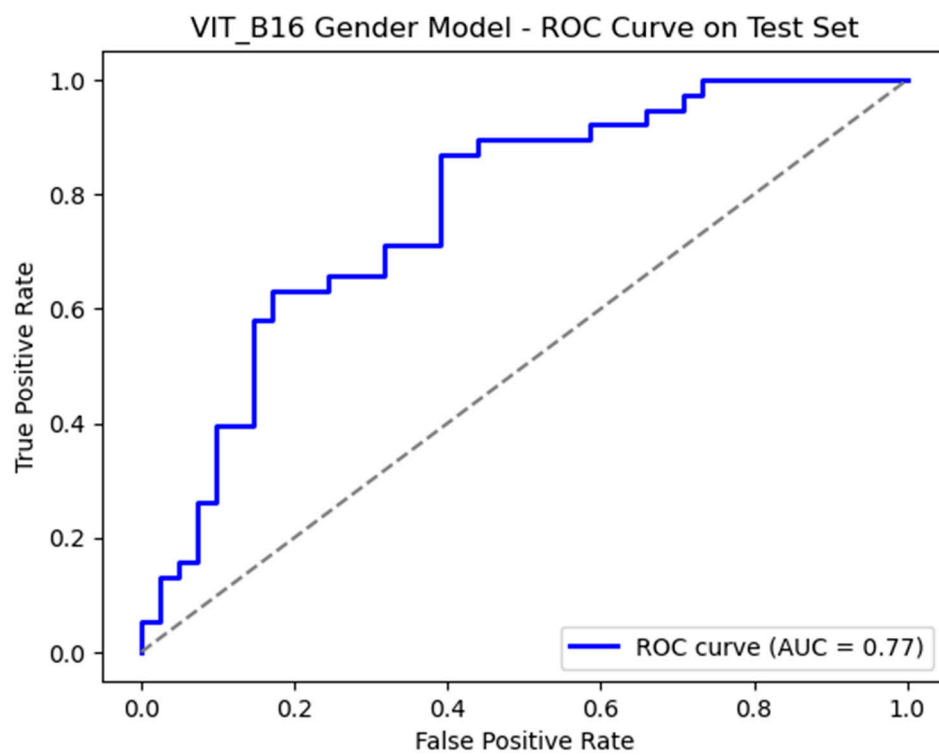
Figure 10. (**A**) ROC graph for CNN sex model; (**B**) ROC graph for VGG16 sex model; (**C**) ROC graph for VGG19 sex model; (**D**) ROC graph for ResNet50 sex model; (**E**) ROC graph for ResNet101 sex model; (**F**) ROC graph for ResNet152 sex model; (**G**) ROC graph for MobileNet sex model; (**H**) ROC graph for DenseNet121 sex model; (**I**) ROC graph for DenseNet169 sex model; (**J**) ROC graph for VIT_B16 sex model; (**K**) ROC graph for Moondream2 model.

3.6. Confusion Matrices of Sex Models

The confusion matrices of the evaluated models have been aggregated and are presented in Table 5.

Table 5. Confusion matrices of predicted vs. actual outcomes for the models in predicting biological sex.

Models	Male		Female	
	Predicted	**Actual**	**Predicted**	**Actual**
CNN	35	41	32	38
VGG16	34	41	31	38
VGG19	37	41	25	38
ResNet50	15	41	33	38
ResNet101	20	41	32	38
ResNet152	36	41	23	38
MobileNet	38	41	22	38
DenseNet121	37	41	27	38
DenseNet169	36	41	26	38
Vision Transformer	34	41	23	38
Moondream2	36	41	3	38

The CNN exhibited the most balanced performance, accurately predicting 35 out of 41 males and 32 out of 38 females, indicating its robustness in handling both classes. MobileNet achieved the highest accuracy for male predictions (38 correct predictions) but demonstrated suboptimal performance for females, with only 22 correct predictions.

While DenseNet121 performed strongly for males (37 correct predictions), it fell short for females, with only 27 correct predictions. Among the ResNet-based models, ResNet50 significantly underperformed for males (15 correct predictions) despite a relatively strong performance for females (33 correct). ResNet101 achieved a slightly better balance, with 20 correct male and 32 correct female predictions, whereas ResNet152 demonstrated improved accuracy for males (36 correct) but showed limitations for females (23 correct). The Vision Transformer exhibited a similar pattern to ResNet152, with 34 accurate male predictions but only 23 accurate predictions for females, suggesting a bias toward the male class. Finally, Moondream2 did not perform well, with only three correct female predictions, highlighting its inadequacy for this task. Overall, the CNN emerged as the most balanced and reliable model, while ResNet variants and Vision Transformer showed varying degrees of class imbalance, favouring male predictions.

3.7. Inference Time

Table 6 presents the inference times of the evaluated models. MobileNet and the CNN had the fastest inference times at 12.658 ms, making them well suited for real-time applications. Conversely, Vision Transformer (164.557 ms) and Moondream2 (4481.013 ms) had significantly slower inference times. Moondream2's prolonged processing time further underscores its inefficiency for sex prediction tasks, indicating a need for optimisation in future research efforts.

Table 6. Inference time for sex prediction.

Model	Time Taken in Milliseconds
CNN	12.658
VGG16	88.608
VGG19	113.924
ResNet50	37.975
ResNet101	50.633
ResNet152	101.266
MobileNet	12.658
DenseNet121	25.316
DenseNet169	37.975
Vision Transformer	164.557
Moondream2	4481.013

4. Discussion

The current study assessed multiple deep learning neural networks and VLMs for their ability to estimate patient sex and age from panoramic dental radiographs. The study aimed to evaluate the performance, limitations, and potential applications of these models in clinical sciences and jurisprudence.

Artificial intelligence (AI) has demonstrated an ability to match or even surpass human performance in specific tasks such as triage, age estimation, and effectively communicating findings [43,44]. However, the appropriate use of AI is hindered by challenges related to data governance [45]. A key limitation lies in the lack of regulations for sharing deidentified radiograph data, as no established consent process currently allows patients to control how their data are used for AI training. This highlights the practical importance of open access datasets, such as the Tufts dataset, and emphasises the need for methodological advancements to develop robust, practical AI models [46].

AI systems, particularly those using deep learning methods, operate on gradient-based learning, extracting features deemed significant for their tasks [37]. However, as models grow more complex, their interpretability diminishes, reducing transparency in

the decision-making process. This trade-off between complexity and interpretability is particularly critical in sensitive applications, such as forensic age estimation, where clarity and accountability are paramount [47].

The models evaluated in this study demonstrated varying levels of success in sex classification and age regression tasks. The CNN, MobileNet, and DenseNet169 achieved higher accuracy and AUC scores for sex classification, making them strong performers in this domain. In contrast, models such as ResNet50 and ResNet101 struggled with classification, likely due to their inefficiency in handling smaller datasets effectively.

Moondream2, the sole multimodal deep learning model in the study capable of processing both text and image inputs simultaneously, was expected to estimate both age and sex. While it successfully identified anatomical landmarks from radiographic images, it was unable to perform age regression, limiting its utility to sex classification. However, even for sex classification, Moondream2 exhibited limitations in generalizing predictions effectively. This shortcoming can be attributed to its lightweight architecture and lack of optimisation for domain-specific tasks, highlighting an area for further research and development.

Limitations and Future Recommendations

Despite efforts to remove class imbalances, the sex classification models in this investigation performed poorly, especially for female predictions. The dataset's sex distribution may have contributed to biased outcomes, since certain models showed lower precision and recall for females.

Having a larger and more diverse dataset with age labels could potentially lead to improved results in the regression task. A greater number of labelled images would provide the model with more examples to learn from, enhancing its ability to generalize and reduce overfitting. Additionally, a more varied dataset would help ensure better representation across different age groups, improving the model's accuracy and reducing potential bias in age prediction.

The study did not explore larger VLMs, which may have provided a more comprehensive understanding of the classes due to their advanced capabilities. However, their significant computational requirements and resource constraints rendered them impractical for the scope of this research.

The following recommendations can be made for future research:

1. Addressing sex classification prejudice: Future research should focus on finding ways to reduce the sex prejudice found in this study. This could include approaches such as class balance through data augmentation, using class weights in model training, or investigating domain-specific architectures that perform better on sex categorisation tasks. Additional strategies for equalizing sex representation in the dataset could increase model fairness and accuracy.

2. Expanding and diversifying the dataset: To improve regression task performance, future studies should use a larger, more diversified dataset with a wider variety of age labels. A larger dataset would provide the model with more generalization capacity, lowering the danger of overfitting and enhancing predictions across age groups. Furthermore, collecting data from several demographic groups may result in greater representation and reduced bias, especially for age-related predictions.

3. Exploring the potential of larger VLMs: Given the potential capabilities of larger VLMs, future studies could investigate their use in medical imaging applications, particularly sex and age prediction. Although these models necessitate significant computational resources, developments in model optimisation, transfer learning, and distributed computing may make them more viable. Researchers could also examine

hybrid models that combine the strengths of both smaller and larger VLMs, potentially improving prediction accuracy while maintaining computational economy.

5. Conclusions

The following conclusions can be drawn from the current research:

1. Sex classification: Convolutional neural networks (CNNs) and DenseNet models demonstrated strong performance, achieving an accuracy of approximately 85%.
2. Age estimation: DenseNet outperformed other models, achieving the highest performance with the lowest mean squared error in estimating age from panoramic radiographs.
3. Inference time analysis: MobileNet and CNNs emerged as the fastest models, making them suitable for real-time applications. DenseNet models, while slightly slower, offered an optimal balance between computational efficiency and accuracy.
4. Vision–language models (VLMs) and multimodal systems: These models require further research and development to enhance their competitiveness and reliability for clinical applications.

Author Contributions: Conceptualization: S.S.A., N.R., T.A.F., S.A., R.A.H., J.D. and T.H.F.; Methodology: S.S.A., N.R., T.A.F., S.A. and R.A.H.; Software: S.S.A., N.R., T.A.F., S.A. and R.A.H.; Validation: T.H.F.; Investigation: S.S.A., N.R., T.A.F., S.A. and R.A.H.; Data Curation: T.H.F.; Writing Original Draft: S.S.A. and S.A.; Writing—Review and Editing: T.H.F. and J.D.; Supervision: S.A. and T.H.F.; Project Administration: J.D. and T.H.F.; Funding Acquisition: J.D. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: The study was approved by the University of Adelaide Human Research and Ethics Committee (HREC-2023-073, 2023-4-26) and the Institutional Review Board of North South University (2023/OR-NSU/IRB/0503, 2023-6-15).

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/Salem1901/Comparative-Analysis-of-Age-and-Gender-Prediction-Using-Deep-Learning-Architectures-.git (last accessed on 21 November 2024).

Conflicts of Interest: Author Taseef Hasan Farook was employed by the company Research and Innovations, Dental Loop Pty Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bassed, R.B.; Briggs, C.; Drummer, O.H. Age estimation using CT imaging of the third molar tooth, the medial clavicular epiphysis, and the spheno-occipital synchondrosis: A multifactorial approach. *Forensic. Sci. Int.* **2011**, *212*, 273.e1. [CrossRef] [PubMed]
2. Čular, L.; Tomaić, M.; Subašić, M.; Šarić, T.; Sajković, V.; Vodanović, M. Dental age estimation from panoramic X-ray images using statistical models. In Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis, Ljubljana, Slovenia, 18–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 25–30.
3. Mathew, N.S.; Chatra, L.; Shenoy, P.; Veena, K.M.; Prabhu, R.V.; Sujatha, B.K. Gender determination in panoramic radiographs, utilizing mandibular ramus parameters: A cross-sectional study. *J. Dent. Res. Rev.* **2017**, *4*, 32–35. [CrossRef]
4. Ratson, T.; Dagon, N.; Aderet, N.; Dolev, E.; Laviv, A.; Davidovitch, M.; Blumer, S. Assessing Children's Dental Age with Panoramic Radiographs. *Children* **2022**, *9*, 1877. [CrossRef]
5. Farook, T.H.; Rashid, F.; Ahmed, S.; Dudley, J. Clinical machine learning in parafunctional and altered functional occlusion: A systematic review. *J. Prosthet. Dent.* 2023, *in press*. [CrossRef]
6. Silva, G.; Oliveira, L.; Pithon, M. Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert. Syst. Appl.* **2018**, *107*, 15–31. [CrossRef]

7.   Jader, G.; Fontineli, J.; Ruiz, M.; Abdalla, K.; Pithon, M.; Oliveira, L. Deep instance segmentation of teeth in panoramic X-ray images. In Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Paraná, Brazil, 29 October–1 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 400–407.

8.   Verma, V. Introducing Moondream2: A Tiny Vision-Language Model. 2024. Available online: https://www.analyticsvidhya.com/blog/2024/03/introducing-moondream2-a-tiny-vision-language-model/ (accessed on 24 November 2024).

9.   Ikechukwu, A.V.; Murali, S.; Deepu, R.; Shivamurthy, R.C. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. *Glob. Transit. Proc.* **2021**, *2*, 375–381. [CrossRef]

10.   Khan, Z.; Khan, F.G.; Khan, A.; Rehman, Z.U.; Shah, S.; Qummar, S.; Ali, F.; Pack, S. Diabetic retinopathy detection using VGG-NIN a deep learning architecture. *IEEE Access* **2021**, *9*, 61408–61416. [CrossRef]

11.   Liu, D.; Liu, Y.; Dong, L. G-ResNet: Improved ResNet for brain tumor classification. In Proceedings of the Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, 12–15 December 2019; Proceedings, Part i 26. Springer: Berlin/Heidelberg, Germany, 2019; pp. 535–545.

12.   Riasatian, A.; Babaie, M.; Maleki, D.; Kalra, S.; Valipour, M.; Hemati, S.; Zaveri, M.; Safarpoor, A.; Shafiei, S.; Afshari, M.; et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Med. Image Anal.* **2021**, *70*, 102032. [CrossRef]

13.   Aboshi, H.; Takahashi, T.; Komuro, T. Age estimation using microfocus X-ray computed tomography of lower premolars. *Forensic. Sci. Int.* **2010**, *200*, 35–40. [CrossRef] [PubMed]

14.   Asif, M.K.; Nambiar, P.; Mani, S.A.; Ibrahim, N.B.; Khan, I.M.; Sukumaran, P. Dental age estimation employing CBCT scans enhanced with Mimics software: Comparison of two different approaches using pulp/tooth volumetric analysis. *J. Forensic Leg. Med.* **2018**, *54*, 53–61. [CrossRef]

15.   Asif, M.K.; Nambiar, P.; Mani, S.A.; Ibrahim, N.B.; Khan, I.M.; Lokman, N.B. Dental age estimation in Malaysian adults based on volumetric analysis of pulp/tooth ratio using CBCT data. *Leg. Med.* **2019**, *36*, 50–58. [CrossRef]

16.   Arian, M.S.H.; Rakib, M.T.A.; Ali, S.; Ahmed, S.; Farook, T.H.; Mohammed, N.; Dudley, J. Pseudo labelling workflow, margin losses, hard triplet mining, and PENViT backbone for explainable age and biological gender estimation using dental panoramic radiographs. *SN Appl. Sci.* **2023**, *5*, 279. [CrossRef]

17.   Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:14091556.

18.   He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

19.   Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:170404861.

20.   Li, Y.; Huang, Z.; Dong, X.; Liang, W.; Xue, H.; Zhang, L.; Zhang, Y.; Deng, Z. Forensic age estimation for pelvic X-ray images using deep learning. *Eur. Radiol.* **2019**, *29*, 2322–2329. [CrossRef] [PubMed]

21.   Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

22.   Oktay, A.B. Tooth detection with convolutional neural networks. In Proceedings of the 2017 Medical Technologies National Congress (TIPTEKNO), Antalya, Türkiye, 27–29 October 2016; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.

23.   Kuo, Y.-F.; Lin, S.-Y.; Wu, C.H.; Chen, S.-L.; Lin, T.-L.; Lin, N.-H.; Mai, C.-H.; Villaverde, J.F. A convolutional neural network approach for dental panoramic radiographs classification. *J. Med. Imaging Health Inform.* **2017**, *7*, 1693–1704. [CrossRef]

24.   Farhadian, M.; Salemi, F.; Saati, S.; Nafisi, N. Dental age estimation using the pulp-to-tooth ratio in canines by neural networks. *Imaging Sci. Dent.* **2019**, *49*, 19–26. [CrossRef] [PubMed]

25.   Sironi, E.; Taroni, F.; Baldinotti, C.; Nardi, C.; Norelli, G.-A.; Gallidabino, M.; Pinchi, V. Age estimation by assessment of pulp chamber volume: A Bayesian network for the evaluation of dental evidence. *Int. J. Legal Med.* **2018**, *132*, 1125–1138. [CrossRef]

26.   Lu, J.; Liong, V.E.; Zhou, J. Cost-sensitive local binary feature learning for facial age estimation. *IEEE Trans. Image Process.* **2015**, *24*, 5356–5368. [CrossRef]

27.   Milošević, D.; Vodanović, M.; Galić, I.; Subašić, M. Automated estimation of chronological age from panoramic dental X-ray images using deep learning. *Expert. Syst. Appl.* **2022**, *189*, 116038. [CrossRef]

28.   De Back, W.; Seurig, S.; Wagner, S.; Marré, B.; Roeder, I.; Scherf, N. Forensic age estimation with Bayesian convolutional neural networks based on panoramic dental X-ray imaging. In Proceedings of the Machine Learning Research 2019, MIDL 2019, London, UK, 8–10 July 2019.

29.   Zhang, G.; Kurita, T. Age Estimation from the Age Period by Using Triplet Network. In Proceedings of the Frontiers of Computer Vision: 27th International Workshop, IW-FCV 2021, Daegu, Republic of Korea, 22–23 February 2021; Revised Selected Papers 27. Springer: Berlin/Heidelberg, Germany, 2021; pp. 81–92.

30.   Tuzoff, D.V.; Tuzova, L.N.; Bornstein, M.M.; Krasnov, A.S.; Kharchenko, M.A.; Nikolenko, S.I.; Sveshnikov, M.M.; Bednenko, G.B. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiol.* **2019**, *48*, 20180051. [CrossRef] [PubMed]

31. Fan, F.; Ke, W.; Dai, X.; Shi, L.; Liu, Y.; Lin, Y.; Cheng, Z.; Zhang, Y.; Chen, H.; Deng, Z. Semi-supervised automatic dental age and sex estimation using a hybrid transformer model. *Int. J. Legal Med.* **2023**, *137*, 721–731. [CrossRef] [PubMed]

32. Chu, P.; Bo, C.; Liang, X.; Yang, J.; Megalooikonomou, V.; Yang, F.; Huang, B.; Li, X.; Ling, H. Using octuplet siamese network for osteoporosis analysis on dental panoramic radiographs. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2579–2582.

33. Mualla, N.; Houssein, E.H.; Hassan, M.R. Dental Age Estimation Based on X-ray Images. *Comput. Mater. Contin.* **2020**, *62*, 591–605. [CrossRef]

34. Vila-Blanco, N.; Carreira, M.J.; Varas-Quintana, P.; Balsa-Castro, C.; Tomas, I. Deep neural networks for chronological age estimation from OPG images. *IEEE Trans. Med. Imaging* **2020**, *39*, 2374–2384. [CrossRef]

35. Atas, I. Human gender prediction based on deep transfer learning from panoramic radiograph images. *arXiv* **2022**, arXiv:220509850.

36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:201011929.

37. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]

38. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:230213971.

39. Salahin, S.M.S.; Ullaa, M.D.S.; Ahmed, S.; Mohammed, N.; Farook, T.H.; Dudley, J. One-Stage Methods of Computer Vision Object Detection to Classify Carious Lesions from Smartphone Imaging. *Oral* **2023**, *3*, 176–190. [CrossRef]

40. Shreffler, J.; Huecker, M.R. *Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios*; StatPearls Publishing: Treasure Island, FL, USA, 2020.

41. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1958**, *20*, 215–232. [CrossRef]

42. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]

43. Razzaki, S.; Baker, A.; Perov, Y.; Middleton, K.; Baxter, J.; Mullarkey, D.; Sangar, D.; Taliercio, M.; Butt, M.; Majeed, A.; et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv* **2018**, arXiv:180610698.

44. Goodman, R.S.; Patrinely, J.R.; Stone, C.A.; Zimmerman, E.; Donald, R.R.; Chang, S.S.; Berkowitz, S.T.; Finn, A.P.; Jahangir, E.; Scoville, E.A.; et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw. Open* **2023**, *6*, e2336483. [CrossRef]

45. Farook, T.H.; Dudley, J. Automation and deep (machine) learning in temporomandibular joint disorder radiomics. *A Syst. Rev. J. Oral. Rehabil.* **2023**, *50*, 501–521. [CrossRef] [PubMed]

46. Panetta, K.; Rajendran, R.; Ramesh, A.; Rao, S.P.; Agaian, S. Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 1650–1659. [CrossRef] [PubMed]

47. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2020.